UNIVERSITY

of

GLASGOW

Department of Computing Science

MSc in Advanced Information Systems

Reflecting user information needs through query biased summaries

Anastasios Tombros

Glasgow

September 1997

Abstract

It is often the case that information retrieval (IR) systems do not present users with enough information in order to assist them in judging the relevance of documents retrieved in response to a query. Typical IR systems usually output the title and the first few sentences of the source text as ‘relevance clues’. This usually leads to the users having to refer to the full text of the documents in order to locate the relevant to the query content that each document conveys.

The aim of the work reported in this thesis, is to investigate the effectiveness of automatically generated summaries customised to a specific query, in assisting users to judge the relevance of documents retrieved in response to that query. More specifically, a summary will automatically be generated and presented for each retrieved document, aiming to provide users with enough evidence about the relevance (or non-relevance) of each document to the query.

In order to examine the effectiveness of this approach, a comparative evaluation against a typical IR system response is performed. The results from this evaluation indicate that the presence of the automatically generated summaries as accompanying information for each retrieved document, improves the accuracy of the relevance judgements while at the same time reduces to a minimum the need to refer to the full text of the documents.

Acknowledgements

Indebted to Mark Sanderson for his supervision, help, and motivation on my research work.

Also to Phil Gray for his help in conceptualising and designing the experiment.

Many thanks to Michael Pryde for his help in getting all the people together for the experiment. Also many thanks to everybody who participated in the experiment. Their help was invaluable.

Indeed obliged to Marcos Theophylactou for hints & tips on Unix scripting. He got me out of lots of trouble.

Finally, thankful to all those who put up with my always-changing moods throughout this year.

…to those who believed in me

Table of Contents

Chapter 1

Introduction

1.1 Motivation

The motivation for the research work reported in this thesis is twofold. It emanates both from the need of users to be presented with enough evidence about the relevance of documents to their information needs, and from the observation that the evaluation of automatic text summarisation systems usually does not take into account qualitative characteristics of summaries (such as their utility in presenting users with relevance evidence). The connection of the two points should become apparent through the discussion evolved in the following paragraphs.

A *communication process* can be thought of as a sequence of events resulting in the transmission of something called *information* from one object (source) to another (destination). [Goffman & Newill, 1967]. Information[1] can actually be of various forms. However, in the context of this thesis we shall restrict the notion of information to that conveyed in textual documents, and we shall therefore regard a *document* as a basic unit of information.

[1] In fact we can not adequately define information; nevertheless we comprehend its properties and effects. [Saracevic, 1969]

Systems whose function is the carrying out of a communication process are usually referred to as *information systems*. In the context of this thesis we shall concentrate on a specific category of such systems, information retrieval (IR) systems. It is very often the case that users are engaged in a communication process with such a system. The stimulus for the initiation of the process, is the formulation of a specific information need by the user[2]. The user subsequently inquires the system on his request for information, through a formulated *query* that he inputs to the system. In the context of the communication process, it is then the system's responsibility to transmit the requested information to the user. In order for the process to be carried out successfully, the user must be presented with information that is *relevant* to his request. At this point, the issue of how the user will effectively judge the relevance of the presented information (documents in our case) becomes important. Clearly, the IR system must inform the user on the relevance, or not, of the presented documents to his information need, and therefore allow for an effective decision on the relevance of the presented information.

It is this very aspect of the communication process, combined with the observation that typical IR systems do not satisfactorily address this issue, that constitutes the focus of motivation for the research reported in this thesis. It is proposed that an automatically generated summary of each presented document, customised to the user's information need, has the potential to effectively inform the user about the document's relevance to his request.

The automatic summarisation of documents is an issue that has been addressed by a number of researchers. Two major aspects of this research issue can be discerned: the design and implementation of summarisation systems, and the evaluation of their effectiveness. It is the latter aspect that constitutes the second point of motivation for the research work reported here. Work on the evaluation of summarisation systems so far, has been mostly restricted to measuring quantitative features of the summaries (e.g. similarity between automatically generated summaries and human prepared ones). The evaluation of such systems in an operational, task-based environment, where qualitative

[2] Little is known about the actual process that leads to the formulation of a request for information.

features of the summaries can be measured, is a rather neglected aspect of the related work in the field.

1.2 Thesis aims

The aim of the work reported in this thesis, is to investigate the effectiveness of automatically generated summaries customised to a specific query, in assisting users to judge the relevance of documents retrieved in response to that query. More specifically, a summary will automatically be generated and presented for each retrieved document, aiming to provide users with enough evidence about the relevance (or non-relevance) of each document to the query.

In this way, the connection of the two focal points of motivation presented previously becomes apparent. The development of a task-based evaluation scheme for summarisation systems can be viewed as the medium for examining the effectiveness of summaries in informing users on the relevance of documents to a request for information.

1.3 Thesis outline

Chapter 2 presents an introduction to information retrieval (IR) systems. In the context of this chapter, the basic concepts and terminology that will be used throughout this thesis are established. The basic IR techniques that are applied on documents and on queries will be described. Moreover, the typical output of such systems will be discussed, criticising the way that users are informed on the relevance of the presented documents. In this way, the first source of motivation for this thesis will be clarified.

The work carried out in the field of automatic text summarisation is presented in Chapter 3. A critical overview of the approaches followed, both in the implementation and the evaluation of summarisation systems, will be given in that chapter. Finally, the proposed task-based evaluation scheme will be outlined, providing insight to the second focal point of motivation of the reported work.

The summarisation system that was developed for the purposes of the research work reported in this thesis shall be described in Chapter 4. The system architecture will initially be presented. The discussion will then evolve around that architecture, describing its constituent parts. The rationale of the basic system design will finally be presented, justifying the main design decisions taken.

In Chapter 5, the experimental design adopted for the evaluation of the developed summarisation system will be presented. First, an introduction to basic issues pertaining to the design of experiments will be given. The way that these issues were addressed in the specific experimental procedure will then be described, and finally the actual evaluation scenario will be elaborated.

The presentation and analysis of the experimental results will take place in Chapter 6. The measures of user performance will first be established, and then the way that these measures were quantified through the data collected from the experimental procedure will be described. The presentation and analysis of the results will then follow, and a discussion on the conclusions drawn from these results is given.

Finally, Chapter 7 indicates some points that future research work could concentrate on, and presents the conclusions that can be drawn from the work reported in this thesis.

Chapter 2

Basic concepts of information retrieval systems

2.1 Introduction

> 'An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his enquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.' [Lancaster, 1968].

This concise definition of the basic functions of an information retrieval (IR) system, offers a perfectly appropriate starting point around which the discussion presented in this chapter shall evolve. Within the few lines of the above definition, the three major components of an IR system have already been identified: The user with his request for information, the collection of documents to apply this request on, and finally the response of the IR system in relation to the user's request for information.

The initial motivation for the development of information retrieval (IR) systems was the need to cope with the huge scientific and research literature developed since 1940. Nevertheless, the 'information revolution' that has recently taken place, especially the explosive growth of the Internet, has greatly expanded the scope of IR systems. It is now not only research literature that is of interest to such systems, but also a wide spectrum of heterogeneous types of information, including multimedia data (e.g. pictures, speech, animation, etc.). However, in this chapter we shall restrict the discussion of IR systems to the context of textual documents only.

It would be unrealistic to attempt to cover every aspect of IR systems within a few introductory pages. Such an approach is beyond the aim of this thesis. Thus, this chapter proposes to provide an introduction to IR systems presented in a way that is adjusted to the context of the work presented in the following chapters. The principal automatic indexing techniques that are used to derive a meaningful computer representation of the documents are first presented in section 2.2. The basic operations applied on the user's request for information are subsequently discussed in section 2.3. In section 2.4 a number of issues concerned with the response of the IR system in relation to the user's query are examined, and finally section 2.5 gives an overview of the issues discussed in this chapter.

2.2 Automatic indexing of documents

This section proposes to deal with the issue of how documents are represented in a form suitable for a computer to use. The process for the construction of this internal representation of the documents is called the indexing process. The essence of this process is to assign to each document a set of terms, or concepts, that are capable of representing the document's content, and subsequently to assign to each term a weight reflecting its importance for characterising the document[2].

In the early years of information retrieval the indexing process was conducted manually by human indexers. However, when indexers manually assign terms to documents, the indexing process becomes rather difficult to control. The indexer performs his task biased by his perception of the set of index terms a user would make use of in order to retrieve the specific document[3]. For this reason an effective indexing procedure is probably best conducted by using objective term selection criteria that are applied in a uniform manner for all the documents of the collection.

[2] This measure of importance is usually met in the literature as *discriminating* or *resolving power* of the index terms[Luhn, 1958; Van Rijsbergen, 1979].

[3] See [Van Rijsbergen, 1979], pp. 6.

Automatic indexing techniques fall under two broad categories, depending on the approach they use to select the index terms: they can be either statistical (using word frequency information) or linguistic (exploiting syntactic or semantic relationships of the words in a given context). In the present discussion we shall concentrate solely on the statistical approach to indexing[4].

