Summaries and the Process of Summarization

from
Evaluation of Automatic Text Summarization - A practical implementation
Martin Hassel 2004
Licentiate Thesis, KTH NADA

1 Introduction

Text summarization (or rather, automatic text summarization) is the technique where a computer automatically creates an abstract, or summary, of one or more texts. The initial interest in automatic shortening of texts was spawned during the sixties in American research libraries. A large amount of scientific papers and books were to be digitally stored and made searchable. However, the storage capacity was very limited and full papers and books could not be fit into databases those days. Therefore summaries were stored, indexed and made searchable. Sometimes the papers or books already had summaries attached to them, but in cases were no readymade summary was available one had to be created. Thus, the technique has been developed for many years (see Luhn 1958, Edmundson 1969, Salton 1988) and in recent years, with the increased use of the Internet, there have been an awakening interest for summarization techniques. Today the situation is quite the opposite from the situation in the sixties. Today storage is cheap and seemingly limitless. Digitally stored information is available in abundance and in a myriad of forms to an extent as to making it near impossible to manually search, sift and choose which information one should incorporate. This information must instead be filtered and extracted in order to avoiding drowning in it.

1.1 The World According to ISO

According to the documentation standard ISO 215:1986, a summary is a "brief restatement within the document (usually at the end) of its salient findings and conclusions, and is intended to complete the orientation of a reader who has studied the preceding text" while an abstract is, according to the same standard, a "Short representation of the content of a document without interpretation or criticism". In this paper, however, they will be used somewhat interchangeably. In the field of automatic text summarization it is customary to differentiate between extraction based, or cut-and-paste, summaries where the summary is composed of more or less edited fragments from the source text (this is the task of text extraction), as opposed to abstraction based summaries ("true abstracts") where the source text is transcribed into some formal representation and from this regenerated in a shorter more concise form, see Hovy and Lin (1997). A good overview of the field can be found in Mani and Maybury (1999).

1.2 In Defense of the Abstract

Why do we need automatic text summarization, indeed, why do we need summaries or abstracts at all? In the words of the American National Standards Institute (ANSI

1979) – "A well prepared abstract enables readers to identify the basic content of a document quickly and accurately, to determine its relevance to their interests, and thus to decide whether they need to read the document in its entirety". Actually the abstract is highly beneficial in several information acquisition tasks, some examples are given in (Borko and Bernier 1975):

- Abstracts promote current awareness
- Abstracts save reading time
- Abstracts facilitate selection
- Abstracts facilitate literature searches
- Abstracts improve indexing efficiency
- Abstracts aid in the preparation of reviews

Furthermore, human language is highly redundant, probably to facilitate error recovery in highly noisy channels. Mathematician and electrical engineer Claude E. Shannon has, for example, using a training data of 583 million words to create a trigram language model and corpus of 1 million words for testing, shown a 75% redundancy of English on letter level (Shannon 1951). Shannon initially defined redundancy as "the discovery of long-windedness" and accordingly it is not the amount of information that is increased, but the probability that the information reaches the recipient.

Fittingly, entropy experiments have also shown that humans are just as good at guessing the next letter – thus discerning the content of the text on a semantic level – after seeing 32 letters as after 10,000 letters (Burton and Licklider 1955). Other experiments (Morris et al. 1992) concerning reading comprehension of extraction based summaries compared to full documents have shown that extracts containing 20% or 30% of the source document are effective surrogates of the source document. Performance on 20% and 30% extracts is no different than informative abstracts.

Then, how does one go about constructing an abstract? Cremmins (1996) give us the following guidelines from the American National Standard for Writing Abstracts:

- State the purpose, methods, results, and conclusions presented in the original document, either in that order or with an initial emphasis on results and conclusions.
- Make the abstract as informative as the nature of the document will permit, so that readers may decide, quickly and accurately, whether they need to read the entire document.
- Avoid including background information or citing the work of others in the abstract, unless the study is a replication or evaluation of their work.
- Do not include information in the abstract that is not contained in the textual material being abstracted.
- Verify that all quantitative and qualitative information used in the abstract agrees with the information contained in the full text of the document.
- Use standard English and precise technical terms, and follow conventional grammar and punctuation rules.
- Give expanded versions of lesser known abbreviations and acronyms, and verbalize symbols that may be unfamiliar to readers of the abstract.
- Omit needless words, phrases, and sentences.

