International Journal of Computational Intelligence Systems, Vol.2, No. 3 (October, 2009), 202-218

Received: 23/08/09
Accepted: 12/09/09

EXPERIENCES WITH AND REFLECTIONS ON TEXT SUMMARIZATION TOOLS

SHUHUA LIU*
Academy of Finland and Abo Akademi University
IAMSR, Joukahaistenkatu 3-5A, 4th floor, FIN-20520 Turku, Finland
E-mail: sliu@abo.fi

Text summarization is a process of distilling the most important content from text documents. While human beings have proven to be extremely capable summarizers, computer based automatic abstracting and summarizing has proven to be extremely challenging tasks. In this paper we report our experience with applying extractive summarization techniques to process news articles, economic reports and nursing narratives. We present analysis of the effect of different summarization methods and parameters on the summarization results. We also compare the performance of the summarizers across the three different document genres. The learned lessons are discussed and the possibilities for applying the theory of Computing with Words in text summarization are elaborated.

Keywords: Text summarization, sentence extraction methods, summarization evaluation, computing with words.

1. Introduction

Text and natural language processing capabilities are of increasing importance in the information society of today. Text summarization is, among many other language technology applications, a highly anticipated function in almost all kinds of information access systems.

Key issues in automated text summarization include how to identify the most important content out of the rest of the text and how to synthesize the substance and formulate a summary text. There are in general two different approaches to text summarization: the selection-based approach and the understanding-based approach. Text summarization systems can choose to make use of shallow text features at sentence level, discourse level, or corpus level to locate the important sentences that will make-up an "extractive" summary. Such extractive methods often treat the "most important" as the "most frequent" or the "most favorably positioned" content, thus avoid any efforts on deep text understanding. They are easy to implement and generally applicable to different text genres but it is usually very hard to achieve performances that exceed a generally attained level. A lack of coherence, lack of balance and lack of cohesion can be evident in the output due to the presence of dangling references. Such readability issues may cause incorrect comprehension, even after smoothing.[1,2,3,4,5]

On the other hand, text summarization systems may also choose to be based on an understanding of text meaning and content, to imitate human summarization process. The output of such a process will be an "abstractive" summary. However, understanding-based methods require a reasonably complete and accurate formal specification of text content, which is only possible in situations when the concern for content is focused on a rather restricted subject field, when information needs can be pre-defined, and when a large enough knowledge support can be constructed in advance. The systems tend to be domain dependent and application-specific, thus will be difficult to function properly in a different context. As in the case of many other applications, the better quality of a summary is traded off for the general applicability and flexibility of the system.

The earliest studies on text summarization dated back to the late 1950's with the pioneering work of Hans Peter Luhn,[6] in which he invented a statistical sentence extraction method that calculates a significance score for each sentence based on counting word frequency. Luhn's work was followed by an early effort in the

*IAMSR, Abo Akademi University, Joukahaistenkatu 3-5A, 4th floor, FIN-20520 Turku, Finland

60's.[7,8] Very active and intensive research efforts are seen especially since the 1990s from the computational linguistics community.[3,4,5,9,10,11,12,13,14] Much progress has been made in exploring a variety of text summarization methods and techniques. A number of rather impressive text summarization systems emerged such as the MEAD system from Univ. of Michigan,[14,15] the SUMMARIST system from Univ. of Southern California,[3] and the Newsblaster from Univ. of Columbia,[13] all employ a collection of extractive summarization techniques.

In most recent years, research on summarization continues in the direction of incorporating more and more progress made in computational linguistics/natural language processing, domain specific ontology development efforts, advanced machine learning methods, as well as summary evaluation methods. In the mean time, in addition to news summarization, application research starts appearing for example in the summarization of emails, product reviews, medical dialogues,[16,17,18] plus multilingual, multimodal sources of varying types on the Web such as blogs and talk show transcriptions (the DARPA funded GALE project).

In our previous research, we have applied the sentence extraction techniques of the MEAD system in summarizing country economic reports (the IMF staff reports)[19,20] and intensive care nursing narratives.[21] In this paper we give a summary of our previous results and extend our previous work with newly added elements, which includes experiments with news text and an examination of the possibilities of applying the theory of computing with words in text summarization. The rest of the paper contains five sections. Section 2 describes the text data. Section 3 introduces summarization methods and tools that we use. Section 4 reports the experiments and results. Section 5 presents further analysis and elaboration on the potentiality of applying new theory such as computing with words in text summarization. Section 6 the conclusion.

2. Source Documents

The text collections used in our study include three different genres: (a) the IMF Staff Report, an important source of information concerning country economic development and monetary policy; (b) intensive care nursing narrative, which keeps daily account of a patient's medical situations and related treatment and care while in ICU; and (c) a news collection from a newspaper.

2.1. *The IMF staff reports*

The IMF (International Monetary Fund) staff reports are written by its mission-teams (the Fund economists) as the product of their missions. The reports are highly professionally written and always carefully reviewed through an elaborate process by relevant departments in the Fund. They are very carefully articulated and the nuances of the language as used in staff appraisal are important and could impact upon how a report was read by the Executive Board.[22] Structurally, they basically follow a very consistent format. All staff reports begin with a one-page executive summary, followed by 3-4 main sections that cover:

- General economic setting, which often contains the conclusion of the last mission and report of the economic developments since then. It gives an overview of the current economic setting and raises issues that should be addressed during the current mission and subsequently discussed in the report.
- Policy discussions, which contains information about discussions held with the state authorities regarding country economic policies, especially monetary and fiscal policy.
- Staff appraisal, which presents recommendations to the member country by the IMF, in order to achieve long-term growth and balance of payments stability.

The written patterns for the reports are also very clear. Each section usually contains a set of numbered paragraphs. Each paragraph seems to start with a one sentence "summary" of the paragraph. The rest of the sentences in the paragraph will contain supplementary information regarding the issue presented in the first sentence. The length of a paragraph tends to vary between 5 and 20 lines of text. In addition, the paragraph may have embedded tables or figures that are related information in some way. There are also footnotes and text boxes that provide additional information on some specific topic, and appendices of tables and figures. Following the main body of the report, there are written statements by various staff

involved in the process, and a Public Information Notice containing the assessment from the executive board regarding the issues addressed in the report. Examples of the reports can be found on the IMF website (http://www.imf.org).

In our study, we included five IMF staff reports, the Article IV Consultation reports from year 2004 and 2005, for China, Finland, Norway and Sweden. The reports are downloaded directly from the IMF publication database accessible at the IMF website. Pre-processing removes table of contents, appendices, executive summary, tables, text boxes, figures, formulas, footnotes and various supplements attached to each staff report. The executive summary provides a model of what the automatically produced summary would ideally contain, thus is used for summary evaluation. Table 1 gives an overview of the word count in each executive summary and in the corresponding report after preprocessing. The amount of words in the executive summaries is between 395 and 466, equivalent to a compression ratio of 5-10 % of words.[19,20]

Table 1 Length of staff reports and executive summaries

	Whole Report	Executive Summary	Compression Ratio
China	8470 words	414 words	4.9%
Finland	5211 words	419 words	8.6%
Norway	4687 words	451 words	9.6%
Sweden 04	5311 words	395 words	7.4%
Sweden 05	4713 words	466 words	9.9%

2.2. *Intensive care nursing narratives*

Nursing narratives are the written story that describes a patient's clinical situation and its progress, care plans, nursing interventions and the evaluation of the interventions, in chronological order covering a specific time frame. In intensive care units (ICUs), the nursing documentation has to be detailed and frequent due to the complex health problems the patients are suffering from, and the typically rapid changes in the patients' condition. Nurses often access, process and take down a lot of information while providing care to the patient. When a patient stays in the ward for several days, the amount of information and documentation is large.