2.2.1 Selection of index terms

The first step towards obtaining a set of index terms is to remove from the input text all high frequency words. Such words are often called ‘stop words’, and their contribution to the characterisation of the document’s content is minimal. Typical examples of stop-words are articles (e.g. ‘the’) and prepositions (e.g. ‘in’, ‘at’). The removal of the stop-words is usually performed by comparing the input text with a ‘stop list’ of words which are considered to have no value as indexing terms. The benefit of this method is that without losing any significant information it is possible to achieve a reduction of the text volume of up to 50 percent[5].

The next step to the automatic selection of index terms is to remove the suffixes from the remaining words of the input text. This can be achieved through the application of a stemming algorithm[6], that will reduce related words to a common root form (stem); for example, the words ‘manufacture’ and ‘manufacturing’ will be mapped to the same entity, ‘manufactur’, in the vocabulary of index terms. The set of words resulting from this process forms the set of index terms (or keywords) of the document collection. A descriptor for each document can subsequently be defined as the set of its constituent index terms. A usual way to represent the document descriptor is by means of a binary vector. Each element of the vector corresponds to an index term of the

[4] Interested readers may refer to [Sparck Jones & Kay, 1973] for a comprehensive overview of linguistic approaches to automatic indexing.

[5] See [Van Rijsbergen, 1979], pp. 17.

[6] An extensive discussion about stemming algorithms can be found in [Frakes & Baeza-Yates, 1992], pp 131-151.

document collection, and the values of the elements are 1 or 0 depending on whether the corresponding index term is present in the document or not.

2.2.2 Term weighting techniques

Once the set of the descriptors has been determined for each one of the documents within the collection, a way to discern the discriminating power of the descriptor terms must be defined. For this purpose several term weighting methods have been developed. The general consensus of these methods is the attempt to utilise information both from within the document itself, and from the document collection.

As early as 1958, H.P. Luhn related the discriminating power of an index term with its frequency of occurrence within the document. In his influential research, Luhn concluded that the highest discriminating power should be associated with middle frequencies. We can therefore define the first term weighting criterion, the term frequency, as follows:

$$\mathrm{WEIGHT}_{ik} = \mathrm{FREQ}_{ik}$$

where

WEIGHT_{ik} is the weight of term k in document i

FREQ_{ik} is the frequency of occurrence of the word k in the context of the document i, normalised to the length of document i

The term frequency criterion, however, makes no attempt to utilise information from the document collection, i.e. it does not examine the way index terms distribute over the entire collection. It consequently makes no distinction between words that occur in every document of a collection and those that occur in only a few items. The inverse document frequency (IDF) attempts to capture the rather intuitive assumption that the discriminating power of an index term increases with the frequency of the term within the document, but decreases with the number of documents $\mathrm{DOCFREQ}_k$ to which the term is assigned. Thus, a definition of this weighting criterion is:

$$\mathrm{WEIGHT}_{ik} = \mathrm{FREQ}_{ik} * [\log_2(n) - \log_2(\mathrm{DOCFREQ}_k) + 1]$$

where

$FREQ_{ik}$ is the frequency of term i in document k

n is the number of documents in the collection

$DOCFREQ_{k}$ is the number of documents to which term k is assigned

The IDF measure in the above formula, is the second factor of the product

It should be noted that the issue of term weighting is an extensively researched area of information retrieval. As a consequence, a great number of variations of the above weighting schemes can be found in the literature. For a comprehensive discussion on the various weighting schemes and their implications on the process of text retrieval, readers who wish to may refer to [Van Rijsbergen, 1979], and [Salton & Buckley, 1988].

2.3 Query operations

Until now we have examined how index terms are selected from the documents of interest, and how weights are assigned to these terms according to their discriminating, or resolving, power. Recalling the description of the basic functions of an IR system given in the beginning of this chapter, we can see that we have only probed one of the three major issues so far: the collection of documents. It is the purpose of this section to deal with the manipulation of the user's query by the IR system.

A query can be defined as a formal statement of information need put to the IR system by a user. The end-user usually enters his query either in a Boolean form (connecting the query terms with logical operators such as 'AND' or 'NOT'), or in a free text form where a simple query such as a sentence or a phrase is being formulated. In the context of this thesis we shall limit the discussion on queries of the latter form.

Once the query has been input to the IR system, it has to be represented in a form suitable to be manipulated by the computer. This is usually done in a manner similar to the construction of the document descriptor: the query is being parsed in its constituent elements (i.e. the query terms), possibly by removing the stop-words from the input text, and by applying an appropriate stemming algorithm. As a result a vector representation of the query is obtained, where the presence of a word from the collection's index vocabulary in the query is denoted by a value of 1 in the respective vector position, while a 0 is used to indicate a lack of such a word.

Once the query descriptors have been obtained, a mechanism that would somehow calculate the relevance of a document to the specific query must be applied. Whereas it may be intuitively possible for a human to establish the notion of relevance, for a computer to do so there must be a well defined model by means of which relevance decisions can be quantified. The two principle models used for such a purpose are the vector-space, and the probabilistic[7]. Without attempting to present the theoretical basis of the models, it would suffice for the purposes of this section to say that their role in an IR system is to generate a measure indicative of the matching between the query and the documents of the collection. This matching, or similarity, measure is derived by means of comparison between the documents' and the query's descriptors, and by exploiting knowledge about the distribution of the index terms throughout the collection.

2.4 The output of an IR system

The similarity measure calculated by the above process attempts to quantify the relevance of a document in relation to a specific query. Based on that quantification, the IR system will usually present the documents from the collection that match the query in a ranked list, in decreasing order of the similarity measure. In this way one could say that the system 'suggests' the documents that are most likely to be relevant to the user's request for information. The actual resolution about the document's relevance is a task that the user alone will have to perform.

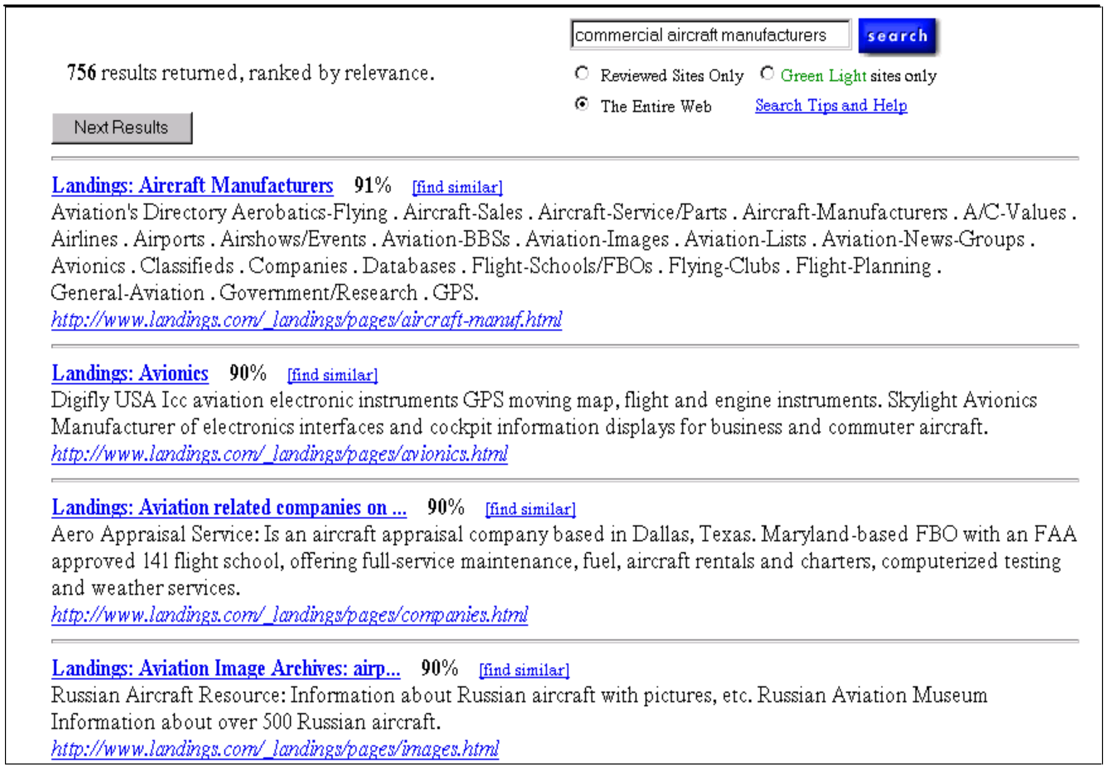
[7] Readers who are interested in the theoretical basis of the two models should refer to Chapter 6 of [Van Rijsbergen, 1979] (probabilistic retrieval), and to Chapter 4 of [Salton, 1983] (vector-space model).