In automatic abstracting or summarization, however, one often distinguishes between informative and indicative summaries, where informative summaries intend to make reading of source unnecessary, if possible. Indicative summaries, on the other hand, act as an appetizer giving an indication of the content of the source text, thus making it easier for the reader to decide whether to read the whole text or not.

2 Automatic Text Summarization

Summarization approaches are often, as mentioned, divided into two groups, text extraction and text abstraction. Text extraction means to identify the most relevant passages in one or more documents, often using standard statistically based information retrieval techniques augmented with more or less shallow natural language processing and heuristics. These passages, often sentences or phrases, are then extracted and pasted together to form a non-redundant summary that is shorter than the original document with as little information loss as possible. Sometimes the extracted fragments are post-edited, for example by deleting subordinate clauses or joining incomplete clauses to form complete clauses (Jing and McKeown 2000, Jing 2000).

Text abstraction, being the more challenging task, is to parse the original text in a deep linguistic way, interpret the text semantically into a formal representation, find new more concise concepts to describe the text and then generate a new shorter text, an abstract, with the same information content. The parsing and interpretation of a text is an old research area that has been investigated for many years. In this area we have a wide spectrum of techniques and methods ranging from word by word parsing to rhetorical discourse parsing as well as more statistical methods or a mixture of all.

2.1 Application Areas

The application areas for automatic text summarization are extensive. As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. Information is published simultaneously on many media channels in different versions, for instance, a paper newspaper, web newspaper, WAP[1] newspaper, SMS[2] message, radio newscast, and a spoken newspaper for the visually impaired. Customisation of information for different channels and formats is an immense editing job that notably involves shortening of original texts.

Automatic text summarization can automate this work completely or at least assist in the process by producing a draft summary. Also, documents can be made accessible in other languages by first summarizing them before translation, which in many cases would be sufficient to establish the relevance of a foreign language document. Automatic text summarization can also be used to summarize a text before an automatic speech synthesizer reads it, thus reducing the time needed to absorb the key facts in a document. In particular, automatic text summarization can be used to prepare information for use

[1] *Wireless Application Protocol*, a secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers and communicators.

[2] *Short Message Service*, the transmission of short text messages to and from a mobile phone, fax machine and/or IP address. Messages must be no longer than 160 alpha-numeric characters.

in small mobile devices, such as a PDA,[3] which may need considerable reduction of content.

2.2 Approaches to Automatic Text Summarization

Automatic Text Summarization is a multi-facetted endeavor that typically branches out in several dimensions. There is no clear-cut path to follow and summarization systems usually tend to fall into several categories at once. According to (Sparck-Jones 1999, Lin and Hovy 2000, Baldwin et al. 2000), among others, we can roughly make the following inconclusive division.

Source Text (Input):

- Source: single-document vs. multi-document
- Language: monolingual vs. multilingual
- Genre: news vs. technical paper
- Specificity: domain-specific vs. general
- Length: short (1-2 page docs) vs. long (> 50 page docs)
- Media: text, graphics, audio, video, multi-media

Purpose:

- Use: generic vs. query-oriented
- Purpose: what is the summary used for (e.g. alert, preview, inform, digest, provide biographical information)?
- Audience: untargeted vs. targeted (slanted)

Summary (Output):

- Derivation: extract vs. abstract
- Format: running text, tables, geographical displays, timelines, charts, etc.
- Partiality: neutral vs. evaluative

The generated summaries can also be divided into different genres depending on their intended purpose, for example: headlines, outlines, minutes, biographies, abridgments, sound bites, movie summaries, chronologies, etc. (Mani and Maybury 1999). Consequently, a summarization system falls into at least one, often more than one, slot in each of the main categories above and thus must also be evaluated along several dimensions using different measures.

3 Summarization Evaluation

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 at least two properties of the summary that must be measured

[3] *Personal Digital Assistant* small mobile hand-held device that provides computing and information storage and retrieval capabilities, often contains calendar and address book functionality.

when evaluating summaries and summarization systems: the Compression Ratio (how much shorter the summary is than the original);

$$CR = \frac{length\ of\ Summary}{length\ of\ Full\ Text} \tag{1}$$

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

$$RR = \frac{information\ in\ Summary}{information\ in\ Full\ Text} \tag{2}$$

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

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

3.1 Intrinsic Evaluation

Intrinsic evaluation 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.

3.1.1 Summary Coherence

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, with the scores for the source sentences, or for that matter with the scores for other summarization systems.