The focus of intensive care is on the maintenance and monitoring of patient's vital functions such as respiratory rate, blood pressure, pulse and temperature, pain, excretion, and level of consciousness. The content of the daily narratives reflects these concerns. It contains nuances and messages that numeric or strictly structured data entries cannot capture. Structurally, the daily narratives are always organized into text passages following the nursing shifts: *morning shift, day shift and night shift (or the long shifts)*. Within each shift, the documentation is structured according to the concerned topics such as: *breathing, hemo-dynamics, elimination, consciousness, family contacts* (family visits or telephone discussions), and *other issues* (e.g. special treatments, skin condition and body temperature).[21]

When the patient moves out of intensive care to a bed ward, a discharge report is written, which summarizes the patient's current situation and covers the most important issues happened in the intensive care unit. The discharge reports usually begin with background information including e.g. reason for admission to the ICU and a short description about the patient's medical history. The intensive care period is summarized mostly with respect to such topics: *breathing, hemodynamics, consciousness and mood, examinations, infection situation, excretion, skin condition and care, pain treatment, special treatments, family contacts, personal belongings, and other advices,* which are a little different from the nursing narratives alone.[21]

Data used in our research were collected from an ICU in a Finnish hospital. The gathered data included the free text parts of the daily nursing documentation over the whole intensive care period, plus the discharge reports written at the discharge to the bed ward. The language used in all the documents is Finnish. All patient identification information was removed. The complete data set includes patients who stayed in the ICU for at least five days during years 2005-2006. In this study we only included data for two groups of patients with the same ICD code of the primary disease: patient group with brain blood vessel diseases (44 patients, average length of stay in the ICU around 13 days) and patient group with intracranial injuries (40 patients, average length of stay in the ICU around 10 days). These two patient groups have the longest stay at ICU. In addition,

word frequency count shows that the daily narratives for these two patient groups are rather similar in terms of most frequently used words. On average these daily narratives consist of about 2500 words.[21]

In contrast to the highly professionally written economic report, nursing narratives are more of the nature of "working notes" instead of well-polished economic reports. Each nurse uses his/her own writing and wording style; abbreviations (standard or arbitrary), slang, misspellings are common. Sentences may not follow most of the grammatical rules, and so on.

2.3. *News collection: Helsinki Sanomat English edition*

News as a type of text discourse is thoroughly studied in Ref 23. News can be viewed as a kind of narrative, just like stories. But news also differs clearly from the narratives of our everyday conversations or in novels, or in our case, the daily nursing narratives. News in the press is a specific kind of mass media discourse that delivers wide ranges and large amounts of social and political knowledge and beliefs.[23] Like staff reports, news is written by trained writers, follows strict grammatical and linguistic rules. A news collection usually contains articles writing about a wide range of topics while the topics for staff reports concentrate solely on economic developments and monetary policy. On the other hand, news is much shorter and simpler in formats comparing to staff reports and nursing narratives.

Nonetheless, news is not totally free flowing text as they seem to be. News typically tends to follow certain (hidden vs explicit in nursing narratives) hierarchical schemata in their structures that keep topics organized. They often consist of conventional content categories such as *Headline, Lead, Main Events, Context and History* (together forming the Background category), plus *Verbal Reactions, and Comments.* The major events and statements are often expressed in the headline and the lead paragraph. A good Headline tends to sum up the main information of the text and signal what are important for the news source. The Lead paragraph and the subsequent content in the Main Event category, together with a brief History and some general Context, provide further details of topics highlighted in the headline. The Comments category is often expressed discontinuously, as installments, throughout the text. Of each category the most important information is expressed following a top-down presentation strategy.[23]

The data collection used in this study contains the text parts of daily news articles in the English edition of *Helsinki Sanomat* (the major daily newspaper of Finland), during one-month period from November 21 to December 21, 2007. All together 185 articles are collected from the website of *Helsinki Sanomat* (http://www.hs.fi/english/archive/). The length of the articles ranges from 142 words to 1388 words, with the majorities between 200 to 700 words. The one-sentence headlines count from minimum of 4 words to maximum of about 15 words, with the majorities between 7 to 10 words.

3. Methods and Tools

3.1. *Extractive summarization methods*

Most of the practical text summarization systems today are extraction-based. A summary is created based on sentence extraction and then the sequential re-organization of the extracted sentences, without re-written, but sometimes may be smoothed to certain degree (for example, by taking source sentences preceding the key sentences containing anaphoric references; sometimes complete paragraphs instead of sentences).

Different extractive methods make use of different text features to represent the text content. These features may include: *thematic features* based on term frequency statistics, *location features* such as position in the text, position in the paragraph or the particular section, *background features* such as terms from title and headings in the text, *cue words and phrases* such as in-text summary cues "in summary", "our investigation", *bonus and stigma terms* such as "significant", "impossible", and so on. Such features can be analyzed individually or combined selectively to form a function that is used to identify important words and significant sentences in the text. Sentence scores are computed based on the evaluation of word importance. Sentences that are concentrates of high score words (significant words) are often the target sentences to be extracted.

Position based method and the Lead method: Position based methods weight the words and sentences in the different parts of the document differently. Often sentences under headings, sentences near the beginning and end of a document or a paragraph are given extra weights than those in the middle; sometimes they are simply selected automatically, e.g. in the Lead system, sentences are added to the summary based on their position in the source articles alone. Sometimes the location of a sentence in text is used to adjust the normal sentence score.

Random method: simply put together sentences randomly selected from a document as a summary based on a random value between 0-1 that is assigned to each sentence. Random methods and Lead method are often used as baseline summarizers.

Query method: Given a query (e.g. a set of single words, phrases or short passages), query-based method will calculate the similarity between the query and the sentences in the documents. Sentences with highest similarity values will be selected to compose a summary.

Machine learning based method: a corpus-based approach, when both the collection of original document as well as the corresponding collection of model summaries (especially extractive summaries) are available, empirical rules for extracting text segments from the documents can be learned using text classification methods. The problem of summarization becomes a problem of two-class classification problem.

3.2. *Summary evaluation methods*

Summary evaluation is as challenging a task as automated summarization itself, especially because different people can judge the quality of a summary differently, one summary can be of different value to different tasks, and there is no golden metrics for measuring the quality of a summary. Evaluation methods can be divided into extrinsic and intrinsic: extrinsic evaluation being task-based evaluation through investigating how a summary affects the completion of some task, and intrinsic evaluation being content based examination by comparing a summary to a target (often called a reference summary or a model summary). Examples of extrinsic evaluation work can be found in Ref. 24 and Ref. 25. So far in our study we have focused on intrinsic evaluation.