2.4.1 A typical interaction with an IR system

The scope of IR systems has expanded in the last few years, primarily due to the rapid growth of the Internet. Hundreds of thousands of documents are being made accessible in an electronic form, and the domains these documents cover are of great diversity. As a consequence, users of IR systems today are not only academic researchers or trained professionals (e.g. librarians), but also simple end-users who wish to locate interesting sources of information in the globe of documents made available. It is therefore of great importance for an IR system to provide its users with easily recognised clues about the relevance to their information need of the retrieved documents.

In a typical interaction with an IR system the user has a specific information need, which he expresses through a query to the system. Let us hypothesise that a user is interested in finding out information about “commercial aircraft manufacturers”. In Figure 2.1 the response of an IR system in relation to that query is presented. For each of the four documents presented in the ranked list, its title, its first few sentences, and its actual location (from where the full text of the document can be fetched) is shown to the user. A quantification of its possible relevance to the query is also shown next to the document’s title. For example, the first document in the list is given a 91 percent ‘similarity’ with the query. The user then, utilising the information being presented to him, has to decide which of the retrieved documents convey the information he is trying to locate through his query.

Ideally, that would be feasible without having to refer to the full document text. In that case, the information presented to the user would be evident enough of the document’s relevance. But it is rather dubious whether the first few sentences of a document and its title are able to give a clear view of the document’s content in relation to the user’s query. As a result, users frequently have to refer to the full text of the document, rendering the process of relevance judgement time consuming. Even when users refer to the full text of a document, the very nature of the documents may have a confounding effect: They may be large and difficult to manage, and the relevant information to the query they may convey can be widely scattered, and therefore hard for the user to extract.



commercial aircraft manufacturers

756 results returned, ranked by relevance.

○ Reviewed Sites Only ○ Green Light sites only

● The Entire Web

Landings: Aircraft Manufacturers 91% [find similar]
Aviation's Directory Aerobatics-Flying . Aircraft-Sales . Aircraft-Service/Parts . Aircraft-Manufacturers . A/C-Values . Airlines . Airports . Airshows/Events . Aviation-BBSs . Aviation-Images . Aviation-Lists . Aviation-News-Groups . Avionics . Classifieds . Companies . Databases . Flight-Schools/FBOs . Flying-Clubs . Flight-Planning . General-Aviation . Government/Research . GPS.
http://www.landings.com/_landings/pages/aircraft-manuf.html

Landings: Avionics 90% [find similar]
Digifly USA Icc aviation electronic instruments GPS moving map, flight and engine instruments. Skylight Avionics Manufacturer of electronics interfaces and cockpit information displays for business and commuter aircraft.
http://www.landings.com/_landings/pages/avionics.html

Landings: Aviation related companies on ... 90% [find similar]
Aero Appraisal Service: Is an aircraft appraisal company based in Dallas, Texas. Maryland-based FBO with an FAA approved 141 flight school, offering full-service maintenance, fuel, aircraft rentals and charters, computerized testing and weather services.
http://www.landings.com/_landings/pages/companies.html

Landings: Aviation Image Archives: airp... 90% [find similar]
Russian Aircraft Resource: Information about Russian aircraft with pictures, etc. Russian Aviation Museum Information about over 500 Russian aircraft.
http://www.landings.com/_landings/pages/images.html

Figure 2.1 Typical output of an IR system

Recognising the cognitive overhead imposed on users, there have been attempts to concentrate the user’s attention on parts of the text that possess a high density of relevant information. These methods, known as passage retrieval [Callan, 1994; Knaus et al., 1995], instead of retrieving the full document text in response to the query, they identify and present to the user individual text passages that are more responsive to particular user needs than the full document texts. The main advantage of these approaches is that they provide an intuitive overview of the distribution of the relevant pieces of information within the documents. As a result, it may be easier for the user to decide on the relevance to his query of the retrieved documents. However, even this approach does not alleviate the need to refer to the full text of the retrieved documents.

It is one of the aims of the research work reported in this thesis, to investigate a different approach to presenting the user with clues about the relevance of the retrieved documents to his information need. Such an approach will aim at minimising the need to refer to the full document text, while at the same time providing enough information to the user so as to support his retrieval decision. It is proposed that an automatically generated summary for each document, customised to the query input by the user, can provide such a function within the frame of an IR system.

2.5 Summary

This chapter provided an overview of the main concepts and functions of an IR system. More specifically, the basic IR techniques applied on a document collection (selection and weighting of index terms), and on the user's query (calculation of similarity measures between the documents and the query) have been described. Subsequently, the output of typical IR systems was discussed, emphasising on the observation that users are not presented with enough information about the relevance of the retrieved documents to their information need. The emphasis was specifically put on that issue in order first to justify a major source of motivation for the research work reported, and then to clarify one of its principal aims: To investigate the effect of automatically generated, query-biased summaries of the retrieved documents on the process of judging the relevance of these documents.

Chapter 3

Related work on the automation of text summarisation

3.1 Introduction

The need to automate the process of text summarisation has become imperative, mainly due to the rapid growth of the amount of textual information available in machine readable form. However, despite the significant advances of information science and natural language processing over the last years, work on automatic text summarisation so far has been rather limited and heterogeneous, resulting in domain restricted systems.

A document summary conventionally refers to an abstract-like condensation of a full text document, that presents succinctly the objectives, scope, and findings of a document [Maizell et al., 1971]. The minimal function that any useful summary should provide is being *indicative* of the source's content, hence helping a reader to decide whether looking at the whole document will be worthwhile. In this sense, summaries can serve as a preview format to support a retrieval decision on the full text of the document. Many summaries also contain *informative* material, such as main results and conclusions. In this case, summaries can act as stand-alone document surrogates that allow the reader to recover useful information without having to refer to the whole document at all. A document summary could also be used to improve the precision of text search, as searching against a condensed version of a document may decrease the probability that irrelevant text will match a query [Maizell et al., 1971; Brandow et al., 1995].

Numerous researchers have addressed the automation of document summarisation. Since its beginnings, automatic text summarisation has been performed primarily by the selection of sentences from the original document [Luhn, 1958; Edmundson, 1969; Rush et al., 1971; Paice, 1981; Brandow et al., 1995; Kupiec et al., 1995; Salton et al, 1997]. This approach can be better termed as *sentence extraction* rather than summarisation. Despite its problems and flaws, it is capable of producing acceptable summaries that are domain independent. Section 3.2 of the present chapter will discuss the principal sentence extraction methods, while section 3.3 shall present their main problems, and some proposed solutions to these problems.

Although there have been attempts to produce coherent summaries by language generation, and artificial intelligence techniques [DeJong, 1982; Jacobs & Rau, 1990; McKeown et al., 1995; McKeown & Radev, 1995; Aretoulaki, 1997], they are capable of processing texts only within a narrow domain whose characteristics are predictable and well understood (e.g. news stories, financial and commercial reports). There is not enough evidence that such systems will be able to manipulate domain independent text in the foreseeable future. This category of approaches to text summarisation will be presented in section 3.4 through a description of a basic architecture they follow.

As far as the question of the evaluation of summarisation systems is concerned, schemes that have been proposed [Edmundson, 1969; Brandow et al., 1995; Kupiec et al., 1995; Salton et al., 1997] are only superficial. They try to evaluate systems by means of how many similar sentences exist in auto-summaries and in human prepared extracts. Section 3.5 will deal with the issue of evaluation, discussing the approaches mentioned above, and proposing a different evaluation procedure that was followed for the purposes of this thesis. In this section, a comparative evaluation of sentence extraction methods will also be given.

3.1.1 A model for automatic summarisation

It is advantageous to have a general model of text summarisation as a process that can be used to judge future research approaches. This general model makes a distinction between *source text interpretation* and *summary text generation* [Maybury, 1995; Sparck Jones & Enders-Niggemeyer, 1995]. The process of source text interpretation can be subdivided into: understanding the content of the document and identifying important information contained in it. Statistical and linguistic approaches have been used for the identification of significant information within a document, but automatic comprehension of a document's content is, until now, beyond the state of the art [Brandow et al., 1995; Salton et al, 1997]. The second part of the summarisation model, summary text generation, also presents difficulties in being automated. That is the main reason why only few systems have focused on language generation techniques in order to produce the summary text. The most commonly used approach is to present sentence extracts of the source text, and then attempt to make them look as if they belong together by applying linguistic rules [Paice, 1990].

The model described above seems to have a pipeline-like architecture: source analysis, identification of useful information, formation of summary representation, and finally generation of the summary text. On the other hand, observation of summarising carried out by professional human abstractors, suggests that the modelled processes occur concurrently and are co-constraining [Maybury, 1995]. For example, in human summarising, the structure of the source document and the abstractor's presuppositions and personal bias tend to emphasise some items of information while suppressing others, thus influencing processing and structure of the resulting summary. At the same time, the specifications of the targeted summary (type, content, and purpose) also affect the kind of processing that occurs. This contradiction should provide strong motivation towards more intensive observation and research on the nature of human summarising.