3.1.2 Summary Informativeness

One way to measure the informativeness of the generated summary is to compare the generated summary with the text being summarized in an effort to assess how much information from the source is preserved in the condensation. Another is to compare the generated summary with a reference summary, measuring how much information in the reference summary is present in the generated summary. For single documents traditional precision and recall figures can be used to assess performance as well as utility figures (section 3.1.5) and content based methods (section 3.1.6).

3.1.3 Sentence Precision and Recall

Sentence recall measures how many of the sentences in the reference summary that are present in the generated summary and in a similar manner precision[4] can be calculated. Precision and recall are standard measures for Information Retrieval and are often combined in a so-called F-score (Van Rijsbergen 1979). The main problems with these measures for text summarization is that they are not capable of distinguishing between many possible, but equally good, summaries and that summaries that differ quite a lot content wise may get very similar scores.

3.1.4 Sentence Rank

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 (Marcu 1999, Jing and McKeown 1999).

3.1.5 The Utility Method

The utility method (UM) (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. Furthermore, the UM methods can be expanded to allow extraction units to exert negative support on one another. This is especially useful when evaluating multi-document summaries, where in case of one sentence making another redundant it can automatically penalize the evaluation score, i.e. a system that extracts two or more "equivalent" sentences gets penalized more than a system that extracts only one of the aforementioned sentences and a, say, less informative sentence (i.e. a sentence that has a lower confidence score).

This method bears many similarities to the Majority Vote method (Hassel 2003) in that it, in contrast to 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).

3.1.6 Content Similarity

Content similarity measures (Donaway et al. 2000) can be applied to evaluate the semantic content in both extraction based summaries and true abstracts. One such measure is the Vocabulary Test (VT) where standard Information Retrieval methods (see Salton and McGill 1983) are used to compare term frequency vectors calculated over stemmed or lemmatized summaries (extraction based or true abstracts) and reference summaries

[4]Precision is in this case defined as the number of sentences in the generated summary that are present in the reference summary.

of some sort. Controlled thesauri and "synonym sets" created with Latent Semantic Analysis (Landauer et al. 1998) or Random Indexing (Kanerva et al. 2000, Sahlgren 2001) can be used to reduce the terms in the vectors by combining the frequencies of terms deemed synonymous, thus allowing for greater variation among summaries. This is especially useful when evaluating abstracts.

The disadvantage of these methods is, however, that they are quite sensitive to negation and word order differences. With LSA[5] or RI[6] one must also be aware of the fact that these methods do not necessarily produce true synonym sets, these sets typically also include antonyms, hyponyms and other terms that occur in similar semantic contexts (on word or document level for RI and document level for LSA). These methods are however useful for extraction based summaries where little rewriting of the source fragments is done, and when comparing fragmentary summaries, such as key phrase summaries.

3.1.7 BLEU Scores

The idea here is that, as well as there may be many "perfect" translations of a given source sentence, there may be several equally good summaries for a single source document. These summaries may vary in word or sentence choice, or in word or sentence order even when they use the same words/sentences. Yet humans can clearly distinguish a good summary from a bad one.

The recent adoption of BLEU/NIST[7] scores (Papineni et al. 2001, NIST 2002) by the MT community for automatic evaluation of Machine Translation, Lin and Hovy (2003) have applied the same idea to the evaluation of summaries. They used automatically computed accumulative n-gram matching scores (NAMS) between ideal summaries and system summaries as a performance indicator. Only content words were used in forming n-grams and n-gram matches between the summaries being compared where treated as position independent. For comparison, IBM's BLEU evaluation script was also applied to the same summary set. However, this showed that direct application of the BLEU evaluation procedure does not always give good results.

3.2 Extrinsic Evaluation

Extrinsic evaluation on the other hand measures the efficiency and acceptability of the generated summaries in some task, for example relevance assessment or reading comprehension. Also, if the summary contains some sort of instructions, it is possible to measure to what extent it is possible to follow the instructions and the result thereof. Other possible measurable tasks are information gathering in a large document collection, the effort and time required to post-edit the machine generated summary for some specific purpose, or the summarization system's impact on a system of which it is part of, for example relevance feedback (query expansion) in a search engine or a question answering system.

[5]Latent Semantic Analyzis; sometimes also referred to as Latent Semantic Indexing.

[6]Random Indexing.

[7]Based on the superior F-ratios of information-weighted counts and the comparable correlations, a modification of IBM's formulation of the score was chosen as the evaluation measure that NIST will use to provide automatic evaluation to support MT research.

Several game like scenarios have been proposed as surface methods for summarization evaluation inspired by different disciplines, among these are The Shannon Game (information theory), The Question Game (task performance), The Classification/Categorization Game and Keyword Association (information retrieval).