At the core of intrinsic evaluation is the choosing of an appropriate content measurement unit, as well as a significance and similarity measurement for matching the content units.[26] Different evaluation methods address these issues in different ways. Semantic similarity analysis aims to measure content similarity in terms of meaning while lexical similarity measures only considers the words used without concern for the actual meaning.

Lexical similarity is often measured by the word or small n-gram overlapping between two summary texts. Examples include simple cosine similarity with a binary count of word overlap, cosine similarity with tf*idf weighted word overlap,[27] measurement of the longest-common subsequence[15] bigram and n-gram[28] overlap measurements such as BLEU.[29] The most recent development of n-gram based evaluation is the ROUGE method, which was found to produce evaluation rankings that correlate reasonably with human rankings.[30]

Measuring and comparing content at word and sentence level granularity was found not precise enough and unsatisfactory.[26] The BE (Basic Elements) method[26] and the Pyramid method,[31,32] acknowledge the fact that there is no single best model summary and they offer ways to address the content variation across multiple model summaries of the same source text. The BE method, as a further development of the ROUGE method, tries to evaluate a summary by the matching of the Basic Elements it contains, which are extracted from the parse trees of the text, to a set of Basic Elements extracted from a number of reference summaries and then integrated. Matched BEs' scores are calculated based on the number of reference summaries they appear in. By adding the scores of each BE in the summary an overall score can be assigned to the summary.[26]

The Pyramid method offers another solution for semantic similarity analysis. But it depends on manually identify and annotate text units of different size that are considered to contain only "important content" of the summary texts. Summaries are compared manually

based on these "semantic units", which are usually approximately clause-length chunks of continuous or discontinuous sequences words or phrases shared by the reference summaries.[31,32] Thus, the content is identified based on "shared meaning" instead of shared words or word strings (n-grams). Pyramid evaluation was found capable of differentiating system summary from human summaries.[31,32] However, creating the pyramid and evaluating the peer summaries are very demanding tasks as they require heavy manual work. By contrast, the BE tries to automate the identification of semantic units with the help of syntactic parsing. It overcomes the subjectivity/variability problems resulting from manually identifying content units.

3.3. *The MEAD system*

The research tool we used in our study is based on the MEAD system developed at University of Michigan.[14,15] MEAD is a public domain multi-document summarization system. It contains a number of sentence extraction methods for summarization, such as position-based, query-based, centroid method, plus two baseline methods: the lead-based method and the random method.[15]

The Centroid method is the key summarization method of MEAD. Given a document or a collection of documents to be summarized, a cluster is formulated and all sentences in the cluster are represented using tf*idf vector space model. A pseudo sentence is then calculated, which is the average of all the sentences in the cluster. This pseudo sentence is regarded as the centroid of the document cluster, which is considered to be the best representation of the entire document cluster. The significance of each sentence is then determined by calculating how similar each sentence is to the centroid.

The query method is implemented in MEAD in two ways. The query-sentence similarity can be calculated using a simple cosine similarity measure. If a sentence contains any of the words in the query, the sentence will get a score adjusted according to the length of the sentence. A short sentence will get a higher score and a longer sentence will get a lower score for containing the same word. A word that occurs many times in the same sentence is counted the same as if it only occurs once. An alternative scoring method is to weight the simple-cosine value with the IDF value of the word. If the query contains words with high IDF values, the sentences containing these words will get high scores, and are more likely to be chosen to the summary.

The overall architecture of MEAD system consists of five types of processing functions: Preprocessing, Features Scripts, Classifiers, Re-rankers and Evaluators.[14] *Prerocessing* takes as input the documents to be summarized in text or HTML format, identifies sentence boundaries and transforms them into an XML representation of the original documents. Then, a *set of features* will be extracted for each sentence to support the application of different summarization methods such as position-based, centroid based or query-based sentence extraction methods. Following feature calculation, a *classifier* will compute a composite score for each sentence based on weighted combination of the sentence features in a way specified in the classifier, which can potentially refer to any features that a sentence has. After the classifier, each sentence has been assigned a significance score. A *re-ranker* selects the sentences to compose the summary by considering cross-sentence similarities or possible dependencies. Sentences are ordered by score from highest to lowest then iteratively decide whether to be added to the summary or not. At each step, if the quota of words or sentences has not been filled, and the sentence is not too similar to any higher scoring sentence already selected for the summary, the sentence in question is added to the summary. Among the re-ranking mechanisms in the software package, Cosine re-ranker simply discards sentences above a certain similarity threshold; the MMR re-ranker on the other hand adjusts sentence scores so that similar sentences end up receiving a lower score. Finally, a summary is formed by chaining selected sentences together in order of their appearance in the original text.[19]

The MEAD system includes a number of lexical similarity based summary evaluating instruments for evaluating summaries in pairs. Most of the algorithms use relatively similar approaches to value the word-overlap, and the results from each algorithm seem to correlate strongly with the results from the other algorithms.

3.4. *MEAD-GUI: a graphical user interface for MEAD*

Using MEAD requires in-depth technical knowledge. To make it convenient to use, we constructed a graphical user interface MEAD-GUI that would add the needed support for using MEAD easily and for carrying out extensive experiments.[19,20]

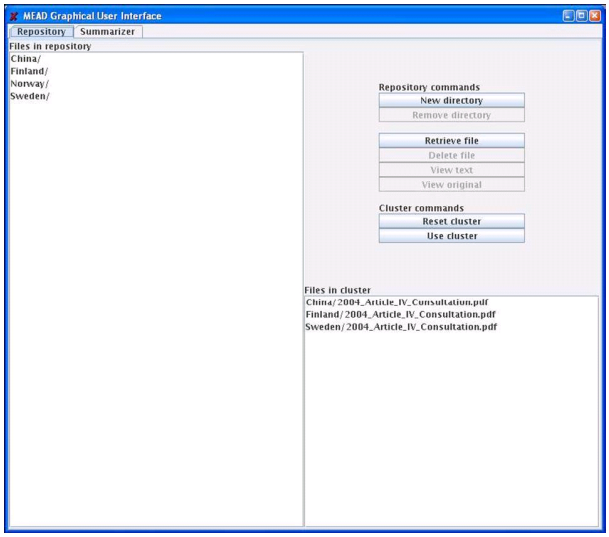

Fig. 1. MEAD-GUI repository window

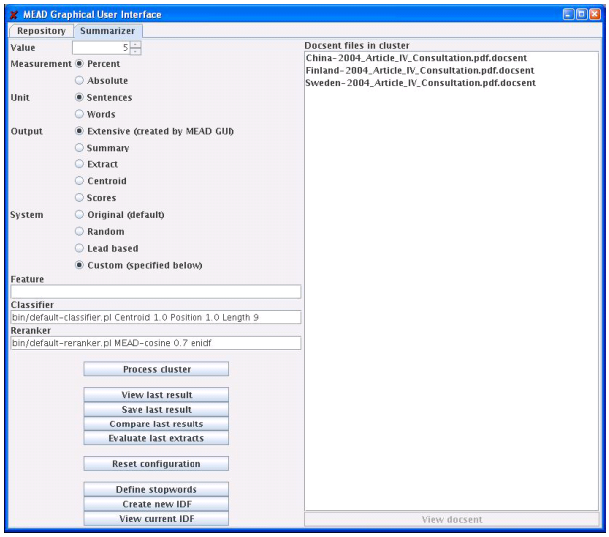

Fig. 2. MEAD-GUI summarizer window

The MEAD GUI supports accessing online documents, preprocessing common document types in HTML, PDF, and Microsoft Word format, provides a flexible local document repository and tools for post-processing and displaying the results of the summarization. Last but not least, it provides support to users in configuring and applying MEAD summarization methods as well as summary evaluation methods. With the help of MEAD-GUI, users are no longer required to manually perform each step in MEAD's summarization process. Tasks are performed automatically to an as wide extent as possible. Optimally, the user should only have to type in a web address and be able to easily view a summary of the document.[19]