3.2 Automatic text summarisation by sentence extraction

An early attempt that tried to address automatic summarisation was reported in a paper by H.P. Luhn published in 1958 [Luhn, 1958]. This work concentrated on the generation of *extracts*, that is sets of sentences from the source text, selected to provide a good indication of the source’s main subject. Later research, particularly that having been reported up to about 1970, was clearly influenced by Luhn’s approach. This rather simplistic solution allows for domain independent summarisation, but as we shall see later on is far from being perfect, and is subject to various improvements.

3.2.1 Automatic sentence extraction methods

The rationale of sentence extraction methods is to find a subset of the source document that is indicative of its contents, typically by scoring words and then sentences according to specific rules. Those rules are mainly concerned with the identification of clues for the importance of each of the source’s sentences.

The criteria for attributing significance to words of the source text, may be *positional* in virtue of their occurrence in titles or section headings, or *semantic* in virtue to their relation with words like ‘summary’, or perhaps even *pragmatic* in the case of names of specialists mentioned in the text, footnotes or bibliography. They may also be *statistical* by involving frequency of occurrence or *non-statistical* by involving only the fact of occurrence as for example the occurrence of a word in a title. A score for each sentence is then obtained by means of a function of the scores of its constituent words. The generation of the summary text is then reduced to the presentation of sentences with the highest scores in order of occurrence in the original text. The methods used for identifying clues to sentence significance that have been tried by various researchers will now be discussed.

The keyword method

This approach was introduced by Luhn [Luhn, 1958] and is based on the hypothesis that high-frequency words are indicative of the document's content and thus considered as positively relevant. Sentences that contain these frequently used words are then scored using functions of their frequency counts. A requirement for this method is the existence of a list containing index terms for the document. In all of the work reported so far, the index terms are single words, which are sometimes conflated, and are called keywords.

The procedure suggested by Luhn involves passing the complete document through a stoplist in order to remove common words (pronouns, articles, prepositions etc.), and sorting the remaining words into a descending frequency ordered list. A specific frequency value is then chosen as a threshold value, and words whose frequencies are higher than this threshold are designated as keywords. This procedure aims at giving high significance to normally rare words which occur frequently in the document, while giving low significance, among others, to normally rare words which occur rarely in the document.

The final step is to extend the measure of significance from words to sentences and to compute a score for each sentence. For this purpose various methods of varying complexity have been used. Luhn noted the importance of phrases for denoting concepts, and the criterion he used was the relationship of the significant words to each other rather than their distribution over a whole sentence. In doing that, he set a limit for the distance at which any two significant words should be in order to be considered significantly related. Through experimentation and analysis of many documents, he concluded that a useful limit would be four or five non-significant words between significant words. He then based the overall sentence score on the groups or clusters of keywords contained in each sentence. Edmundson (1969) on the other hand, used a rather simple scoring function, by weighting keywords according to their frequency in the document and by summing the keyword weights for each sentence.

The cue method

H.P. Edmundson [Edmundson, 1969] experimented with the hypothesis that certain of the words occurring in a sentence provide an indication of whether the sentence deals with important concepts. These words are not in themselves keywords as defined in the previous method, and can be categorised as *bonus words, stigma words* and *null words*.

Bonus words according to this method increase the sentence score (they are considered as being 'positively relevant'), and they mainly include superlatives and value words, such as "greatest" and "significant". Stigma words on the other hand have a diminishing effect on the sentence score (they are considered as 'negatively relevant'). Such words mainly include anaphors and belittling expressions, such as "impossible" and "hardly". Null words finally, are irrelevant words that contribute neither negatively nor positively to the sentence score. The final weight for each sentence is calculated as the sum of the cue weights of its constituent words.

A modification of this method [Rush et al., 1971] was motivated by the belief that opinions and subjective notions should not be included in a summary. According to this approach, there are certain 'cue words' that provide unequivocal clues to such things as opinion and subjectivity, as well as positive notions. Their method involves the construction of a 'word control list' which contains individual words and short phrases. Each entry in the list is accompanied by a semantic code, denoting whether the word or phrase is a positive/negative indicator or whether it has some other significance. Negative indicators are prevalent in this list, so that the process of extraction relies more on rejection than on selection of sentences.

The indicator-phrase method

Indicator phrases, which are more elaborate constructs than cue words, contain words that are likely to accompany indicative or informative summary material. Such phrases commonly convey explicit statements about the topic of the text, for example: "The aim of this paper is to examine…", "The purpose of this article is to review…" or "In this report, we outline an investigation into…".

As it is obvious, there are numerous indicator phrases that one may come up with. However, it is suggested that a satisfactory coverage of such phrases can be achieved by defining as few as seven or eight distinct basic types [Paice, 1981]. Each of these basic types can be seen as a ‘template’, and all other indicator phrases can be produced by reference to the templates. Consider for instance the first two examples given above: they belong to the same basic template type, since they possess the same basic structure. A template type can be viewed as a generic formula upon which basic transformations can be applied (e.g. article, preposition, or word substitution) in order to derive a specific indicator phrase. Weights are then assigned in a cumulative manner to various points in a template. In this manner a score may be returned even if the template is not matched right to the end.

The title method

This approach is based on the hypothesis that the author of a document ‘reveals’ the main concepts in the title of his writing. Also, when the author partitions the body of the document into major sections, he summarises it by selecting appropriate headings.

Like the keyword method, a list of index terms is created for the document prior to the sentence scoring process. In the title method, candidate terms are selected from the title, subtitle and headings of the document. Edmundson [Edmundson, 1969], who experimented with this method, assigned heavier weights to content words of the title than to content words of headings and subtitles. He then computed the final title weight for each sentence as the sum of the title weights of its constituent words.

The location method

This approach results from the observations that sentences occurring under certain headings of a document convey significant content and are thus relevant, and that topic sentences tend to occur very early or very late in a document and its paragraphs. For example, within a paragraph, the first sentence is usually central to the subject of the text, while in some cases it is the last.

The location method assigns positive weights to words occurring under certain headings of the document (heading weight), e.g. “Introduction”, “Conclusions”. In addition to this, it also assigns positive weights to sentences according to their ordinal position in the text (ordinal weight), i.e. in first and last paragraphs of the document and as first and last sentences of paragraphs. The final location weight for each sentence is the sum of its heading weight and its ordinal weight.

Relational criteria

The reasoning of this approach is that the most important sentences in a document are those which are related to the largest number of other sentences. This novel approach is based on the generation of a 'semantic structure' for the document. This is a graph representation of a document, where sentences are vertices and significant inter-sentence links are represented by edges. Links are considered to exist wherever distinct sentences refer, or contain reference, to the same subject. The score for each sentence is then computed as a function of the number of distinct sentences to which it is significantly related, and the degree of change in the semantic structure which would result if the sentence was deleted [Skorokhod'ko, 1971].

A similar approach, based on the notion of paragraph rather than sentence significance, was introduced by [Salton et al., 1993]. Paragraph significance is defined in the same way as previously: the most important paragraphs are those which are related to the largest number of other paragraphs in the source document. In order to actually identify the important paragraphs, ideas from the automatic hypertext link generation research are being used [Salton et al., 1997], but instead of producing inter-document links between various related documents, links between various paragraphs of the document are produced (intra-document links). By using a similar graph representation of the document, paragraphs with the larger number of intra-document links are selected for extraction. In this way it is possible to produce longer summaries, but it is equally possible that these summaries will contain more coherent text since a paragraph contains more context [Salton et al., 1997].

3.3 Points for discussion

Sentence extraction systems allow summarisation in arbitrary domains, since they do not use any field-specific knowledge. These methods on the other hand, suffer from problems originating from the fact that sentences are extracted from the source document and are merely presented in order of occurrence in the original text. No attempt to perform a synthesis of the summary text is made. The main problems of sentence extraction methods will now be discussed, and some potential enhancements will also be described.

3.3.1 Textual cohesion

The main problem with automatic extraction methods is that they produce summaries that usually contain incoherent or inconsistent text. Disruption of textual cohesion of the generated summary may result from the extraction of explicit references within sentences, which can only be understood by reference to material elsewhere in the text [Paice, 1990]. Anaphoric references are the most common example of such a disruption. To illustrate this problem better, consider the following example of an automatically extracted summary:

“The work undertaken examines… Only low carbon steels were selected for experimentation. **This** accounts for the classic appearance of ductile failure with the centre of the wire ….”.

As it is made obvious by reading the text, the anaphoric reference introduced with the word ‘this’ refers to material elsewhere in the document, not to the fact that “Only low carbon steels were selected for experimentation.”. The extraction methods described previously do not take into account this aspect. Another form of incoherent summary text, results from the fact that extracted sentences may not be consecutive in the original text and may not naturally follow one another. That is, the role of a particular sentence in an extract may not be consistent with the roles of its adjacent sentences.