3.2.1 The Shannon Game

The Shannon Game, which is a variant of Shannon's measures in Information Theory (Shannon 1948), is an attempt to quantify information content by guessing the next token, e.g. letter or word, thus recreating the original text. The idea has been adapted from Shannon's measures in Information Theory where you ask three groups of informants to reconstruct important passages from the source article having seen either the full text, a generated summary, or no text at all. The information retention is then measured in number of keystrokes it takes to recreate the original passage. Hovy (see Hovy and Marcu 1998) has shown that there is a magnitude of difference across the three levels (about factor 10 between each group). The problem is that Shannon's work is relative to the person doing the guessing and therefore implicitly conditioned on the reader's knowledge. The information measure will infallibly change with more knowledge of the language, the domain, etc.

3.2.2 The Question Game

The purpose of the Question Game is to test the readers' understanding of the summary and its ability to convey key facts of the source article. This evaluation task is carried out in two steps. First the testers read the source articles, marking central passages as they identify them. The testers then create questions that correspond to certain factual statements in the central passages. Next, assessors answer the questions 3 times: without seeing any text (baseline 1), after seeing a system generated summary, and after seeing original text (baseline 2). A summary successfully conveying the key facts of the source article should be able to answer most questions, i.e. being closer to baseline 2 than baseline 1. This evaluation scheme has for example been used in the TIPSTER SUMMAC text summarization evaluation Q&A[8] task, where Mani et al. (1998) found an informativeness ratio of accuracy to compression of about 1.5.

3.2.3 The Classification Game

In the classification game one tries to compare classifiability by asking assessors to classify either the source documents (testers) or the summaries (informants) into one of N categories. Correspondence of classification of summaries to originals is then measured. An applicable summary should be classified into the same category as its source document. Two versions of this test were run in SUMMAC (Mani et al. 1998).

[8]Question and Answering; a scenario where a subject is set to answer questions about a text given certain conditions, for example a summary of the original text.

3.2.4 Keyword Association

Keyword association is an inexpensive, but somewhat shallower, approach that relies on keywords associated (either manually or automatically) to the documents being summarized. For example Saggion and Lapalme (2000) presented human judges with summaries generated by their summarization system together with five lists of keywords taken from the source article as presented in the publication journal. The judges were then given the task to associate the each summary with the correct list of keywords. If successful the summary was said to cover the central aspects of the article since the keywords associated to the article by the publisher were content indicative. Its main advantage is that it requires no cumbersome manual annotation.

3.3 Evaluation Tools

In order to allow a more rigorous and repeatable evaluation procedure, partly by automating the comparison of summaries, it is advantageous to build an extract corpus containing originals and their extracts, i.e. summaries strictly made by extraction of 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 meaning of the text as good as possible. Since the sentence units of the original text and the various summaries are known entities, the construction and analysis of an extract corpus can almost completely be left to computer programs, if these are well-designed. A number of tools have been developed for these purposes.

3.3.1 Summary Evaluation Environment

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.) depending on the granularity of the evaluation. For example, when evaluating an extraction based summarization system that works on the sentence level, the texts are pre-processed by being broken up into sentences.

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 summaries. To measure content, the assessor proceeds through the summary being evaluated, unit by unit, and clicks on one or more associated units in the model summary. For each click, the assessor can specify whether the marked units express all, most, some or hardly any of the content of the clicked model unit. To measure quality, assessors rate grammaticality, cohesion, and coherence at five different levels: all, most, some, hardly any, or none. Quality is assessed both for each unit of the peer summary and for overall quality of the peer summary (coherence, length, content coverage, grammaticality, and organization of the peer text as a whole). Results can, of course, be saved and reloaded and altered at any time.

A special version of SEE 2.0 has for example been used in the DUC-2001 (Harman and Marcu 2001) intrinsic evaluation of generic news text summarization systems (Lin and Hovy 2002). In DUC-2001 the sentence was used as the smallest unit of evaluation.

3.3.2 MEADeval

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 it can also perform some general content comparison. It supports a number of standard metrics, as well as some specialized (see table 1).

A strong point of Perl, apart from platform independency, is the relative ease of adapting scripts and modules to fit a new summarization system. MEADeval has, for example, been successfully applied to summaries generated by a Spanish lexical chain summarizer and the SweSum[9] summarizer in a system-to-system comparison against model summaries (see Alonso i Alemany and Fuentes Fort 2003).

