

# Generation of Reference Summaries

**Martin Hassel**

IPLab, KTH KOD  
xmartin@nada.kth.se

**Hercules Dalianis**

DSV-KTH / Stockholm University  
hercules@kth.se

## Abstract

We have constructed an integrated web-based system for collection of extract-based corpora and for evaluation of summaries and summarization systems. During evaluation and examination of the collected and generated data we found that in a situation of low agreement among the informants the corpus gives unduly favors to summarization systems that use sentence position as a central weighting feature. The problem is discussed and a possible solution is outlined.

## 1. Background

When developing text summarizers and other information extraction tools it is extremely difficult to assess the performance of these tools. One reason for this is that evaluation is time-consuming and needs large manual efforts. When changing the architecture of the summarizer one needs to carry out the evaluation process again.

Therefore it would be fruitful to have a tool that directly can assess the result from a text summarizer repeatedly and automatically. We have for this reason constructed the KTH extract tool to create an extract corpus that can be used to evaluate text summarizers.

To create the extract corpus we need a large group of human informants. When the extract corpus is in place it can be used repeatedly with little effort. One other advantage is that one can create an extract corpus in any language and evaluate any language-dependant text summarizer, as long as one is sure about the quality of the corpus. In order to use the extract corpus for evaluation of a summarizer one needs careful preparation of the corpus, also it is important to discuss in what sense the extract corpus can correspond to the output of the summarizer.

The specific target for our evaluation is the SweSum text summarizer for Swedish news text and the DanSum<sup>1</sup> text summarizer for Danish news text.

SweSum is a text summarizer mainly developed to summarize Swedish news text (Dalianis 2000). SweSum works on sentence level – i.e. extracting sentences, judging the relevance of each sentence and then creating a shorter text (non-redundant extract) containing the highest-ranking sentences from the original text.

SweSum has been ported to English, Spanish, French, Danish, Norwegian, German and Farsi so far. SweSum is freely available online at <http://swesum.nada.kth.se>, and we have today around 2 200 visitors per month using it, mostly from American and Spanish universities.

### 1.1 Previous Research

Evaluating summaries and automatic text summarization systems is not a straightforward process. What exactly

makes a summary beneficial is an elusive property. Generally speaking there are two properties of the summary that must be measured when evaluating summaries and summarization systems: the Compression Ratio (how much shorter the summary is than the original);

$$CR = \text{length of Summary} / \text{length of Full Text}$$

and the Retention Ratio (how much information is retained);

$$RR = \text{information in Summary} / \text{information in Full Text}$$

Retention Ratio is sometimes also referred to as Omission Ratio, (Hovy 1999). An evaluation of a summarization system must at least in some way tackle both properties.

### 1.2 Evaluation methods

A first broad division in methods for evaluation automatic text summarization systems, as well as many other systems, is into intrinsic and extrinsic evaluation methods (Spark-Jones and Galliers 1995).

**Extrinsic evaluation** measures the efficiency and acceptability of the generated summaries in some task, for example relevance assessment or reading comprehension.

**Intrinsic evaluation** on the other hand measures the system in of itself. This is often done by comparison to some gold standard, which can be made by a reference summarization system or, more often than not, is man-made using informants. Intrinsic evaluation has mainly focused on the coherence and informativeness of summaries.

Summaries generated through extraction-based methods (cut-and-paste operations on phrase, sentence or paragraph level) sometimes suffer from parts of the summary being extracted out of context, resulting in coherence problem (e.g. dangling anaphors or gaps in the rhetorical structure of the summary). One way to measure this is to let subjects rank or grade summary sentences for coherence and then compare the grades for the summary sentences with the scores for reference summaries.

For single documents traditional precision and recall figures can be used to assess performance as well as utility figures and content based methods. Precision and

---

<sup>1</sup> DanSum is SweSum ported to Danish

recall are standard measures for Information Retrieval and are often combined in a so-called F-score. The main problems with these measures for text summarization is that they are not capable of distinguishing between many possible, but possibly equally good, summaries and that summaries that differ quite a lot content wise may get very similar scores.

Sentence rank is a more fine-grained approach than precision and recall (P&R), where the reference summary is constructed by ranking the sentences in the source text by worthiness of inclusion in a summary of the text. Correlation measures can then be applied to compare the generated summary with the reference summary. As in the case of P&R this method mainly applies to extraction based summaries, even if standard methods of sentence alignment with abstracts can be applied (see Marcu 1999, Jing and McKeown 1999).

