Older Versions of the ROUGEeval Summarization Evaluation System were Easier to Fool

Jonas Sjöbergh [*]

KTH CSC, SE-10044 Stockholm, Sweden

Abstract

We show some limitations of the ROUGE evaluation method for automatic summarization. We present a method for automatic summarization based on a Markov model of the source text. By a simple greedy word selection strategy, summaries with high ROUGE-scores are generated. These summaries would however not be considered good by human readers. The method can be adapted to trick different settings of the ROUGEeval package.

Key words: automatic summarization, automatic evaluation, Markov models

1 Introduction

Automatic text summarization has been an active research area for many years. Evaluation of summarization is a quite hard problem. Often, a lot of manual labour is required, for instance by having humans read generated summaries and grading the quality of the summaries with regards to different aspects such as information content and text clarity. Manual labour is time consuming and expensive. Summarization is also subjective. The conception of what constitutes a good summary varies a lot between individuals, and of course also depending on the purpose of the summary.

ROUGEeval (Lin, 2003) is an attempt at automating the evaluation of summaries. Given a source text and a set of model summaries, normally manually written summaries, it calculates different scores for how well an automatically

[*] *Email address:* jsh@nada.kth.se (Jonas Sjöbergh).

generated summary corresponds to these model summaries. Most measurements are word n-gram based, calculating for instance the word overlap or word bigram overlap between a summary and the model summaries. Longest common (word) subsequence measurements are also available.

ROUGE-scores can easily be calculated for any number of systems or for new versions of a system under development, once the model summaries are available. This makes it a very useful tool for automatic summarization system development.

Intuitively it seems likely that given a reasonable attempt at summarizing a text, a high word n-gram overlap with known good summaries of the same text should correlate with human judgments of what a good summary is. ROUGE-scores have also been shown to correspond quite well with human judgments (Lin and Hovy, 2002, 2003a,b).

However, it is also quite easy to generate texts that score highly using the ROUGE-measurements without being good summaries. This is not very surprising, since the ROUGE-measurements are quite unsophisticated. Using unsophisticated measurements is normally a good thing, as long as they tell us something interesting, but often makes it possible to trick the measurements.

2 Markov Model Summarization

Our summarization system could be presented with many possibly impressive sounding words: it is not extraction based, but based on a language model of the source document; it is very language independent, resource lean and fast; it has the nice property that a 100 words summary of a text will be the first 100 words of a 200 words summary of the same text; it achieves higher ROUGE-scores than the agreement between model summaries; etc. The main point to note though is that it uses a simple greedy strategy that fools the ROUGE-measurements but unfortunately generates "bad" summaries. The summaries usually give the reader some idea of what the source text is about, though, but the text quality is low.

The summarizer first collects all the word bigram frequencies of the source text. The most frequent bigram is selected to start the summary. Then the basic strategy is to select the next word in such a way that the newly created word bigram in the summary is the one that has the highest frequency count of those possible given the last word in the summary.

Every time a bigram is used in the generated summary, the frequency count of this bigram is adjusted to account for the fact that it now occurs also in the

```
of the hurricane andrew had been injured and the storm caused
by the gulf of mexico and louisiana and at least 150 000 people
died on the us insurers expect to the florida and there are
likely to be concentrated among other insurers have been badly
damaged a result of damage caused in the state and to dollars
20bn of new orleans then to pay out on monday and new iberia
associated tornadoes devastated inhabitants of miami where
a million people and texas and then however the president's
insurance industry is the costliest disaster in florida as
a quarter of hurricane hugo which hit the industry analysts
cautioned that the bahamas on tuesday night or shut down
```

Fig. 1. A system generated summary.

summary. The frequency is adjusted by subtracting the length of the source document divided by the wanted length of the summary. So if a 100 words summary of an 8 000 words text is generated, each time a bigram is used its frequency count is lowered by 80.

This means that a simple Markov model has been constructed of the source text, and the most probable bigram is selected in each step; and that the probabilities are also adjusted in each step. The generated text will thus have roughly the same proportion of bigrams as the source text. An example of a system generated summary is shown in Figure 1.

Since in the ROUGE evaluation package it is common to specify the maximum allowed summary length in bytes, our method also works as specified above, but instead of using the frequency counts, the frequency counts divided by the byte length of the word to add to the text is used. Picking a complete new bigram instead of just one word is also allowed, if the frequency count per byte is higher for the new bigram than for the best single word, but this rarely happens. This method favors short words over slightly more common but longer words, since it might be more useful to have two shorter content words than one long one.

The ROUGE evaluations can also be done using stemming of the generated summaries and the model summaries. This means that even more content words can be squeezed into the summary, since no bytes need to be "wasted" on inflections in the summary. The summarizer in this case works exactly as above, the only extra thing that needs to be done is to feed it a stemmed version of the source text.