Extracts only	General text
precision	unigram overlap
recall	bigram overlap
normalized precision[10]	cosine[11]
normalized recall[12]	simple cosine[13]
kappa[14]	
relative utility[15]	
normalized relative utility	

Table 1: Metrics supported by MEADeval.

3.3.3 ISI ROUGE - Automatic Summary Evaluation Package

ROUGE, short for Recall-Oriented Understudy for Gisting Evaluation, by Lin (2003) is a very recent adaption of the IBM BLEU (see section 3.1.7) for Machine Translation that uses unigram co-occurrences between summary pairs. 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. Also, ROUGE does not apply any length penalty (brevity penalty), which is natural since text summarization involves compression of text and thus rather should reward shorter extract segment as long as they score

[9]SweSum mainly being a Swedish language text summarizer, also supports plug-in lexicons and heuristics for other languages, among these Spanish.

[10]Like precision, but normalized by the length (in words) of each sentence.

[11]The 2-norm (Euclidean Distance) between two vectors.

[12]Like recall, but normalized by the length (in words) of each sentence.

[13]Cosine without adjustments for Inverse Document Frequency (IDF).

[14]The simple kappa coefficient is a measure of interrater agreement compared to what could be expected due to chance alone.

[15]The Relative Utility and Normalized Relative Utility metrics are described in Radev and Tam (2003), also see section 3.1.5.

well for content. ROUGE has been verified for extraction based summaries with a focus on content overlap. No correlation data for quality has been found so far.

3.3.4 KTH eXtract Corpus and Tools

At the Royal Institute of Technology (KTH), Hassel has developed a tool for collection of extract based summaries provided by human informants and semi-automatic evaluation of machine generated extracts (Hassel 2003, Dalianis et al. 2004) in order to easily evaluate the SweSum summarizer (Dalianis 2000). The KTH eXtract Corpus (KTHxc) contains a number of original texts and several manual 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. 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 analysed automatically in the sense that the inclusion of sentences in the various extracts for a given source text can easily be compared. 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. Further statistical analysis can evaluate how close a particular extract is to the ideal one. The tool also has the ability to output reference summaries constructed by Majority Vote in the format SEE (described in section 3.3.1) uses for human assessment.

Obviously, the KTHxc tool could easily be ported to other languages and so far corpus collection and evaluation has been conducted for Swedish as well as Danish. The University of Bergen has initiated a similar effort for Norwegian and has developed some similar tools (Dalianis et al. 2004).

3.4 Famous Last Words

Most automatic text summarization systems today are extraction based systems. However, some recent work directed towards post-editing of extracted segments, e.g. sentence/phrase reduction and combination, thus at least creating the illusion of abstracting in some sense, leads to the situation where evaluation will have to tackle comparison of summaries that do not only differ in wording but maybe also in specificity and bias.

Furthermore, in automatic text summarization, as well as in for example machine translation, there may be several equally good summaries (or in the case of MT - translations) for one specific source text, effectively making evaluation against one rigid reference text 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 automatic evaluation methods presented in this paper mainly

deal with content similarity between summaries. Summary quality must still be evaluated manually.

Today, there is no single evaluation scheme that provides for all these aspects of the evaluation, so a mixture of methods described in this paper should perhaps be used in order to cover as many aspects as possible thus making the results comparable with those of other systems, shorten the system development cycle and support just-in-time comparison among different summarization methods. Clearly some sort of standardized evaluation framework is heavily in need in order to ensure replication of results and trustworthy comparison among summarization systems.

However, it is also important to keep users in the loop, at least in the end stages of system evaluation. One must never forget the target of the summaries being produced.

References

Alonso i Alemany, L. and M. Fuentes Fort (2003). Integrating Cohesion and Coherence for Automatic Summarization. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary.

ANSI (1979). American National Standard for Writing Abstracts. Technical report, American National Standards Institute, Inc., New York, NY. ANSI Z39.14.1979.

Baldwin, B., R. Donaway, E. Hovy, E. Liddy, I. Mani, D. Marcu, K. McKeown, V. Mittal, M. Moens, D. Radev, K. Sparck-Jones, B. Sundheim, S. Teufel, R. Weischedel, and M. White (2000). An Evaluation Road Map for Summarization Research. http://www-nlpir.nist.gov/projects/duc/papers/summarization.roadmap.doc.

Borko, H. and C. Bernier (1975). *Abstracting Concepts and Methods*. Academic Press, New York.

Burton, N. and J. Licklider (1955). Long-range constraints in the statistical structure of printed English. *American Journal of Psychology*, 68:650–655.