The utility method (UM) (see Radev et al. 2000) allows reference summaries to consist of extraction units (sentences, paragraphs etc.) with fuzzy membership in the reference summary. In UM the reference summary contains all the sentences of the source document(s) with confidence values for their inclusion in the summary.

This method bears many similarities to the Majority Vote method (Hassel 2003) in that it, in contrast to standard P&R and Percent Agreement, allows summaries to be evaluated at different compression rates. UM is mainly useful for evaluating extraction-based summaries; more recent evaluation experiments has led to the development of the Relative Utility metric (Radev and Tam 2003).

### 1.3 Evaluation Tools

We have described a number of evaluation methods, now we need tools to use these methods. These tools will support us in creating a framework for more rigorous and repeatable evaluation procedure, partly by automating the comparison of summaries.

It is advantageous to build an extract corpus containing original full texts and their corresponding extracts, i.e. summaries strictly made by extraction of, in our case, whole sentences from an original text. Each extract, whether made by a human informant or a machine, is meant to be a true summary of the original, i.e. to retain the central points of the text to as large extent as possible

A number of tools have been developed for these purposes. Summary Evaluation Environment (SEE; Lin 2001) is an evaluation environment in which assessors can evaluate the quality of a summary, called the peer text, in comparison to a reference summary, called the model text. The texts involved in the evaluation are pre-processed by being broken up into a list of segments (phrases, sentences, clauses, etc.) During the evaluation phase, the two summaries are shown in two separate panels in SEE and interfaces are provided for assessors to judge both the content and the quality of model summaries. The assessor rates each unit and the overall structure of the model summary.

MEADeval (Winkel and Radev 2002) is a Perl toolkit for evaluating MEAD- and DUC-style extracts, by comparison to a reference summary (or "ideal" summary). MEADeval operates mainly on extract files, which describe the sentences contained in an extractive summary: which document each sentence came from and the number of each sentence within the source document – but can also perform some general content comparison. It supports a number of standard metrics, as well as some specialized

The ISI ROUGE - Automatic Summary Evaluation Package. ROUGE, short for Recall-Oriented Understudy for Gisting Evaluation, by Lin (2003). According to in-depth studies based on various statistical metrics and comparison to the results DUC-2002 (Hahn and Harman 2002), this evaluation method correlates surprisingly well with human evaluation (Lin and Hovy 2003). ROUGE is recall oriented, in contrast to the precision oriented BLEU script, and separately evaluates 1, 2, 3, and 4-grams. ROUGE has been verified for extraction-based summaries with a focus on content overlap. No correlation data for quality has been found so far.

However, none of the above tools have any support to help informants to create extracts, thus aiding in corpora building as well as evaluation.

## 2. KTH eXtract Corpus tool

At KTH, Stockholm, the KTH eXtract Corpus tool has been constructed (Hassel 2003, Dalianis et al. 2004). The tool assists in the collection of extract-based summaries provided by human informants and semi-automatic evaluation of machine generated extracts in order to easily evaluate the SweSum summarizer (Dalianis 2000). The KTH eXtract Corpus (KTHxc) contains a number of original full texts and several man-made extracts for each text. The tool assists in the construction of an extract corpus by guiding the human informant creating a summary in such a way that only full extract units (most often sentences) are selected for inclusion in the summary, (see figure 1). The interface allows for the reviewing of sentence selection at any time, as well as reviewing of the constructed summary before submitting it to the corpus.

Once the extract corpus is compiled, the corpus can be analyzed automatically in the sense that the inclusion of sentences in the various extracts for a given source text can easily be compared. Also available is the possibility of comparison on word level, a so-called vocabulary test. This allows for a quick adjustment and evaluation cycle in the development of an automatic summarizer. One can, for instance, adjust parameters of the summarizer and directly obtain feedback of the changes in performance, instead of having a slow, manual and time-consuming evaluation.

The KTH extract tool gathers statistics on how many times a specific extract unit from a text has been included in a number of different summaries. Thus, an ideal summary, or reference summary, can be composed using only the most frequently chosen sentences.