The same thing can be done for stop word removal. When evaluating without stop words, simply remove all stop words from the source text and then summarize in the normal way, making room for more content words when no bytes are "wasted" on stop words.

Since the bigram model only tries to include the important bigrams, it achieves quite low scores using the longer n-gram measurements. It is of course straightforward to extend the bigram model to a trigram model instead, by simply choosing new words so as to generate the highest frequency trigram possible. The data is of course much more sparse for trigrams than for bigrams, which leads to quite low evaluation scores, especially for short summaries.

Another method to improve the longer n-gram measurement scores was also implemented. After a summary has been generated using the bigram model described above some post processing was done. If a word occurs at least three times in the summary, the summary can be cut into chunks that start and end with this word. Changing the order of these chunks does not change the bigrams of the summary, but it changes the trigrams.

For all words occurring at least three times, all pairs of chunks starting and ending with this word were examined. If the sum of the frequency counts for the trigrams in the summary was higher if the two chunks changed places, they were interchanged.

The same can of course be done by dividing the summary on common word bigrams, to improve the quadrigram frequency counts. For 100 words summaries this has very little effect, though. It results in only one change in the evaluation corpus of about 100 documents. When generating longer summaries the effect is of course larger.

3 Evaluations

Here the ROUGE-scores using different settings for the ROUGEeval software and for the summarizer are presented. A comparison is also made with a baseline summarizer: Lead, the beginning of the source text up to the maximum allowed summary size limit. The agreements between the human written model summaries are also reported. These were calculated by simply removing one model summary at a time, treating it as a system summary and evaluating it using the rest of the model summaries. This was done for all model summaries and the mean value was calculated.

Texts from the Document Understanding Conference 2004 were used (DUC, 2006; Over and Yen, 2004). 114 documents with four manually written 100 words summaries each were summarized. The system summaries were generated to be 665 bytes long.

Four different evaluation runs were done. All runs used the following options: `ROUGE-1.5.5.pl -a -b 665 -n 4 -w 1.2`. The differences between the runs

System	1	2	3	4	L	W-1.2
Lead	31.6, 30.6	6.9, 6.6	2.4, 2.3	1.1, 1.0	31.8, 33.1	7.7, 13.5
Human	40.3, 40.3	11.3, 11.3	4.3, 4.3	2.0, 2.0	36.2, 36.3	12.4, 22.5
Trigrams	32.9, 30.7	7.0, 6.5	2.5, 2.4	0.9, 0.8	26.2, 24.5	9.1, 15.4
Bigrams	39.4, 35.4	11.8, 10.6	2.8, 2.5	0.8, 0.7	30.5, 27.5	10.4, 17.0
Short bigr.	41.2, 30.6	12.0, 8.9	2.5, 1.9	0.6, 0.5	33.5, 24.9	11.4, 15.3
Post proc.	41.2, 30.6	12.0, 8.9	2.9, 2.2	0.9, 0.7	33.7, 25.1	11.4, 15.4

Table 1
Recall and Precision values in % for different ROUGE-scores. 114 documents from DUC 2004, each with four human written 100 words summaries, were used. Summaries were truncated to 665 bytes when evaluating. Evaluated using ROUGE 1.5.5.

System	1	2	3	4	L	W-1.2
Lead	20.0, 20.0	5.8, 5.8	1.9, 1.8	0.7, 0.7	19.8, 19.8	6.2, 10.1
Human	31.5, 31.5	9.3, 9.3	3.0, 3.0	1.1, 1.1	28.8, 28.9	12.4, 20.1
Trigrams	19.9, 20.2	5.2, 5.2	1.7, 1.6	0.5, 0.4	17.2, 17.4	7.8, 12.8
Bigrams	26.2, 27.6	6.6, 6.8	1.6, 1.6	0.4, 0.4	21.2, 22.3	9.3, 15.9
Short bigr.	24.3, 24.6	5.7, 5.7	1.2, 1.2	0.2, 0.2	20.4, 20.8	9.0, 14.8
Bigr.+stop	29.4, 16.9	9.0, 5.2	2.3, 1.3	0.7, 0.4	25.6, 14.7	11.2, 10.4
Post pr.+stop	29.4, 16.9	9.0, 5.2	2.4, 1.4	0.7, 0.4	25.6, 14.7	11.2, 10.4

Table 2
As in Table 1, but evaluated with stop word removal.

were the use of stop word removal or stemming, the options `-s` and `-m`.