A first simple solution to the problem of unresolved anaphoric references was given in the work reported by [Rush et al., 1971]. According to this approach, if a selected sentence contains an anaphor it is necessary to determine if the previous sentence (possibly the resolution of the anaphor) has been selected for inclusion in the extract, and to reinstate it if it has been rejected. If the restored sentence also contains an anaphor, the procedure must be repeated. But if several sentences (more than three) have to be reinstated because of the required antecedents of the initial sentence, then that sentence is rejected.

A different approach proposed by [Paice, 1990], deals with the detection of the unresolved anaphoric references. By applying a set of rules whenever an anaphor is recognised, this approach attempts to estimate whether the antecedent (i.e. the resolution of the anaphor) lies within the current sentence or not. According to this approach, sentences or passages that do not resolve anaphors within their own context are defined as '*not tidy'*. In order to produce a summary text which contains no dangling anaphors, when a 'not tidy' sentence is encountered, adjacent sentences are added until a 'tidy' passage has been constructed. If the passage becomes too long to be included in the summary, then it must be rejected. A point of criticism for this method is that the rules used are mostly empirical in nature and do not rely on grammatical characteristics of anaphoric words. For example, they usually rely on positional data: "if the word *these* is preceded by less than ten other words, its antecedent is assumed to lie in the previous sentence." It can be argued that the above rule is not on a firm linguistic ground, but nevertheless it provides a reasonable solution to the problem.

3.3.2 Coverage and balance

The lack of coverage and balance [Paice, 1990] of the resulting summary, is yet another weakness of extracting systems. The issue of *coverage* deals with containing every significant item of information in the generated summary. A document may have two, or even more, main topics. With the methods so far described it is relatively easy to include some of the main concepts, while omitting some others that may even be more important. The question of *balance* relates to presenting every important aspect of the source in the generated summary. A complete summary for a research paper for instance, should provide information about the "Purpose of the study", the "Procedures of research", the "Findings" and the "Conclusion". Preservation of summary's balance does not entail the identification of important information in these sections, it just ensures that all these aspects are covered. However closely related the two notions may seem, it should be made clear that coverage refers to the actual content of the source document, where balance actually pertains to its structural organisation.

In order to ensure that every important aspect of the source text is presented in the generated summary (i.e. the document's balance), various 'stylised' arrangements are utilised. The notion of the formatted abstract was introduced by [Paice, 1990]: headings such as "Purposes of study", "Procedures", and "Findings", can be helpful in identifying relevant portions of the summary of a research report, and also in achieving balance.

Obviously, it is not only summaries that possess a structure. Almost any piece of text can be analysed into a number of major components, which tend to occur predictably in texts of that type. This structure has been given the name 'superstructure' of the text. The superstructure of a summary should correspond rather directly to the superstructure of the original document, and this correspondence can be used to guide the composition of the summary. If all the elements of the source's superstructure are included in the summary, then balance can be ensured.

In order to tackle the issue of coverage, information within documents should be utilised. Such information can be provided by the document's sectional organisation and 'orientation' material. Sectional organisation of a document, although not always explicit, can provide useful information: different sections deal with different separate aspects of a document's message. One or more important concepts should be identified in each section to help focus the process of selecting extract-worthy sentences. The title method that was described in section 3.2 may be thought of dealing with the problem, but it only seeks for significant information at the section's heading. 'Orientation' material typically contains sentences that inform the reader of the document's structure. For example section 3.1 of this chapter contains orientation material that inform the reader for the contents of the chapter.

3.3.3 General comments

Some general remarks one can make on the automatic extraction methods, is that they are often heuristic and empirical in nature. Indeed, some of the methods described earlier have been formed through experimentation and intuition. Such an example is the *keyword* method as described in section 3.2, and more specifically the arbitrary choice of a threshold frequency value that constitutes a term significant. It is therefore clear that further research is needed in order to improve the linguistic basis of such methods.

Moreover, it can be argued that sentence extraction systems are actually ‘corpus-dependent’, that is they rely on the exploitation of special characteristics of the document corpus, and therefore their strategies are adjusted accordingly resulting in special extraction algorithms [Aretoulaki, 1997]. This kind of dependency should not be confused with a restriction of such systems to a specific application domain. It only allows for an ‘optimisation’ of the system on that specific document corpus. For example, given that a summarisation system will be applied on documents consisting of news articles, it would be fully justifiable to take advantage of their journalistic idiosyncrasy, and to give an extra ‘bonus score’ at the leading sentences of each document [Brandow et al., 1995].

A final remark about sentence extraction is that many of the criteria used to establish a measure of sentence significance rely on the assumption that the author of the document does reveal important information in specific parts of the document. Thus, the indicator phrase and cue methods assume that when the author uses specific linguistic structures he wishes to convey significant information; the title and location methods on the other hand rely on the structural arrangement of the presented information within the document. It can be argued that in some cases these assumptions could mislead the extraction methods.

3.4 Language generation approaches

An alternative to sentence extraction approaches for automatic summarisation is the use of natural language understanding and generation methods borrowed from the general field of Natural Language Processing (NLP) [DeJong, 1982; Jacobs & Rau, 1990; McKeown et al., 1995; McKeown & Radev, 1995; Aretoulaki, 1997]. Such approaches are different from, and certainly more complicated than, sentence extraction techniques. They involve a deeper and more extensive analysis of the input text in order to obtain its meaning. The representation of this meaning may then be manipulated in order to identify its more important constituents. Such approaches do not rely simply on keywords or phrasal patterns for the determination of importance, as is the case with sentence extraction methods. Full grammatical and semantic processing is usually involved, as well as syntactic information derived from analysis of the sentence structure. However attractive the idea of such systems may seem, they have so far proven capable of processing information only within a narrow domain whose characteristics are predictable and well understood. The main characteristics of such approaches will be discussed in the following sections.

3.4.1 Basic architecture of language generation systems

Traditionally, language generation systems are divided into two modules [DeJong, 1982; McKeown et al., 1995] : a *content planner* which attempts to interpret the source text by selecting information to include in the summary from an underlying knowledge base, and a *sentence generator* which takes the conceptual representation of text produced by the content planner and realises it in natural language. More specifically, the generator determines the sentence structure for each input proposition, selects the appropriate words, and orders them in a sentence by building syntactic structures and applying syntactic constraints.

The content planner uses conceptual facts to select information from an underlying knowledge source and to determine its overall organisation in the text. The output of the content planner is an intermediate representation of the source text. The intermediate form may range from surface representations like syntactic parse trees, to representations involving conceptual primitives [DeJong, 1982]. Approaches that demand domain knowledge are usually required in order to derive a conceptual representation of the source, thus enforcing domain dependency to such systems. Moreover, since the amount of knowledge that needs to be stored for robust processing is large, systems are usually compelled to specialise in a single conceptual world and application [Aretoulaki, 1997]. As a result, the flexibility of such systems is greatly restricted, along with their portability to other domains. Their components are not easily extended to accommodate additional cases or applications.

The facts that the content planner selects are subsequently passed to the sentence generator module, which generates a sentence for each fact. In order to do so, the sentence generator must decide on the sentence structure, choosing for example whether to generate a question or declarative sentence. It must also select a main verb, select words for each verb argument, build a syntactic tree for the sentence and enforce syntactic constraints. While the usual approach is to generate one sentence for each fact, it is possible that in limited circumstances systems can combine a few consecutive facts within a single sentence. Furthermore, since the production of connected text (and not simply single sentences) is required, issues of discourse coherency and structure should be addressed [McKeown, 1985]. A system employing text generation should be able to determine how to organise the individual sentences by adhering to a specific organisational framework. That framework would describe the discourse goals to be included in the sentences, and should provide the basis for producing coherent output text.

3.4.2 Some general remarks

The decomposition of language generation systems into the previously mentioned two modules follows the summarising model that was presented in paragraph 3.1.1 of this chapter. Clearly, the content planner interprets the source document by identifying significant information and by comprehending, to some extent, the source’s content.

The sentence generator is then responsible for the source text generation and the presentation of the summary to the user. This aspect of such systems, the actual generation of the text, is the most problematic one. Natural language interpretation requires examination of the evidence provided by a particular text, in order to determine the meaning of the text and the intentions of its author. Interpretation does not require a formulation of reasons for selecting between various options for the construction of the text (e.g. why in a specific point within a sentence active voice is used instead of passive) [McKeown, 1985]. In generation of natural language, however, that is exactly what is required. A generator must be able to construct the best expression for a given situation by choosing between many possible options. This requires utilising a wide range of knowledge sources that provide such options, and that store them for usage in a similar case in the future.

Finally, recent research [Enders-Niggemeyer et al., 1993] has suggested that automatic text summarisation should focus on the observation of human summarisation skills, and on linguistic work on discourse and text in order to derive a satisfactory theory of text structure. Observation of human skills can possibly assist in effectively modelling the process of summarisation, therefore directing automatic systems towards more flexible human-mimic approaches.

3.5 Evaluation

The question of how an automatic summarisation system should be evaluated seems to be a thorny one, since only few researchers have tried to address it. The difficulty arises from the very nature of a summary, as it is difficult to say what the properties of a ‘good’ summary are. In this section, the commonly adopted approaches for the evaluation of summarisation systems will be presented. Then a novel scheme for system evaluation will be proposed, and finally a presentation of a comparative evaluation of sentence extraction methods will be given.