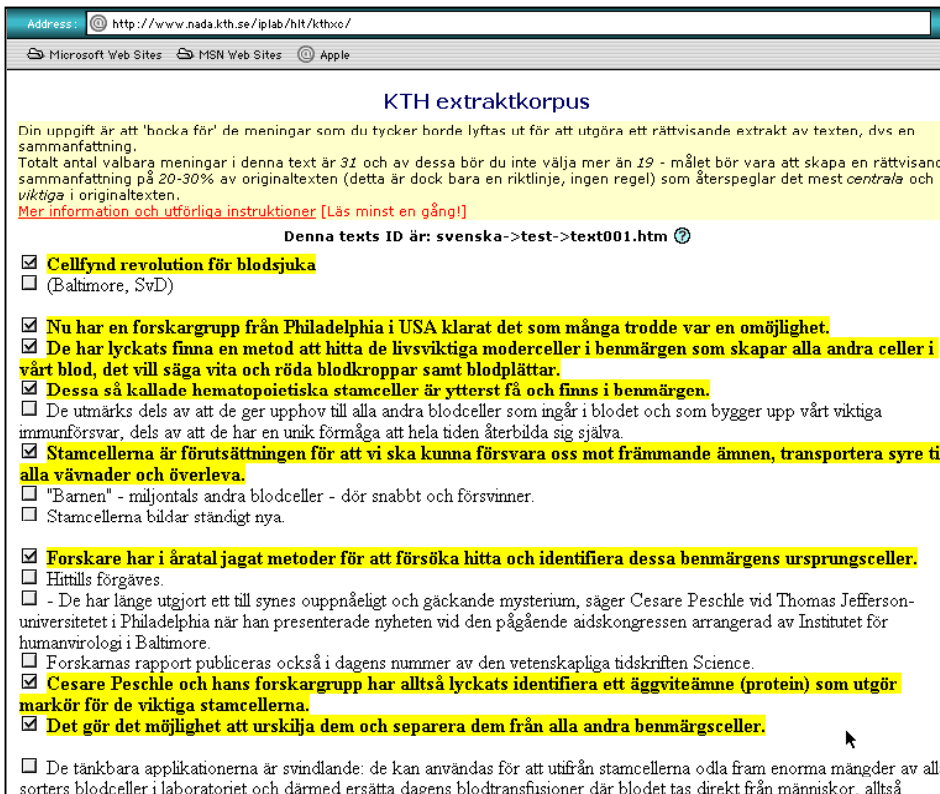


Figure 1. The KTH extract tool in action assisting an informant in creating an extract.

The reference summary can be generated at an arbitrary compression rate, i.e. the most high-ranking extract units up to a desired percentage of the original text. When several units with an equal number of votes and not all of them will fit into the reference summary units are extracted in order to prevent dangling anaphoric references.

Further statistical analysis can evaluate how close a particular extract is to a reference summary constructed by majority vote. The tool also has the ability to output reference summaries in the format SEE (see above) uses for human assessment. The KTHxc tool can easily be ported to other languages as the interface is completely separated from the code, and so far corpus collection and evaluation has been conducted for Swedish as well as Danish<sup>2</sup> news texts.

### 3. Creating the extract corpus

Three groups of texts have been collected during three iterations, two Swedish and one Danish group.

The Swedish extract corpus consists of a total of 301 Swedish text extracts submitted by 45 informants; average length of submitted extracts is currently 32.5 percent (31% and 34% respectively for group 1 and 2).

The Danish extract corpus at the present consists of 135 Danish text extracts submitted by 15 informants; average length of submitted extracts here is currently 32%.

<sup>2</sup> The University of Bergen has initiated a similar effort for Norwegian and has developed some similar tools (Dalanian et al. 2004).

During the extraction phase the human informants were allowed to submit extract summaries as short as 5 percent and up to 60 percent of the original text. The mean length of the submitted extracts varied between the texts, partly due to the length of the original text but also depending on the nature of the text (number of sentences, percentage of short respectively long sentences and of course also the texts rhetorical structure. However, the mean length of the submitted extracts over respectively of the three different groups was fairly consistent, ranging between 31 and 34 percent.

### 4. Evaluating

This experiment shows that there is not very much agreement between the informants on which sentences to select for the extract summary. The level of

agreement among the informants was calculated with a simple precision function. This is done per text and then a mean value was calculated over all texts in each group.

$$\frac{V_c}{N_s * N_x} * 100$$

In the function above  $V_c$  is the number of votes that are represented in the generated extract,  $N_s$  is the number of sentences represented in the same extract and  $N_x$  is the number of manmade extracts made for the original text the votes and sentences account for. This means that when all informants choose not only the same number of sentences but also exactly the same set of sentences the function will result in a precision, or agreement, of 100%.

We were prepared for a low agreement among the human extractors as to which sentences are good summary sentences as previous studies have shown this (for an overview see Mani 2001). In a previous study by Hassel (2003) using 11 informants and 96 extracts we found that when taking all selected extraction units into account for each text that there was only a mean agreement of 39.6% over ten texts.