For every run, two bigram models and the trigram model were evaluated. These were implemented as described in the previous section, that is the bigram model selects each new word so as to generate the most common bigram possible of those in the original text, while taking previously generated bigrams into account. One bigram model used the frequency count per byte value, thus favoring short words, and one just used the frequency counts. When stemming or stop word removal was used in the evaluations, a version of the bigram model using the same options was also included, as was the same version but also using the trigram post processing method.

The results are shown in Tables 1 to 4. Scores are given for ROUGE-1, word overlap between a system generated summary and human written “gold standard” summaries; ROUGE-2, word bigram overlap; ROUGE-3, word trigram overlap; ROUGE-4, word quadrigram overlap; ROUGE-L, longest common

System	1	2	3	4	L	W-1.2
Lead	33.4, 32.2	7.3, 7.0	2.5, 2.4	1.1, 1.1	33.2, 32.1	7.9, 13.8
Human	42.6, 42.6	11.9, 11.9	4.5, 4.5	2.1, 2.1	38.0, 38.1	13.0, 23.6
Trigrams	34.8, 32.5	7.4, 6.9	2.7, 2.5	1.0, 0.9	27.3, 25.6	9.4, 16.0
Bigrams	41.4, 37.2	12.3, 11.1	2.9, 2.6	0.8, 0.7	31.7, 28.5	10.8, 17.6
Short bigr.	43.0, 31.9	12.5, 9.3	2.6, 1.9	0.6, 0.5	34.7, 25.8	11.7, 15.8
Bigr.+stem	43.9, 30.8	12.7, 8.9	2.5, 1.7	0.6, 0.4	35.3, 24.8	11.9, 15.1
Post pr.+stem	43.9, 30.8	12.7, 8.9	3.0, 2.1	0.9, 0.6	35.7, 25.0	12.0, 15.3

Table 3
As in Table 1, but evaluated with stemming.

System	1	2	3	4	L	W-1.2
Lead	22.6, 22.6	6.3, 6.3	2.1, 2.1	0.8, 0.8	22.4, 22.4	6.7, 10.9
Human	35.1, 35.1	10.2, 10.2	3.3, 3.2	1.2, 1.2	31.9, 32.0	13.5, 21.9
Trigrams	22.6, 22.9	5.7, 5.7	1.8, 1.8	0.5, 0.5	19.1, 19.4	8.5, 14.0
Bigrams	29.0, 30.5	7.1, 7.4	1.6, 1.7	0.4, 0.4	23.2, 24.5	10.1, 17.2
Short bigr.	26.6, 27.0	6.0, 6.1	1.3, 1.3	0.2, 0.2	22.1, 22.6	9.6, 15.9
Bigr.+stem+stop	34.5, 20.4	10.4, 6.1	2.5, 1.5	0.7, 0.4	27.7, 16.5	12.0, 11.5
P.pr.+stem+stop	34.5, 20.4	10.4, 6.1	2.6, 1.5	0.8, 0.5	28.8, 17.0	12.4, 11.9

Table 4
As in Table 1, but evaluated with stemming and stop word removal.

word subsequence between the system summary and the “gold standard” summaries; and ROUGE-W, also the longest common word subsequence, but weighted to give higher scores to words occurring consecutively.

The trigram model does not perform very well, though it sometimes at least has higher ROUGE-3 and ROUGE-4 scores than the bigram model. The reason the trigram model is weak is likely that there is insufficient data to gather reliable trigram statistics in the source text. The short summaries also make it difficult for the trigram model to include the appropriate amount of common trigrams. Since the choice is usually between including zero or one copies of a trigram where most trigrams “ideally” would be included more than zero times but are not frequent enough to warrant one whole copy, it is somewhat arbitrary if trigrams end up in the summary or not.

On the bigram level, many more bigrams “should” be included more than one time, and the statistics are less sparse in the source text. This gives very high recall figures for ROUGE-1 and ROUGE-2 using the bigram models.

Older versions of ROUGE, such as ROUGE 1.4.2, used in the DUC 2004 evaluations, only produced the recall figure. If one looks only at the recall value, it is quite easy to achieve ROUGE-1 and ROUGE-2 scores comparable to the agreement between the human written summaries, many times scoring higher than the human agreement. The longer n-gram measurements are harder to trick. On the other hand, they already give very low scores for the human agreement, in the range 1 – 4 %. Since the difference between baseline performance and human agreement is so small, it is hard to differentiate the performance between different systems using these measurements.

ROUGE-L and ROUGE-W give human agreement higher scores than the simple bigram texts. Removing stop words also tends to affect the human written summaries less than the bigram generated texts, that tend to score highly by including appropriate amounts of common stop words. Stemming does not seem to have that much of an effect.

Adapting the summary to the evaluation procedure gives large effects. Favoring short words is for instance bad when evaluating without stop words. Removing inflections or stop words when summarizing if these are disregarded in the evaluations also improves recall, though of course the precision drops quite a lot, since the summary is in effect longer in words despite being the same number of bytes.