3.5.1 System evaluation

The first attempt to carry out a thorough evaluation of a sentence extraction system was performed by H.P. Edmundson [Edmundson, 1969]. Since then, relatively little work has been carried out on the issue.

Edmundson's evaluation schemes were based on his belief that in any document there are sentences which should be included in every summary (right answers), and there are sentences which should not be included in any summary (false answers) [Edmundson, 1964]. He therefore reduced the problem of automatically extracting sentences to form a summary, to that of reducing the number of false answers to a minimum, while at the same time selecting as many right answers as possible. For his experiments, he used a target extract for each of his test documents which was prepared by human extractors. Each automatically created extract was then evaluated according to the percentage of sentences that were coselected in automatic and manual extracts, and the "*mean similarity rating*" which was based on subjective content judgement (how similar in content automatic and manual extracts were). The results of this evaluation, performed on a sample of 40 documents, showed that 44% of the sentences that the system selected were also selected by the human extractors, and that the mean similarity rating was 66%.

Later attempts to evaluate extraction systems [Brandow et al., 1995; Kupiec et al., 1995; Salton et al., 1997] do not differ significantly to that of Edmundson's. They try to compare the auto-extracts with some target extracts prepared by humans, or try to rate the auto-extract's acceptability by human judges. The acceptability rating is then compared to that obtained for an extraction system which simply outputs sentences in order, until a desired summary length is reached. This evaluation method showed that the leading-text summaries received significantly higher acceptability ratings (87% as opposed to 68% for the automatically generated summaries), suggesting the inadequacy of present extraction methods [Brandow et al., 1995].

In each case, the human factor becomes deciding, either in composing the target extract or in subjectively judging its appropriateness. In the case when auto-extracts are compared to manually prepared ones, the naive assumption is adopted that only one ‘correct’ summary exists for every document. To prove the contrary, experiments conducted on human abstractors [Edmundson, 1964; Kupiec et al., 1995] have shown that sentence extracts, or even paragraph extracts [Salton et al., 1997], selected by different persons have a very low level of agreement. A more striking observation is that for a given abstractor and for a given document to abstract over time, the overlap of extracted sentences is only 55%.

It would therefore seem appropriate to prepare more than one manual extract for each document. Different human extractors would produce different extracts, which would be compared to those automatically generated. An evaluation measure can then be computed by means of a function of the evaluation schemes mentioned previously. Nevertheless, even with this approach only quantitative features of the extracts can be measured.

3.5.2 A qualitative approach to evaluation

Motivated by the state of the art in evaluation schemes, the work reported in this thesis proposes a novel approach to the evaluation of summarisation systems. This approach aims at measuring qualitative characteristics of the auto-extracts by applying a task-based evaluation scheme. The proposed scheme should be performed in an end-user, operational environment, and should allow the integration of the summarisation module into an existing IR system. Evaluation measures would then result from the interaction of the user with the integrated IR-summarisation system, with the notion of the relevancy of a document to a specific information need being the principal one.

The evaluation scenario that is proposed in this thesis, can be briefly described as follows:

- The user queries the IR system.
- The system presents the answers to the query by providing the title and the first few sentences of the retrieved documents.
- At this point measurements of the user’s responses are performed. Such measurements should indicate the ability to identify relevant documents.

- The procedure should then be repeated with the automatically generated summaries being presented to a different user as an indication of the retrieved document’s content.
- Comparison of measurements should provide information about the *indicative* nature of the auto-summaries.

The model described above judges the utility of a summarisation system in the context in which it will eventually be used, and for the purposes for which it has been built. The *indicative* function of a summary is the one which is considered essential [Edmundson, 1969; Paice, 1990], and therefore is the one which should be primarily evaluated. Experiments with various users coming from various backgrounds (academic researchers, authors, simple end-users) should provide a sound basis for a fair evaluation scheme.

The need for a similar approach to evaluation has been acknowledged [Hand, 1997], but yet attempts to perform a task-oriented evaluation have not been well established. An early task-based evaluation attempt was made by [Miike et al., 1994], where timing statistics and relevancy decisions based on summaries for a domain-specific summariser were recorded. However, the presentation of the results was rather equivocal, and the experimental methodology followed was not clear.

3.5.3 Performance evaluation of the extracting methods

Edmundson extended his work on the evaluation of summarisation systems to the evaluation of the performance of the various sentence extraction methods [Edmundson, 1969]. He experimented with four distinct extracting methods: the keyword, cue, title and location methods. His approach involved adjusting parameters for these methods and then combining them in all possible ways. This resulted in fifteen auto-extracts being generated and scored for each document, one for each combination.

In order to rate the extracting methods, he followed the same procedure as in his evaluation method for summarisation systems: human experts prepared a manual extract for each document, and then scored each auto-extract according to how well its sentences matched the target sentences. In this way, the ‘rank order’ of the methods, in decreasing performance, was: location, cue, title, keyword. Furthermore, the combination of the cue, the title and the location methods seemed to have the highest mean sentence co-selection score.

The keyword method gave the poorest results among the extraction methods that were examined. A possible reason for this failure is the fact that the keyword method selects significant information more evenly throughout a text, whereas documents used in Edmundson’s experiment contain indicative material located at the beginnings and ends of text. This fact is a major point of criticism in Edmundson’s work: He only used documents from a single field (Chemistry), which held a specific structure.

The lack of progress in this area indicates that there should be more profound work on evaluating the extraction methods. Various extracting methods and combinations of methods should be tried, and evaluation processes such as those discussed for system evaluation, i.e. task based evaluation schemes, should also be applied.

3.6 Summary

In this chapter a critical overview of the research in the field of automatic text summarisation was presented. The two main categories of approaches to summarisation (sentence extraction and language generation) were discussed, and their most important problems were mentioned. The main aims of this discussion were to establish the ground for a number of issues that will be brought up in following chapters, as well as to provide an insight to the major implications of text summarisation.

The principal approaches to the evaluation of summarisation systems were also presented, emphasising on the inadequacy of the evaluation criteria employed. Through this presentation the evaluation approach followed in this thesis was outlined, thus justifying one of the aims of the research work undertaken.

Chapter 4

The Summarisation System

4.1 Introduction

This chapter describes the actual design and implementation of the system used for the automatic summarisation of documents. It would be beneficial for the purposes of the discussion presented here to initially give an outline of the architecture of the summarisation system. By gradually describing the various components of the architecture, the reader should attain an overall view of the system and its design. Having examined the constituent elements of the system, the rationale of its basic design is provided, justifying the principal design decisions taken. A sample output of the system is then presented, offering the opportunity for a discussion on some of its features. Finally, a summary of the issues discussed in the present chapter is presented.

4.2 System architecture

Figure 4.1 presents the main components and procedures of the system. In summary, the first task is to create the document collection on which the system will actually act. Once the document collection has been defined, the input text of the documents is broken into its constituent sentences. Subsequently, by using a combination of sentence extraction methods that utilise information both from the structural organisation of each document and from the distribution of terms within the documents, a score is calculated for each one of the sentences. The target summaries are required to be customised to the user's information need. Therefore, in addition to the various sentence extraction criteria, information about the queries is also taken into account when calculating the scores for the sentences. The system then outputs a predetermined number of 'top-scoring' sentences, that constitute the actual summary of the document.