Cremmins, E. T. (1996). *The Art of Abstracting*. Information Resources Press, Arlington, VA, 2nd edition.

Dalianis, H. (2000). SweSum - A Text Summarizer for Swedish. Technical report, KTH NADA, Sweden.

Dalianis, H., M. Hassel, K. de Smedt, A. Liseth, T. C. Lech, and J. Wedekind (2004). Porting and evaluation of automatic summarization. In Holmboe, H. (editor), *Nordisk Sprogteknologi 2003: Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums Forlag.

Donaway, R. L., K. W. Drummey, and L. A. Mather (2000). A Comparison of Rankings Produced by Summarization Evaluation Measures. In Hahn, U., 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 the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 69–78. Association for Computational Linguistics.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Hahn, U. and D. Harman (editors) (2002). *Proceedings of the 2nd Document Understanding Conference*. Philadelphia, PA.

Harman, D. and D. Marcu (editors) (2001). *Proceedings of the 1st Document Understanding Conference*. New Orleans, LA.

Hassel, M. (2003). Exploitation of Named Entities in Automatic Text Summarization for Swedish. In *Proceedings of NODALIDA'03 - 14th Nordic Conference on Computational Linguistics*, Reykjavik, Iceland.

Hovy, E. (editor) (1999). *Multilingual Information Management: Current Levels and Future Abilities. Chapter 3 Cross-lingual Information Extraction and Automated Text Summarization.*

Hovy, E. and C.-Y. Lin (1997). Automated Text Summarization in SUMMARIST. In *Proceedings of the ACL97/EACL97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

Hovy, E. and D. Marcu (1998). Automated Text Summarization Tutorial at COLING/ACL'98. http://www.isi.edu/~marcu/acl-tutorial.ppt.

ISO 215:1986 (1986). Documentation – Presentation of Contributions to Periodicals and Other Serials. ISO 215:1986. Technical report, International Organisation for Standardisation.

Jing, H. (2000). Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pp. 310–315, Seattle,WA.

Jing, H. and K. R. McKeown (1999). The Decomposition of Human-Written Summary Sentences. In Hearst, M., 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.

Jing, H. and K. R. McKeown (2000). Cut and Paste-Based Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 178–185, Seattle, WA.

Kanerva, P., J. Kristoferson, and A. Holst (2000). Random Indexing of text samples for Latent Semantic Analysis. In Gleitman, L. and A. Josh (editors), *Proceedings 22nd Annual Conference of the Cognitive Science Society*, Pennsylvania.

Landauer, T. K., P. W. Foltz, and D. Laham (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

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 (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th COLING Conference*, Saarbrücken, Germany.

Lin, C.-Y. and E. Hovy (2002). Manual and Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

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.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Mani, I., D. House, G. Klein, L. Hirshman, L. Orbst, T. Firmin, M. Chrzanowski, and B. Sundheim (1998). The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

Mani, I. and M. T. Maybury (editors) (1999). *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.

Marcu, D. (1999). The Automatic Construction of Large-Scale Corpora for Summarization Research. In Hearst, M., 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.

Morris, A., G. Kasper, and D. Adams (1992). The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research*, 3(1):17–35.

NIST (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176, IBM.

Radev, D. R., H. Jing, and M. Budzikowska (2000). Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In Hahn, U., 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 the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA.

Radev, D. R. and D. Tam (2003). Single-Document and Multi-Document Summary Evaluation via Relative Utility. In *Poster Session, Proceedings of the ACM CIKM Conference*, New Orleans, LA.

Saggion, H. and G. Lapalme (2000). Concept Identification and Presentation in the Context of Technical Text Summarization. In Hahn, U., 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 the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, USA. Association for Computational Linguistics.

Sahlgren, M. (2001). Vector-Based Semantic Analysis: Representing word meanings based on random labels. In *Proceedings of Semantic Knowledge Acquisition and Categorisation Workshop at ESSLLI'01*, Helsinki, Finland.

Salton, G. (1988). *Automatic Text Processing*. Addison-Wesley Publishing Company.

Salton, G. and M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27.3-4:379–423,623–656.

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal*, 30:50–64.

Sparck-Jones, K. (1999). Automatic Summarizing: Factors and Directions. In Mani, I. and M. T. Maybury (editors), *Advances in Automatic Text Summarization*, pp. 1–13. The MIT Press.

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.

Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.

Winkel, A. and D. Radev (2002). MEADeval: An evaluation framework for extractive summarization. http://perun.si.umich.edu/clair/meadeval/.