As shown in Figure 1 and Figure 2, the MEAD-GUI graphical user interface is divided into two different window frames: repository and summarizer. The repository window frame provides an abstraction of MEAD-GUI's document repository and access to tools used to create the repository easily. The summarizer window frame provides the necessary tools for performing summarization and displaying the results of the summarization to the user.

4. Results and Experiences from Summarizing Economic Reports, Nursing Narratives and News using Extractive Techniques

4.1. *Summarizing the IMF staff reports*

This set of experiments involves only five reports, which gives us the possibility to manually examine the results against the original reports and the model summaries closely. The summaries produced by the system are also automatically evaluated against the staff-written executive summaries using the word-overlapping content similarity metrics included in MEAD. A number of different summarization schemes are applied. The detailed accounts of the various experiments are reported in a number of our previous publications.[19,20] Here we give a summary of the experiments and results.

4.1.1. Linear combination of Centroid and Position, with Length cutoff

In this summarization scheme, the classifier weights the centroid, position and length features as equally

important. The re-ranker calculates cosine similarity and uses a cut-off value of 0.7 to decide if a sentence is too similar to sentences already in the summary. This scheme is repeated with two compression rates: 5% and 10% of sentences.

Classifier: bin/default-classifier.pl
Centroid 1.0 Position 1.0 Length 9.0
Reranker: bin/default-reranker.pl MEAD-cosine 0.7 enidf
Compression: 5% (and 10%) of sentences

The result shows that the system tends to pick up sentences that are longer, appear at the very beginning of the report and the earlier parts of the report. It also tends to narrow down the extracted content to only a few issues (especially when with a 5% sentence compression rate), thus missing out many other issues discussed in the report. The summarizer seems to favor redundancy over topic coverage, as we can see much redundancy in the extracted sentences.

4.1.2. *Centroid method; MMR re-ranker*

The Length cut off seems not important as the Centroid method already makes it so that it favors long sentences. Since in the staff reports, the sections that appear at a later stage (e.g. policy discussion, staff appraisal) is as important as, if not more than, the sections that appear earlier, and since the Position feature value too much only the sentences at the very beginning of the whole report, we took away the Position feature also. To balance the topic coverage and reduce redundancy, we replaced the MEAD default re-ranker with the MMR re-ranker with Lambda as 0.2. This summarization scheme is again repeated with two compression rates: 5% and 10% of sentences.

Classifier : bin/default-classifier.pl Centroid 1.0
Reranker : bin/default-reranker.pl MMR 0.2 enidf
Compression: 5% (and 10%) of sentences

The result is significantly different from the first set of experiments. The 5% summary contains significantly more sentences from the policy discussion and staff appraisal than in the last experiment, while also has much wider topic coverage. The MMR re-ranker seems working very well. There are much less redundancy than in the last experiment. However, the complete removing of Position feature results in the missing-out of some important sentences that often appear at the beginning of different sections.

4.1.3. *Centroid + Position; MMR re-ranker*

Classifier : bin/default-classifier.pl Centroid 1.0 Position 1.0
Reranker : bin/default-reranker.pl MMR 0.2 enidf
Compression: 5% (and 10%) of sentences

Comparing the Centroid method with "Centroid + Position" method, we found considerable overlapping of sentences from Section I and Section II of the staff reports. However, the outputs from Section III are completely different sentences in these two experiments. This indicates that Centroid + Position method causes tradeoffs between sentences that appear at the beginning of the section and sentences that are selected in the Centroid method.

4.1.4. *Paragraph Lead-based*

The IMF staff reports are noticeably divided into independent sections, and each section is composed of a set of numbered paragraphs. We consider the evident content structuring formality of the reports could be utilized to benefit the summarization output. In fact it seems that simply extracting all the introductory sentences of the important paragraphs would be able to produce a summary of reasonable quality. Such an approach is explored in this experiment by creating paragraph level Position feature, which sets the first sentence in each paragraph to receive a score of 1; the second sentence a slightly lower score, and so on. When a new paragraph begins, the score is reset to 1.[19] The summarization scheme is shown below:

Classifier : bin/default-classifier.pl
Centroid 1.0 Length 9.0 Position 1.0 and 5.0
Reranker : Default reranker
Compression : 10% sentences

Such a scheme will allow sentences considered important by the Centroid feature also have an opportunity of being included in the summary, even if they are not the first sentence in a paragraph. The experiments shows that, when Centroid and Position features are weighted equally, it does not ensure at gall that only the first sentence of each paragraph would be included in the summary. The Centroid score was in a

considerable amount of cases significantly higher for sentences later in the paragraph.

The reason for this is likely that the first sentence in a paragraph is usually a bit shorter, which tends to result in a lower Centroid score. When the weight for Position feature is set to be five times the weight of Centroid feature, and Length with a length cutoff of 12 words, the first sentence is nearly always selected. The Centroid feature, however, is useful for prioritizing sentences selected by the Position feature. The summaries appear to have included sentences throughout the entire document, and the number of words in the summaries is significantly lower than in the other summaries. The result hinted us that favoring long sentences is not always a good approach; it may in some cases lead to a long summary that is not of very high quality.

4.1.5. *Random method*

The sentences picked up by Random method are usually different from those selected by any non-random method as they tend to be irregular and unpredictable as they are meant to be. Comparing to the results from Centroid + Position method, the latter are indeed more stable and have a better coverage of topics. However, our evaluation result does not give direct indication that the Random method is inferior to other methods. In fact, in some cases the results from applying the Random method were only perhaps slightly inferior to the other methods, and in some cases, it is even slightly better.