The precision scores that newer versions of ROUGEeval calculate is a good way to detect that something is wrong with the bigram generated summaries. Trying to cram in many common short words or not using stop words at all if they are not considered in the evaluations of course lead to very low precision.

A small evaluation of the text quality of the generated summaries was also performed to show that while the ROUGE scores are high, the summaries are not what humans would call good summaries. Three human readers were asked to read five system generated summaries and five of the manually written model summaries. No mention of which group a summary belonged to was made, but it was quite obvious anyway so the readers were likely aware of which summaries were automatically generated.

The system generated summaries used the most readable options, so no stemming was performed and all stop words were left in the text. Summaries generated using other options are even less readable. The human readers assigned three scores to each summary, one for “text flow”, one for “understandability” and one for “overall impression”. These represent roughly how easy it is to read the summary, if the reader understands what the summary is about and finally if the reader subjectively thinks this is a good summary. All scores were given on a scale from 1 (very bad) to 5 (very good).

	Text Flow	Understandability	Overall Impression
System	1.1	1.7	1.2
Human	4.9	4.7	4.9

Table 5
Manual evaluation of text quality.

The results are shown in Table 5. Even though the evaluation corpus was very small, it is obvious that the quality of the generated summaries is very far from the human written summaries. In fact the text quality using the most readable system options is as expected very poor, though a reader might get some understanding of what the general contents are supposed to be.

4 Conclusions

We presented a very simple Markov model based summarizer that does not use text extraction. While it produces summaries that can probably give a good idea of the contents of the source text, the summaries are not what could reasonably be called good summaries. They do however achieve quite impressive scores on some ROUGE measurements, especially with regards to the previously commonly used recall measurements. For example a ROUGE-1 recall score of 41%, compared to only 40% for the agreement between the human written summaries.

The common substring based ROUGE-measurements, ROUGE-L and ROUGE-W, are not as easily tricked by our simple bigram model as the n-gram models, though the produced summaries are still on par with those of other automatic systems. ROUGE-W is of course more favorable for the system generated summaries since it gives higher weight to local word sequences which is what the system is good at generating. If higher scores are needed for the substring measurements, the summarizer can probably be adapted for this. For instance by reordering text chunks to improve the longest common substring with the original text instead of to improve trigram overlap.

Removal of stop words also separates the good model summaries from the bad system generated summaries. The longer n-gram based measurements, such as ROUGE-4, also have this property, but since the difference between human level performance (1 – 2 %) and baseline performance (0.5 – 1 %) is very small, they are of less use. Evaluating small improvements to a system under development using ROUGE-4 when on average humans have only one quadrigram in common would require a quite large test corpus.

Recent versions of ROUGEeval also calculate both precision and recall values, which is a good way to detect that the system generated summaries are not very good. It seems to be much harder to achieve high precision and recall than just a high recall. Achieving high precision instead of recall would probably be quite easy though.

In conclusion, we can say that one should use several of the ROUGE measurements together to get a good picture of the system performance. When looking at only one or two measurements, a system can probably be created that takes advantage of what is being evaluated and thus can trick the evaluation system. This can also be a problem when using automatic evaluation to tune the system performance. Since ROUGE is not an all encompassing measurement (and does not claim to be), the system might become tuned to high ROUGE scores without necessarily generating better summaries.

It should also be noted that in most larger evaluation efforts, such as the DUC conferences, ROUGE is only one part of the evaluation. A system such as the one presented here would get very low scores in the human readbility and other text quality measurements that are normally used.

Acknowledgments

We thank Viggo Kann for contributing valuable ideas.

References

DUC, 2006. Document understanding conferences. http://duc.nist.gov/.

Lin, C.-Y., 2003. ROUGE: Recall-oriented understudy for gisting evaluation. Http://www.isi.edu/~cyl/ROUGE/.

Lin, C.-Y., Hovy, E., 2002. Manual and automatic evaluation of summaries. In: Proceedings of the Workshop on Automatic Summarization, ACL-2002. Philadelphia, USA, pp. 45–51.

Lin, C.-Y., Hovy, E., 2003a. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: Proceedings of the 2003 Human Language Technology Conference (HLT-NAACL 2003). Edmonton, Canada.

Lin, C.-Y., Hovy, E., 2003b. The potential and limitations of automatic sentence extraction for summarization. In: HLT-NAACL 2003 Workshop: Text Summarization (DUC03). Edmonton, Canada.

Over, P., Yen, J., 2004. An introduction to DUC 2004 intrinsic evaluation of generic new text summarization systems. http://www-nlpir.nist.gov/projects/duc/pubs/2004slides/duc2004.intro.pdf.