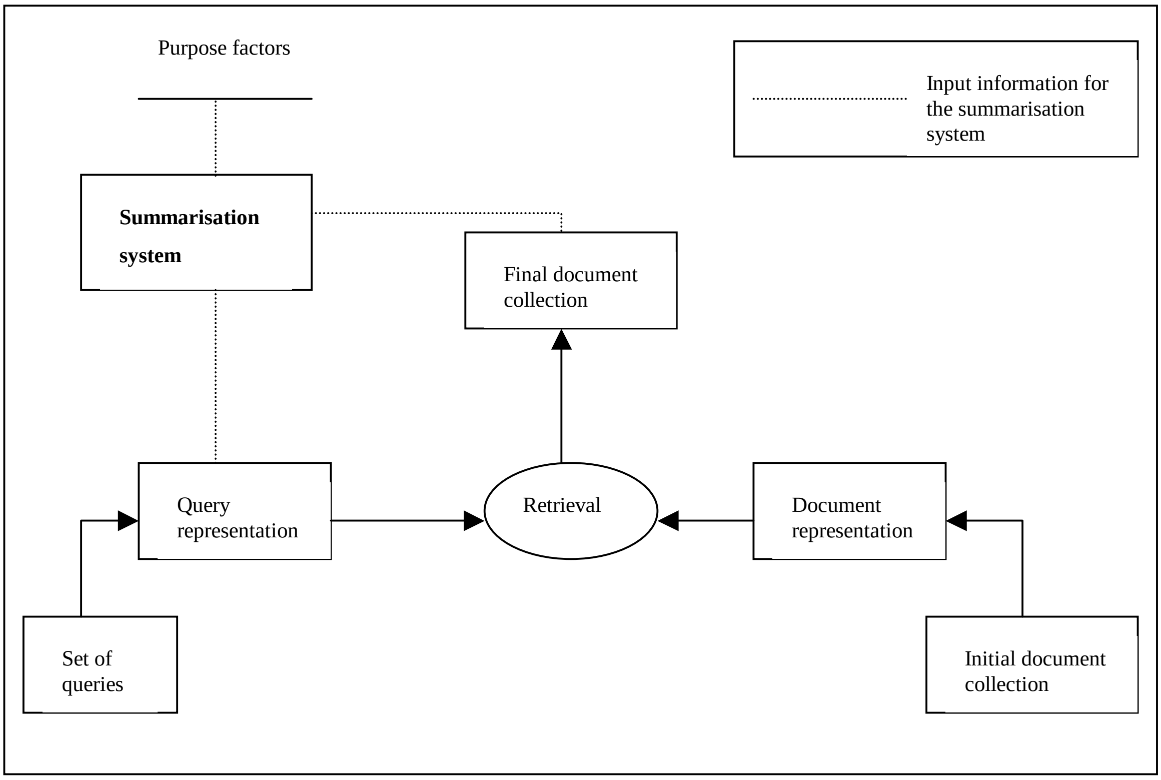


Figure 4.1 System architecture

4.2.1 Purpose factors

There are generally three factors that affect summarising: input, purpose and output [Sparck Jones & Enders-Niggemeyer, 1995]. Input factors pertain to the characteristics of the documents being summarised, purpose factors define the summary requirements, and output factors deal with the presentation of the summaries and are affected by both purpose and input factors. It would therefore be appropriate to view a summary as a function of both the type of documents to be summarised, and the purpose for which it is actually required [Edmundson, 1964; Rush et al., 1971].

Purpose factors have a direct effect on the process of summarisation. By defining the specific function a summary may have, the actual form of the summary is defined as well. For instance, a different summary form is required for informing a user about the relevance of a document to his information need, and a different one for providing the user with a clear view of the document's content thus enabling him to answer specific questions concerning the document. Clearly, in the former case a summary should be of an indicative nature placing emphasis on the user's information need, whereas in the latter case an informative summary that would virtually act as a surrogate for the whole document would have to be generated. The definition of these kinds of factors should precede the actual implementation of the system, providing a strongly defined framework around which the actual implementation of the summarisation system shall evolve.

On the other hand, the influence of the type of documents to be summarised is applied on the actual implementation of the system, and more specifically on the sentence extraction methods. Special features of the document collection usually drive the summarisation system towards customised algorithms that exploit these characteristics. Such effects will come up during the discussion of the sentence selection methods later in this chapter.

It is at this point appropriate to define the purpose factors for the summarisation system presented in this chapter, by describing the operational environment in which the system will be applied. The summaries generated by the summarisation system will aim at providing users with information on the relevance (or non-relevance) of documents retrieved in response to a query. It is assumed that users are willing to spend a limited amount of time to go through the list of the retrieved documents, and to decide on their relevance. Moreover the summaries should be customised to the specific query, thus reflecting the context in which the query terms are used within the document. Based on these requirements, we can conclude that the *intent* of the summaries should be indicative, helping users to judge the relevance of the original document in relation to a specific query, whereas the *focus* should be user-directed, biasing the summaries towards the query.

4.3 Methodology for generating the document collection

Having defined the purpose factors for the summarisation system, we can proceed in describing in detail the other parts of the system architecture depicted in Figure 4.1. The purpose of this section is to describe the methodology for generating the set of documents that the system will summarise. This task is performed through integration with SIRE (System for Information Retrieval Experiments), a modular IR system that is subsequently described.

4.3.1 System for Information Retrieval Experiments

SIRE is an IR tool kit, composed of separate modules (programs) communicating via a common protocol. This section will briefly describe the communication protocol, and will also mention the modules used for the integration of SIRE with the summarisation system.

Communication protocol

The protocol that all the modules of the system use on their input and output is a simple stream of tagged tokens. The tokens are written in plain ASCII text, one token per line with spaces and tabs used to separate the token’s components. Each token holds one piece of information (e.g. an index term) and any additional attributes of that information (e.g. the frequency of that term within the document).

```
a 2     21
b 5     banks
b 5     after
b 8     takeover
```

Figure 4.2 Example stream of tokens.

Figure 4.2 shows an example stream of four tokens. The first component of a token, a single letter, denotes the token's type: the 'a token' is a document start tag, 'b tokens' are the first three words of the text. The number immediately following a token type indicates the character length of the token's main attribute. There is a large number of tokens that SIRE uses, but it is beyond the aim of this paragraph to elaborate on them.

Modules

SIRE comprises a number of modules, simple independent programs, that perform atomic tasks. Although the actual number of modules included in SIRE is large, it would suffice for the purposes of the present discussion to refer only to modules that perform tasks relevant to the integration with the summarisation system. Such modules are:

- Tokenising modules: They parse a specific document collection or query type into a token stream.
- Stop word module: Removes any stop words from the token stream.
- Stemming module: Stems all words in the token stream.
- Indexing module: Indexes all words in the token stream, using document start tokens ('a tokens') as document boundary markers.
- Retrieval module: Based on tokenised query terms, this module performs a retrieval and outputs the document ranking list as a stream of tokens.

4.3.2 Integration with SIRE

The integration of the summarisation system with SIRE primarily involves using its indexing and retrieval modules in order to generate the document collection. The steps of this procedure are presented in the following paragraphs.

- The initial document collection, consisting of over 500 MB of news articles from the Wall Street Journal (WSJ) collection, was indexed using SIRE's indexing module (stop word removal and stemming were first applied). The documents of the WSJ collection possess a specific format that characterises their different sections (e.g. title section, main text section, date section). A section is defined as a portion of text between a pair of tag delimiters. For example the beginning of the body text of an article is marked by a <TEXT> tag, while its end by a </TEXT> tag; the main text of the article is the portion of text included between these two tags. A sample document from the WSJ collection is given in Appendix A**.**

- Fifty queries were randomly selected from a set of 250 queries, and they were tokenised by the appropriate tokenising module included in SIRE. Stop words were first removed, and all remaining query terms were stemmed. The queries act as an expression of a user's information need, on which the automatically generated summaries will be customised. Specific details about the type of queries used by the system will be presented in section 4.5.1 of this chapter.

- Based on the document index files and the tokenised queries, a retrieval was performed for each one of the fifty queries. The retrieved documents for each query were stored in a separate text file, so as to simplify any future accesses to them. It was decided that the final document collection should consist of the first 50 documents of the ranked list generated by each retrieval. Since users would have a limited amount of time to perform their relevance judgements on the documents, it would be superfluous to present them with more than fifty documents. The document collection that was so generated, comprised 2,220 news articles that occupied 12,5 MB of disk space.

- The document collection was accordingly tokenised using a modified version of SIRE’s tokenising module. In this way, information useful for the sentence extraction methods was generated. The next section analyses this procedure.

4.3.3 Extracting information from the collection

The summarisation system is based on a number of sentence extraction methods that attribute significance scores to each of the sentences of the input text. In order for these methods to function, they require access to certain information from the document collection. For this purpose, by using the tokenising module provided by SIRE, a modified version of an index file, thereafter called simply ‘*information file*’, is built. A sample part of the information file is given in Appendix B**.**

Based on the basic token protocol used by SIRE, the information file contains four types of tokens: Beginning of document (‘a’ token), its position in the file that the document collection is stored (‘h’ token), text token (‘b’ type), and title token (‘j’ type). There is a section in the file for each document in the collection. A section is identified as the part of the information file contained between two consecutive ‘a’ type tokens. Within the boundaries of each section, a token for each unique term of the document is written (stop words have been removed, and terms are in stemmed form).

Tokens of text terms are either ‘b’ type or ‘j’ type, depending on which section of the document they occur. The former token type characterises terms that occur in the body text section of the document (portion of text between the <TEXT> and </TEXT> tag delimiters), while the latter terms that occur in the title of the document (portion of text between the <HL>, </HL> tag delimiters). The additional information provided for each token is the number of the term’s occurrences within its containing document (TF[8]). This data is gathered by modifying the standard modules provided by SIRE: ‘b’ and ‘j’ type tokens are recognised during the parsing of the document collection by using the appropriate tokenising module, while TF values are gathered during the tokenising of

[8] For a brief discussion about TF measures, see section 2.2.2 of this thesis.

the document collection. In this way information about the distribution of the index terms within their containing document, as well as information about the location (title or body text) of each document term is made available to the sentence extraction methods. The way that this information is utilised by the summarisation system is the subject of the next section.