4.1.6. *The IDF Dictionary*

The Centroid method is highly dependent on the IDF dictionary. Will the use of different IDF dictionaries have a significant influence on the summarization result? The purpose of this experiment was to examine the performance variation with different IDF dictionaries. To do that we compared the default MEAD IDF dictionary with three other customized IDF dictionaries:

Mead IDF: included in the MEAD package
Small IDF: created from 31 IMF country reports
IMF IDF: reated from 69 IMF country reports
Large IDF: created from 146 economic reports and articles from different sources (IMF, ECB, Bank of Finland)

Overall, the results indicate little difference between the summary outputs based on different IDF dictionaries, at a compression rate of 10% sentences. At a compression rate of 5% sentences, sometimes the MEAD IDF dictionary shows a bit better performance, sometimes the customized IDF show a bit better performance, but these are not truly significant changes. The Mead IDF dictionary generally stays as a good performer thus proves to be a good default and can be used directly in summarizing the IMF staff reports. Generally, spending a large effort on collecting large amount of documents for building a domain specific IDF database does not help much. A small IDF cluster can actually achieve similar results with a proper compression rate and summarization method.

In the mean time, it is useful to notice the effect of different IDF dictionaries differs on different documents. IDF variations seem to cause system outputs become sensitive to small changes in writing styles reflected in wording choice. If a document uses many uncommon or unknown words comparing to the IDF cluster, then the summary result will be poor.[19,20]

4.1.7. *Re-rankers*

Three rerankers are offered in MEAD: the Identity-reranker, Default-reranker and MMR-reranker are tested. Using identity-reranker is equivalent to not using a reranker; it does not modify sentence scores received from the features and classifier. Default-reranker is the MEAD Cosine reranker with a similarity threshold of 0.7. The MMR-reranker is based on maximal marginal relevance principle, applied with a similarity threshold of 0.2, using MEAD-cosine as the similarity function.

Our experiments show that, for the IMF staff reports, MMR re-ranker seems to have a significant advantage over the other two re-rankers. It does a better job in redundancy control of the summary outputs. Looking at the results by individual reports, we can notice obvious difference of the re-ranking effect among some of the reports. Depending on how the executive summaries and the corresponding reports are worded and overlapped, the differences in the effect of the different re-rankers may be minor or significant.

4.2. *Summarizing nursing narratives using MEAD*

In this set of experiments we want to extract discharge information from daily nursing narratives. As we know, the summarization methods and the MEAD system are originally designed for processing texts in English. To handle the Finnish text, we spent some time fine tune the system. Due to the rather free-style nature of writing and wording of the nursing narratives, the sentence segmentation tool in MEAD cannot be used directly. A more suitable tool was created to identify sentence boundaries according to both the sentence ending mark and line breaks, while most of the irregular abbreviations were also dealt with. Then time stamps are added to each sentence; all the text segments in the original narratives are rearranged as in time order.

Following system modification, target summaries are created. As the original discharge reports contain information that is drawn from other sources than nursing narratives, they cannot be used as the target summaries as such. To serve our expectations for the summaries, the target summaries are prepared by removing four parts of the discharge summaries: background information, medical examinations (a list of taken x-rays and other kinds of scans), infection situation (a list of CRP values and antibiotics given) and personal belongings (a sentence about if the patient has some personal belongings and if they have been given to family members, by whom and when), which are not based on the daily narratives. Consequently, the shortened discharge summaries focus on six mostly used topics, and they are used as the target summaries for evaluating the quality of system summaries.[21]

To determine the compression rate (the length of the summaries), the amount of words in the original daily narratives and in the shortened discharge summaries are counted and it was found out that the length of the discharge summaries is on average about 10% of the length of the daily report. So a compression rate of 10% words is basically appropriate. To make use of the MMR re-ranker, however, the system requires a compression rate expressed in terms of sentences. So, a compression rate of 10% sentences is used when MMR rereanker is applied.[21]

Summarization scheme: among the summarization methods available in MEAD, we consider the query-based method would be the best to serve the purpose of extracting discharge information. In the meantime, Centroid method could also play a helpful role in identifying something that is always important to a specific patient. In addition, as discharge information is a bit more about the most recent status of the patient than the developments over time, text segments from the few days before the discharge (usually at the end of the narratives) can be regarded as more important. Thus, the summarization schemes we explored are the query-based method modified by centroid and position methods. Random method (10% words) is used as a baseline. We tested a number of different combinations of the Query, Centroid, Position and Length for the classifier, and two re-rankers, the MEAD cosine reranker and MMR reranker, on a few of the files to be summarized, and evaluated the results using word overlap measures. The difference in performance is not always consistent, but the following two schemes seem to be slightly better performers than others. They are applied in the subsequent experiments on the selected data set.

- Centroid 1.0 Length 6.0 Position 2.0 Query 4.0, default reranker, 10% words;
- Centroid 1.0 Length 6.0 Position 2.0 Query 4.0, MMR reranker, 10% sentences;

Next, query words for each topic category were selected in such a way that they are frequently used words in that category, but not too general ones - i.e. those words that were found to best represent the topic class. For each topic, we counted the number of words used in that topic, and compared it to the total number of words in all the topics. Based on these percentages, a total of 75 query words were divided between the categories, and the number of query words in each category was the following: consciousness 26 words, breathing 17, hemodynamics 16, family contacts 7, diuresis 6, and other issues 3 words.[21]

In total, 252 summaries are generated and evaluated against the corresponding shortened discharge reports one by one, using the evaluation metrics in MEAD. The results show that, query based method with MMR re-ranker is the best performer comparing to Random method and query method with default re-ranker.

However, the performance difference between MMR and Default is rather minor. By contrast, MMR over performs the Random baseline significantly. On the other hand, although combinations of query, position and centroid method show better performance than the Random method, the level of similarity between all the system-generated summaries and the corresponding reference summaries is really rather low, e.g. comparing to the similarity level achieved in summarizing IMF staff reports. One reason may be that the system summaries are "extracts" while human written discharge summaries are in fact "abstracts". Another reason would be that, unlike in the case of IMF staff reports, where both the original report and its executive summary are written by same authors, nursing narratives and the discharge reports may be written by different people, thus it is very possible that the lexical differences between the discharge reports and the system summaries are rather outstanding. The differences between the two types of summaries are inevitably bigger than in the case with formally and grammatically written texts. This experiment generated many questions that will be explored in our next step work.

4.3. *News summarization*

In this experiment, we would like to generate new headlines for the news articles. The experiment is very straightforward. One sentence is extracted from each article and then compared with the original headlines. Keeping in mind that extractive methods was first designed and tested on news articles, this experiment can hopefully provide an upper bound performance level for comparing with results from summarizing the other two types of documents. Four summarization schemes are applied:

- Centroid 1.0, MMR reranker, 1 sentence;
- Centroid 1.0 Position 1.0, MMR reranker, 1 sentence;
- Lead, 1 sentence;
- Random, 1 sentence.

There aren't much surprising results from this experiment. After a few runs of the experiments, it quickly becomes very clear that the Centroid method, Centroid + Position method and Lead method all return the same result, which is the first sentence of the articles. As it happens, the Lead sentences of all the news articles have a good length when meant to be very informative about the news event. Since Centroid method tends to pick up longer sentences, the Lead sentences end up becoming its favourites also. The Centroid + Position method only reinforces the significance score of the Lead sentence. The lexical similarity analysis shows that, the Lead sentence clearly outperforms the sentences selected by Random method. When comparing with the performance level achieved in summarizing IMF staff reports, we can notice that although the system easily sets up an upper bound for summary performance when dealing with news articles, the performance on IMF staff reports is at a rather close level too.