This is however not so bad as it can seem at first glance. When generating a "gold standard" summary by presenting the most selected sentences up to a length of the mean length of all submitted extracts for a given text the precision, or the agreement level, rose to 68.9%. Very few of the sentences chosen for the gold standard were selected by as few as one third or less of the informants. Of course, even fewer sentences were selected by all

## KTH extraktkorpus

Nedan visas tre stycken selektionsmenyer. Den första för *språk* tillgängliga i korpusen, den andra för *texttyper* tillgängliga för ett valt språk och den tredje för *texter* tillgängliga för en viss texttyp (för ett visst språk). Du kan använda dessa för att orientera dig korpusen och välja ut specifika filer som du vill titta på.

svenska ▾ nyhetstexter ▾ text001.htm (28) ▾

Filnamn (text)	Antal extrakt	Kortaste extrakt	Längsta extrakt	Medellängd	Överlapp alla röster	Överlapp, medellängd	Abstrakt	Ändrad
text001.htm	28	22%	54%	37%	36%	60%		
text002.htm	19	15%	38%	27%	33%	60%		
text003.htm	22	14%	57%	30%	33%	57%		
text004.htm	16	19%	55%	31%	36%	70%		
text005.htm	24	19%	58%	33%	34%	59%		
text006.htm	29	20%	60%	32%	33%	62%		
text007.htm	26	20%	56%	39%	35%	60%		
text008.htm	22	24%	53%	37%	35%	62%		
text009.htm	24	20%	53%	32%	33%	56%		
text010.htm	28	15%	59%	41%	34%	65%		
<b>Totalt/Medel</b>	<b>238</b>	<b>19%</b>	<b>54%</b>	<b>34%</b>	<b>34%</b>	<b>61%</b>		

Figure 2. Overview of the web-based statistics of the extract corpus for Swedish (the text above is in Swedish)

informants. In fact, not even all informants could agree upon extracting the title or not when one was present.

Later we obtained more informants in form of students from our courses and we found that mean agreement decreased to 34% for all selected sentences and the mean agreement down to 61% for texts of summary length of the mean length of all man-made extracts, see figure 2.

These results are somewhat agreeing with work in manual indexing of texts (Bäckström 2000, van Dijk 1995). Bäckström found only 30 percent agreement, or index consistency, in selecting index terms in Swedish between two inexperienced human indexers and van Dijk (in French) found 60-80 percent agreement between two experienced indexers.

In order to verify this relationship between non-experienced and experienced informants we collected a second set of extracts using language consultant students as informants. This group has shown an agreement level of 73 percent when the most selected sentences up to a summary length of the mean length of all submitted extracts for each text. This is the highest agreement level of the three groups. This is probably the case since language consultants are well-trained readers and writers.

The extract summaries generated with SweSum<sup>3</sup> were then semi-automatically<sup>4</sup> compared on sentence and word level with the gold standard extracts generated by majority vote. We found that the summaries generated by SweSum and the gold standard summaries had between 47 and 62 percent of the sentences in common.

Of course this does not say much about how useful or coherent the system generated summaries were, only how well the different summaries created by SweSum corresponded to what our informants wanted to see in the summaries. That is, the figures represent how well SweSum mimics human selection.

However, what we find striking is the fact that SweSum apparently performs worse in regards to the reference summary when the agreement level rises. The reason for this seems to be that when the agreement level is high, which means that most votes are concentrated on a few sentences; it is less probable that the summarizer by “chance” hits the selected sentences. If, on the other hand, the agreement level is low and the votes are more evenly spread over the sentences, it might be the case that SweSum has a higher chance of hitting the same sentences as in the reference summary, since both system solve ties<sup>5</sup> by prioritizing the sentence occurring earliest in the text.

What we have here is a case of an evaluation method that evaluates partly along the same premises as the system it is evaluating. A system that does not put a high focus on sentence position might not get scored as favorably. This might, for example, be one reason that SweSum scores better than the Spanish lexical chain summarizer in the system-to-system comparison against Spanish model summaries made by Alonso i Alemany and Fuentes Fort (2003).

## 5. Tie Breaking

A more intelligent tie breaking scheme is clearly in need, preferably one that relies more on submitted data than on a general method that might be exploited by summarization system to be evaluated. One such is what we could call mutual exclusion; another could be called mutual inclusion.

Mutual exclusion as a tie breaking method would occur when in the statistics two or more sentences, or extract units, that have received the same number of votes show no informant overlap in the statistics. That is, when no, or very few, informants who have chosen one sentence have also chosen another we can assume that there is a reason for this, for example information redundancy.