5. Computing with Words for Text Summarization

While selection based methods is better at quickly identifying text segments carrying "important" content, it is the understanding-based methods that would be better at synthesizing the selected information. The theory of fuzzy logic based "computing with words (CW)" offers mathematical tools to formally represent and reason about perceptive information, which are delivered in natural language text by imprecisely defined terms, concepts classes and chains of thinking and reasoning. It thus provides relevant methods for understand-based summarization systems.

Until recently, the application of fuzzy logic in natural language understanding has been discussed only sparsely and scattered in the literature of soft computing and computational linguistics, although the theoretical foundation has been laid out in several articles by Prof. Lotfi Zadeh already decades ago. The term "Computing with Words" (CW) was coined only more recently in mid 1990s and it indicates a relatively new emphasis in the development of fuzzy theory, to answer the needs for better methods for representing and reasoning with perceptive information.[33,34] In this section we examine the possibilities and challenges in applying the concepts and methods of CW in text summarization.

5.1. *Cognitive models of reading comprehension & summarization*

The theoretical foundation for understanding based text summarization systems is found in the cognitive models of reading comprehension. Among the various theories, the microstructure-macrostructure model proposed by Kintsch and van Dijk[35,36] is perhaps the most influential.

The model takes as its input a list of propositions that represent the meaning of a text segment. The output is the semantic structure of the text at both micro and macro levels represented in the form of a coherence graph. Such a structure is believed will enable the full meaning of the text be condensed into its gist. A microstructure refers to the semantic structure of sentences. It reflects the individual propositions and their close neighboring relations. A macrostructure presents the same facts as the whole of microstructures, but describes them from a global point of view. A coherence graph contains a set of ordered and connected propositions. The order of connection is determined particularly by the "referential relations" between the propositions in the form of argument overlapping.[35,36]

Microstructure and macrostructure are related by a set of semantic mapping rules called "macro-rules" such as detail-deletion rule, irrelevance-deletion rule, rule of generalization and rule of construction,[35,36] which are applied in "macro-operations" that derive macrostructures from microstructures. Macro-operations are controlled by a "schema", which is a formalized representation of the reader's goal and it helps to determine the relevance/irrelevance and importance/unimportance of propositions, and thus which part of the text will form its gist. The controlling schema can be determined according to text genre or derived from query description.

Knowledge is indispensable in effective reading and comprehension. A reader's knowledge determines to a large extent the meaning that he or she derives from a text. Knowledge that is important in reading comprehension can be grouped as four types: linguistic knowledge (phonetics, morphological, syntax, semantics, pragmatics knowledge, genre knowledge), general world knowledge (commonsense knowledge as well as socially known properties of some social and natural world), specialized domain knowledge and context/situational knowledge (task context, communication context, location context). World knowledge is of many different types, which do not always apply in discourse processing in the same way but are instead personally and contextually variable.

In reading process, "a reader often also tries to produce a new text that satisfies the pragmatic conditions of a particular task context or the requirements of an effective communication context". "The new text will contain information not only remembered from the original text, but also re-constructively added explanations or comments based on his knowledge and experience." Kintsch and van Dijk[35,36] noted four types of text production activities associated with text reading: (i) Lexical and structural transformations such as lexical substitution, proposition reordering, explication of coherence relations among propositions, and perspective changes! (ii) Through the memory traces, particular contents will be retrieved so that they become part of the new text. (iii) When micro and macro information is no longer retrievable, the reader will try to reconstruct the information by applying rules of inference to the information that is still available. (iv) A new text may also be some meta-comments on the structure or content of the text such as giving comments, opinions, expressing attitudes.

5.2. *Understanding based summarization*

Summarization is sometimes the purpose and sometimes a byproduct of a reading and comprehension process. Understanding based summarization means in essence three things: (i) text understanding; (ii) finding out what is important; (iii) rewrite a number of important messages to form a coherent text, i.e. text production. Key to the computational implementation of understanding-based summarization is the capability to (i) correctly interpret the syntax and semantics of word and sentence (i.e. to relate linguistic forms to meaning, to map natural language expressions onto a formal semantic representation), to (ii) derive the most important information by appropriately operating/reasoning on the formally described content; and to (iii) map the newly derived formal representation of content into natural language expressions. None of these steps are trivial tasks. In fact, every of the steps involve a handful of very challenging issues. Research in computational linguistics addresses the first and third task and has proposed and developed impressive solutions. The theory of computing with words will be a relevant approach in dealing with the second task.

5.3. *Computing with Words (CW)*

Among the many different ideas about how meaning and knowledge can be represented in human mind and machine, logics is more easily received by formal

treatment than frames and semantic networks, and have been playing a significant role in language processing. Natural language text contains rich predicate-argument assertions that can be generally treated as propositions that lend themselves to logical representation and operations.

However, it is an acknowledged fact that there is a sharp contrast and mismatch between the formality and precision of classical logic and the flexibility and variation of natural languages. Natural language text abounds with perceptive information that is intrinsically imprecise and fuzzy. Although widely applied in NLP systems, classic logics such as FOPC (first order predicate calculus) have significant limitations in terms of expressing uncertain or imprecise information and knowledge. It has only rather limited power in expressing qualitative quantifiers, modifiers, or propositional attitudes (associated with words like believe, wants, think, dream, knows, should, and so on). Fuzzy logic, on the other hand, provides the necessary means to make qualitative values more precise by introducing the possibility of representing and operating on various quantifiers and predicate modifiers, which help to maintain close ties to natural language.[33,34] Thus, in principle, the imprecision and vagueness of terms, concepts and meaning in natural language text could largely be treated in a quantitative way using the method of Computing with Words.

One basic notion of CW is that imprecise, uncertain and ambiguous information needs to be precisiated first. Precise meaning can be assigned to a proposition p drawn from a natural language by translating it into a generalized constraint in the form p ! X isr R, where X is a constrained variable and R is a constraining relation that is implied in p and the text context; r is an "indexing variable" whose value intensifies the way in which R constraints X. The principal types of generalized constraints include e.g. equality (r is =), possibilistic (r is blank), probabilistic (r is p), random set (r is rs), usuality (r is u), fuzzy graph (r is fg), etc.[33,34] A natural language proposition p in a generalized constraint form is called precisiated natural language (PNL).

This means that, with PNL, the meaning of a lexically imprecise proposition is represented as an elastic constraint on a variable or a collection of elastic constraints on a number of variables. The translation of p into generalized constraints is a process of making the implicit constraints and variables in p explicit.

An "explanatory database" needs to be constructed first, which will help to "identify" and "explicate" the constrained variable and constraining relation in different types of propositions based on test-score semantics.[33] Once the data for reasoning is ready, the reasoning process is treated as propagation of generalized constraints. Rules of deduction are equated to rules that govern the propagation of generalized constraints. Deduction rules drawn from various fields and various modalities of generalized constraints reside in a deduction database. Generalized constraints in conclusions need to be retranslated into propositions expressed in a natural language.

5.4. *Computing with Words using fuzzy logic: possibilities for application in text summarization*

The Kintsch-Dijk model recognize that a macrostructure could capture the most essential information denoted by a sequence of propositions and thus to represent the gist of a text segment. Thus, a summary can be generated through the deriving of a marcostructure from microstructures by deleting details, deleting irrelevant proposition, generalizing multiple propositions and constructing new propositions.[35,36]

This process goes on recursively on sequences of micro propositions as long as constraints on the rules are satisfied. This in certain sense resembles the process of constraint propagation in CW. Both micro and macro propositions denote both hard facts and soft perceptions that could be formalized as generalized constraints. Natural language propositions can also be transformed into Generalized Constraints with the support of Explanatory Database. Ambiguity resolving mechanisms are the tools for the construction of the explanatory database in the CW framework. Constraint propagation will derive conclusions from facts presented by the collection of propositions. This process is driven by query or topic description propositions. And the process of inducting macro propositions from micro propositions can be realized as propagation of

generalized constraints, as well as abstraction and generalization with protoforms.

The challenges are great, however, when applying the theories in practice. The Kintsch-Dijk model presents a useful framework for us to understand the mental processes taking place in reading comprehension. However, one evident limitation in the model is that the formulation of discourse meaning structure is solely based on a coherent referential relation that does not necessarily equals to meaning relations. A macrostructure must be implied by the explicit microstructure from which it is derived, while the explicit microstructures often only supply incomplete information on the meaning of the content. For human readers, natural language discourse may be connected even if the propositions expressed by the discourse are not directly connected through referential relations because a reader is often able to provide the missing links of a sequence on the basis of their general world knowledge or contextual knowledge as well as inferences. For computing systems this is the hardest challenge. There is no guarantee that the topmost propositions resulted from referential relationship analysis are truly the most important content and information.

Next, to apply the concepts and methods of CW in text summarization presupposes an accurate natural language processing function that can transform free text into natural language propositions and then to the form of generalized constraints. This assumes that the constrained variables and constraining relations can be reliably defined, which presents a serious obstacle. To make it harder, the specification of generalized constraints requires the appropriate granulation of attribute values and calibration of lexical constituents (adverbs, adjectives) of propositions available. In reality however, standard calibration of value terms in many cases does not exist and it is very common that different people calibrate the same value term differently. So even though the operations on membership values is consistent and can be done in a systematic manner according to standard mathematical functions, the initial definition of value terms may be inappropriate. Finding effective ways for constructing the explanatory database for natural language texts in sufficiently specific domains and facilitating the calibration of lexical constituents of propositions will be the key to the implementation of the working systems.

Then, although constraint propagation will make precise reasoning about the presented generalized constraints, there are chances that the derived conclusion may be invalid because (i) the input to the reasoning process (i.e. the generalized constraints) were wrongly identified; (ii) the fact that not every proposition in natural language is precisiable. Incorrect input combined with a formal reasoning process result in false conclusions. To deal with perceptive information, the theory of computing with words using PNL introduced notations of sufficient generality and expressiveness, while will inevitably become complicated and computationally heavy. Such complexity does not necessarily guarantee better results in summarization due to e.g. the lack of other elements (such as sufficient world knowledge).

6. Discussion and Summary

In this paper we reported our experience with applying sentences extraction tools to summarize country economic reports, nursing narratives and news articles. Although the results are still primitive and indicative instead of conclusive, the extractive techniques seem to perform reasonably well when handling the *Helsinki-Sanomat* English news text and the IMF Staff reports, which are all written by highly professional writers. By contrast, results from the use of simple extractive techniques on nursing narratives are rather disappointing. For news articles, simply applying Lead, Centroid or Position method easily achieve good result. For staff reports, it appears that the Paragrah Lead-based with a higher weight for the Position feature (5.0) was a better approach for summarization than others. It is worthwhile to stress the fact that the Centroid method strongly favors long sentences, which is not necessarily the best strategy. It was somehow discouraging in the case of summarizing staff reports that the Centroid feature could not succeed in selecting the first sentence from a paragraph even when helped by the Position 1.0 feature. Comparing to each news article that are usually simpler in format and much shorter in length, and that tends to focus on one event, one topic or has one centroid, longer documents such as the IMF staff reports and nursing narratives usually contain what may be called multi-themes, multi-topics or multi-centroid that

are equally important. This introduces much creativity in applying the summarization methods as well as the importance of proper preprocessing.

Both the summarization methods and summary evaluation methods applied here are largely built upon language models for IR tools that are usually based on "words" instead of "meanings". Such language models loses a lot of syntactic and semantic information about the original document but often maintains enough information to decide if two documents are on the similar topics, thus help in document retrieval. Although they alone are not sufficient tools for performing more intellectual and cognitively demanding tasks such as language understanding, they prove to be very helpful in quickly identifying relevant/important sentences or passages through quick statistical analysis of text surface features. What can be noted is that, as long as it is the extraction strategy that is employed, the nature of the resulted summary remains primitive to many tasks and purpose no matter what further improvements over the extraction methods are added. All the effects of the improvement seem to remain at an insignificant level. In this study, summary evaluation is still limited to intrinsic evaluation using simple lexical similarity measures. During the study, the question of whether such word overlapping based similarity evaluation can be considered reliable at all arose. Automatic evaluation of summaries certain has merits in indicating the relative performance level. But such evaluation approaches merely provide an indication rather than a definite answer regarding the similarity of text. Our next step work will include the ROUGE evaluation, as well as summary evaluations in terms of semantic similarity.

The Kintsch-Dijk model regards gist formation and summary production as an integral part of the cognitive process of reading comprehension and text production. It suggests one way for formulating what is important in a text as a mapping from its micro semantic structures to macro semantic structures. It also pointed out that a summarizer not only reads and interprets text meaning and content, but also actively reconstructs meaning according to his prior knowledge or with respect to his information gaps. The process of meaning construction is often the reasoning or inference that draws upon prior knowledge to fill the gaps of incoming information and that adapts the new knowledge to what is already in memory.

Zadeh's pioneer work on Computing with Words laid the theoretical foundations for representing and reasoning with information in all format, numerical or perceptive, precise or ambiguous. In relation to textual information processing, the theory of CW suggests a very different approach from traditional approaches in computational linguistics. It seems to offer a better framework and more suitable methodology for the representation of and reasoning with meaning, knowledge and strength of belief expressed in natural language than is possible within the framework of classical logic. The proposed constraint propagation reasoning mechanism can help derive new constraint statements from groups of constraint statements. An advantageous application for it would be to derive from natural language text answers to natural language queries. Text summarization could benefit from such an approach when the required summary can be created as the collected answer to a collection of queries about certain topic or object. Depending on how the questions and queries are formulated, the resulted summary may represent the text content from multiple viewpoints and angles. CW is thus one of the many tools that will be helpful in text summarization. There are nonetheless many challenging issues in implementing the CW framework. Large amount of propositions contained in a text can easily prevent the framework of computing with words from being of practical use. Another major barrier has been the vast amount of linguistic and world knowledge needed in natural language understanding, as well as the flexibility and context dependence in the way the knowledge are applied.