Mutual inclusion would, on the other hand, occur when all, or almost all, informants have chosen the same set of

<sup>3</sup> The extracts were generated with SweSum by setting the desired compression rate to equal the mean length of all submitted man-made extracts for each text.

<sup>4</sup> The SweSum generated extracts were pasted into the evaluation view of the corpus interface.

<sup>5</sup> A tie is here defined as when two or more extract units receive equal score or, in the case of the reference summary, selection frequency.

sentences. This means that if a local high agreement occurs within the text this bond should be preserved.

## 6. Conclusions

In automatic text summarization, as well as in for example machine translation, there may be several equally good summaries for one specific source text effectively making evaluation against one rigid reference summary unsatisfactory. Also, evaluation methods that allow for evaluation at different compression rates should be favored, as experiments have shown that different compression rates are optimal for different text types or genres, or even different texts within a text type or genre. The semi-automatic evaluation methods presented in this paper attempts to tackle these properties. The described system mainly deals with content similarity between summaries. Summary quality, i.e. cohesion and coherence, must still be evaluated manually.

However, as we have shown, counting votes is not enough when constructing extraction-based corpora from many extracts. The distribution of the votes should also be taken into account in order to extract text binding clues hidden in the distribution of the votes.

Also, when generating the reference summaries from the selection statistics one must be aware of how the system solves tie breaking in case of equal number of votes and how this may favor the summarization system being evaluated.

## Acknowledgements

We would express gratitude to all the informants who have been willing to submit time and effort in compiling the selection statistics. We would also like to thank the participants of the ScandSum network and especially professor Koenraad de Smedt and Anja Liseth of the University of Bergen, Norway, for fruitful and invigorating discussions on summarization and evaluation.

## References

- Alonso i Alemany, L. and M. Fuentes Fort (2003). Integrating cohesion and coherence for automatic summarization. In Proceedings of EACL 2003, Budapest, Hungary.
- Bäckström, K. 2000. Marknadsundersökning och utvärdering av indexeringsprogram – en delstudie inom projektet Automatisk Indexering Magisteruppsats vid Institutionen för lingvistik Uppsala Universitet. (Master Thesis in Swedish)
- Dalianis, H. 2000. Swesum - a text summarizer for Swedish. Technical report TRITA-NA-P0015, IPLab-174 NADA, KTH, Sweden.
- Dalianis, H., M. Hassel, K. de Smedt, A. Liseth, T. C. Lech, and J. Wedekind. 2004. Porting and evaluation of automatic summarization. In H. Holmboe (editor), Nordisk Sprogteknologi 2003: Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Museum Tusulanums Forlag.
- van Dijk, B. 1995. Parlement Européen: Evaluation des opérations pilotes d'indexation automatique. (Convention spécifique no 52556) Rapport d'évaluation finale. (in French)
- Donaway, R. L., K. W. Drummey, and L. A. Mather 2000.. A Comparison of Rankings Produced by Summarization Evaluation Measures. In U. Hahn, C.-Y. Lin, I. Mani, and D. R. Radev (editors), Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and NAACL 2000.
- Hahn, U- and D. Harman (editors) 2002. Proceedings of the 2nd Document Understanding Conference. Philadelphia, PA.
- Hassel, M. 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In In the Proceedings of NODALIDA'03 - 14th Nordic Conference on Computational Linguistics, Reykjavik, Iceland.
- Jing, H. and K. R. McKeown. 1999. The Decomposition of Human-Written Summary Sentences. In M. Hearst, G. F., and R. Tong (editors), Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 129–136, University of California, Beekely.
- Lin, C.-Y. 2001. Summary Evaluation Environment. <http://www.isi.edu/~cyl/SEE>.
- Lin, C.-Y. 2003. ROUGE: Recall-oriented understudy for gisting evaluation. <http://www.isi.edu/~cyl/ROUGE/>.
- Lin, C.-Y. and E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada.
- Mani, I. 2001. Summarization Evaluation: An Overview. In Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.
- Marcu, D. .1999. The Automatic Construction of Large-Scale Corpora for Summarization Research. In M. Hearst, G. F., and R. Tong (editors), Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 137–144, University of California, Berkely.
- Radev, D. R., H. Jing, and M. Budzikowska. 2000. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In U. Hahn, C.-Y. Lin, I. Mani, and D. R. Radev (editors), Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and NAACL 2000, Seattle, WA.
- Spark-Jones, K. and J. R. Galliers. 1995. Evaluating Natural Language Processing Systems: An Analysis and Review. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.
- Winkel, A. and D. Radev. 2002. MEADeval: An evaluation framework for extractive summarization. <http://perun.si.umich.edu/clair/meadeval/>.