Summarization is always about emphasizing certain information over some others. Different people can abstract out different substance from the same source, and even the same person can formulate different summary information from the same source at different time, in different situations. "What is important" changes with the change of the context. Important information may simply be a direct or derived answer to a query. Important information may be "new information" to fill the cognitive gaps in reader's memory and mental models concerning world objects. Important information may refer to the core content that

author intends to convey to the readers with bulk of background details, repetitions, comments left out.

Text summarization calls for a combination of both the shallow and deep text analysis methods. Good enough and working solutions for text summarization will be found in the middle between an elegant model but infeasible computationally, and a more crude techniques and computationally effective solutions. The development in the field of natural language processing and computational linguistics has produced rather mature parsing techniques and abundant lexical and ontology resources. Significant progress has also been made in compiling and learning knowledge from corpus analysis using manual, statistical or hybrid techniques. All these are enablers for the application of CW methods. Much waits to be explored for text summarization and language understanding systems to take full advantage of these methods and resources for automated content analysis and reasoning.

Acknowledgements

Financial support from Academy of Finland is gratefully acknowledged (Grant Number 111692). The author would also like to thank Johnny Lindroos, Fredrick Sundell and Marketta Hiisa for their contribution to the project and their assistance in carrying out some of the experiments.

References

1. Paice C. D., "Constructing literature abstracts by computer: techniques and prospects", Information Processing and Management: an International Journal, Volume 26 Issue 1, April 1990
2. Salton G., J Allen, C Buckley, A Singhal, "Automatic analysis, theme generation, and summarization of machine-readable texts", Science, 1994
3. Hovy E. and C. Lin, "Automated Text Summarization in SUMMARIST", Mani and Maybury (eds), Advances in Automatic Text Summarization, MIT Press, 1999.
4. Mani I. and M. T. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999.
5. Spärck-Jones K. "Automatic Summarizing : Factors and Directions", in Mani and Maybury (eds), 1999
6. Luhn H. P., "The Automatic Creation of Literature Abstracts", 1958, in Mani and Maybury (eds), Advances in 1. Automatic Text Summarization, MIT Press; 1999
7. Climenson, W. D., Hardwick, N. H., Jacobson, S.N. (1961). Automatic Syntax Analysis in Machine Indexing and Abstracting.
8. Edmundson, "New Methods in Automatic Extracting", 1968, in Mani and Maybury (eds), Advances in Automatic Text Summarization, MIT Press; 1999.
9. Carbonell J. G. and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", In A. Moffat and J. Zobel (eds), Proceedings of the 21st Annual Int. ACM SIGIR Conf' on Research and Development in Information Retrieval, Melbourne, Australia, 1998, pp. 335-336.
10. Endres-Niggemeyer B., Summarizing Information, Springer-Verlag Berlin Heidelberg, September 1998, Hardcover, 374 pp
11. Marcu D., 1999. "The automatic construction of large-scale corpora for summarization research", Proceedings of 1. SIGIR'99, University of Berkeley, CA, August 1999
12. Marcu D. and L. Gerber (2001). "An Inquiry into the Nature of Multidocument Abstracts, Extracts, and Their Evaluation", Proceedings of the NAACL-2001 Workshop on Automatic Summarization, Pittsburgh, PA, June 3, 2001.
13. McKeown K. and D. R. Radev, "Generating Summaries of Multiple News Articles", in Mani and Maybury (eds), 1999.
14. Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Celebi A., Drabek E., Lam W., Liu D., Qi H., Saggion H., Teufel S., Topper M. and A. Winkel, The MEAD Multidocument Summarizer, MEAD Documentation v3.08, 2003.
15. Radev D., Jing H., Stys M. and D. Tam, "Centroid-based Summarization of Multiple Documents", Information Processing and Management, vol. 40, 2004, pp. 919-938.
16. Lacson R., R. Barzilay and W. Long "Automatic analysis of Medical Dialogue in the Home Hemodialysis Domain: Structure Induction and Summarization", In Journal of Biomedical Informatics, 2006.
17. Galley M., "Automatic summarization of conversational multi-party speech", Proceedings of AAAI SIGART Doctoral Consortium, 2006.
18. McKeown K., L. Shrestha and O. Rambow, "Using question-answer pairs in extractive summarization of email conversations", Proceedings of CICLing 2007, volume 4394, pages 542-550
19. Lindroos J. Automated Text Summarization using MEAD: Experience with the IMF Staff Reports, Master Thesis, Åbo Akademi University, 2006
20. Liu, S. and J. Lindroos, "Towards Fast Digestion of IMF Staff Reports with Automated Text Summarization Systems", proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, December 18-22, 2006, HongKong
21. Liu, S., Hiisa M. and F. Sundell, "Automatic summarization of intensive care nursing narratives", manuscript, 2007
22. Harper R., Inside the IMF, 1st Edition, Academic Press, 1998.
23. Van Dijk T. A., News as discourse, 1988

24. Lacson R. C., R. Barzilay and W. J. Long, "Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization", Journal of biomedical informatics (2006), Volume: 39 Issue: 5 Pages: 541-55
25. McKeown K., R. Passonneau, D. Elson, A. Nenkova and J. Hirschberg, "Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization", 28th Annual International ACM SIGIR Conference on Research and Development in Information Retreival (SIGIR 2005), Salvador, Brazil.
26. Hovy E., Chin-Yew Lin, L. Zhou and J. Fukumoto, "Automated Summarization Evaluation with Basic Elements", 2005
27. Salton, G., Wong, A., and Yang, C. S., "A vector space model for automatic indexing", Communications of the ACM, 18(11):613-620. 1975
28. Daniel Jurafsky and James H. Martin, Speech and Language Processing, Prentice Hall
29. Papineni K., Roukos S., Ward T., and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002, pp. 311-318
30. Lin, C-Y., "ROUGE: A package for automatic evaluation of summaries", in M-F. Moens, & S. Szpakowicz (Eds.), Text Summarization Branches Out: Proceedings of the ACL-04 Workshop 2004, pp. 74-81
31. Nenkova, A., & Passonneau R. (2004). "Evaluating content selection in summarization: the pyramid method". In 1. S. Dumais, D. Marcu, & S. Roukos (Eds.), HLT-NAACL 2004: Main Proceedings (pp.145-152). Association for Computational Linguistics.
32. Nenkova A., R. Passonneau, and K. McKeown, "The Pyramid Method: Incorporating human content selection variation in summarization evaluation", ACM Transactions on Speech and Language Processing, 4(2), 2007
33. Zadeh L A., "Fuzzy Logic = Computing with Words", IEEE Transactions on Fuzzy Systems, 2, 103-111, 1996
34. Zadeh L A., "From Computing with Numbers to Computing with Words -- From Manipulation of Measurements to Manipulation of Perceptions", IEEE Transactions on Circuits and Systems 45 (1999) 105-119
35. Kintsch W. and van Dijk T A.: Toward a model of text comprehension and production, Psychol. Review 85 (1978) 363-394
36. Van Dijk, T. A. and Kintsch, W.: Strategies of discourse comprehension. New York: Academic Press (1983)