4.4 Sentence extraction methods

The information that is made available from the process described in the previous section, is accessed by the summarisation system in order to produce a score for each sentence of each one of the documents in the collection. This score is assumed to be indicative of the sentence's significance, and it will eventually, along with information from the queries, determine its presence or absence in the automatically generated summary. The sentence selection methods are discussed in this section, divided into methods that use clues from within the documents (structural organisation), and methods that use evidence from the distribution of the index terms over the documents. The way that information available from the queries is manipulated by the summarisation system, is presented in the next section of the present chapter.

4.4.1 Using evidence from the structure of the documents

The summarisation system must be able to identify and manipulate the constituent elements of each document, whether these are words, phrases or sentences. This is required in order to utilise information that is provided from the structural organisation of the documents.

Sentence parsing

The information file for the document collection provides data about the words that occur within each document. It is the responsibility of the summarisation system to parse each document into sentences, and to store necessary information about the sentences in an easily accessible data structure.

A sentence, as used by the summarisation system, is a string of words terminated by a period, a question mark, a colon, or an exclamation mark. The algorithm for parsing the documents into sentences looks for the terminating symbols in the text, and accordingly marks the limits of the sentences. For each sentence the following information is kept:

- Its beginning and end, in terms of actual position in the file that the document collection is stored.
- The location of the sentence within its containing document.
- A flag indicating whether a sentence is part of a section heading of the document.

Methods of sentence selection[9]

In order to be able to decide which criteria to use for utilising the information that is provided by the documents, a small scale study of the characteristics of the WSJ collection was conducted. The methodology that was followed, involved examining 50 randomly selected documents from the collection, attempting to extract conclusions about the distribution of important information within these documents. Titles, headings, the first few sentences of the documents, and their overall structural organisation were studied. Furthermore, this sample collection was used for experimentation with various system parameters, in order to approximate the best settings for the summarisation system. Although the sample of the documents was small, there was a strong uniformity in the characteristics of the sample that allowed for a generalisation of the conclusions to the entirety of documents in the collection.

It should be noted that two of the selection methods described in Chapter 3, namely the indicator-phrase and the cue methods, were excluded from the system design. The indicator-phrase approach is not entirely applicable to our document corpus, since the lexical constructs it employs are not usually met in news articles. The cue method on

[9] Details about the various sentence extraction methods can be found in Chapter 3 of this thesis.

the other hand, is not thought of as being able to contribute to the identification of sentence significance, since it is dubious whether 'bonus' or 'stigma' words can be identified in the context of news articles.

In the following paragraphs, the conclusions from the analysis of the documents will be related to the various sentence extraction methods that were described in the previous chapter of this thesis, in order to provide an insight to the sentence selection strategy of the summarisation system.

- **Title method**

  As mentioned earlier, the document collection consists of news articles. It is generally known that the titles of news articles tend to reveal the major subject of the article; they usually act as a preview to the whole article. This belief was strengthened by the sample study of the document collection: Titles in the WSJ collection tend to refer to the main subjects of the article.

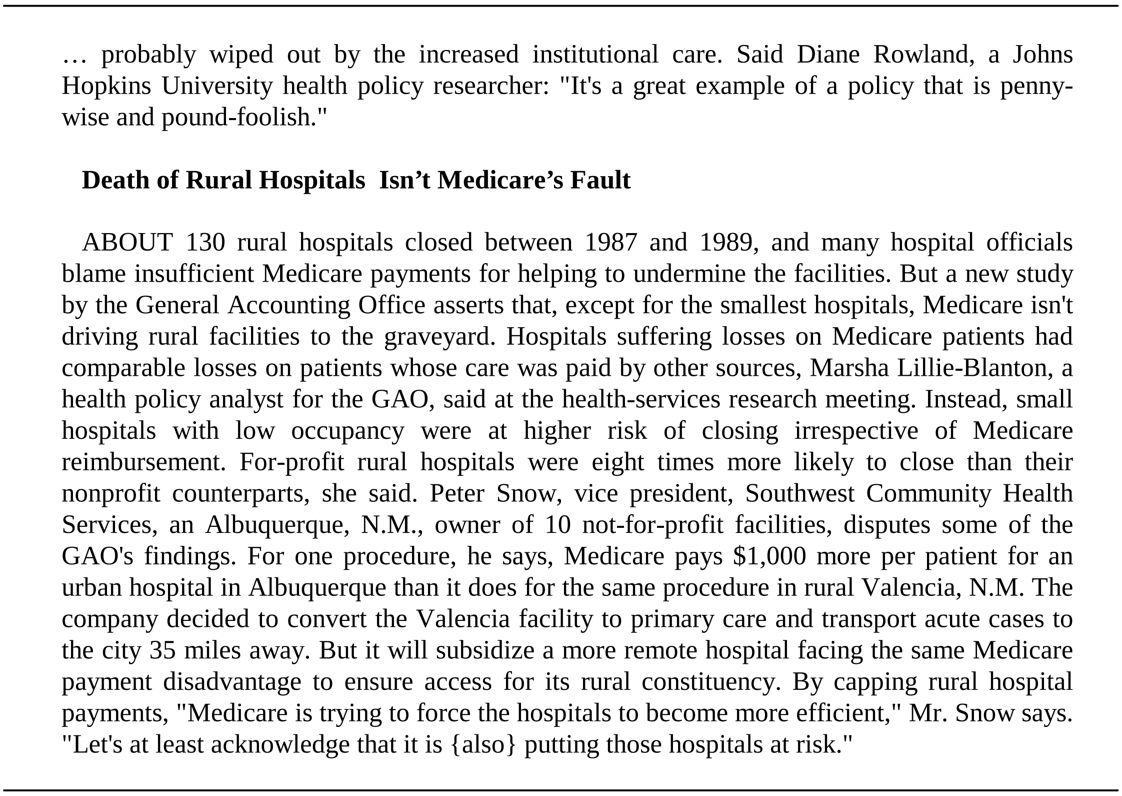
  In order to exploit this feature of the collection, terms that occur in the title section of the documents are assigned a positive weight (title score). The collection's information file allows for the identification of the title terms ('j tokens'). All the terms of a specific document are looked up in the information file (within the boundaries of the specific document). If a term is found in a 'j' type token, then it is assigned a positive 'title score'. The title score of each sentence is then defined as the sum of the title scores of its constituent words.

- **Location method**

  The location method [Edmundson, 1969] is based on the physical arrangement of the linguistic elements of an article. This arrangement can mainly be described in terms of the location of a sentence with respect to the limits either of its containing document, or of sections within that document.

  - *Leading sentences.* It was uniformly noted from the sample study, that the first few sentences of each article provide a fair amount of information about the article's content. This conclusion seems to be in agreement with [Brandow et al., 1995], who suggested that 'improvements (to the auto-summaries) can be achieved by weighting the sentences appearing in the beginning of the articles more heavily'. Bearing in mind the purpose factors mentioned earlier in this chapter, it is believed that the leading sentences of the articles are able to provide the user with a rapid overview of the document's content. In order to quantify this contribution, an ordinal weight is assigned to the first two sentences of each article. No other part of the articles, when examined relatively to their limits, seems to convey significant content information.

- *Paragraphs.* While it has been suggested that the first and last sentence of paragraphs often act as a form of summary for the specific paragraph [Edmundson, 1969] and should therefore be assigned a positive score, this is not the case in the WSJ collection. Its documents are fragmented into a large number of paragraphs, that usually consist of only a couple of sentences. Thus, it does not seem appropriate to further examine the division of the document text into paragraphs since this division carries no semantic information.
- *Headings.* Heading sections within documents on the other hand, provide evidence about their division into meaningful semantic units. This was a uniform conclusion obtained from the sample study of the WSJ collection. In a similar way that the title of an article is indicative of its content, the heading of a section is revealing of its principal information. In Figure 4.3 an example of a section heading, and of its function within the document is given. In order to exploit the evidence provided by section headings, a 'heading score' is assigned to each one of the sentences comprising a heading. Heading sentences are identifiable by the respective flag in the structure that keeps information about the input sentences.



… probably wiped out by the increased institutional care. Said Diane Rowland, a Johns Hopkins University health policy researcher: "It's a great example of a policy that is penny-wise and pound-foolish."

Death of Rural Hospitals Isn’t Medicare’s Fault

ABOUT 130 rural hospitals closed between 1987 and 1989, and many hospital officials blame insufficient Medicare payments for helping to undermine the facilities. But a new study by the General Accounting Office asserts that, except for the smallest hospitals, Medicare isn't driving rural facilities to the graveyard. Hospitals suffering losses on Medicare patients had comparable losses on patients whose care was paid by other sources, Marsha Lillie-Blanton, a health policy analyst for the GAO, said at the health-services research meeting. Instead, small hospitals with low occupancy were at higher risk of closing irrespective of Medicare reimbursement. For-profit rural hospitals were eight times more likely to close than their nonprofit counterparts, she said. Peter Snow, vice president, Southwest Community Health Services, an Albuquerque, N.M., owner of 10 not-for-profit facilities, disputes some of the GAO's findings. For one procedure, he says, Medicare pays $1,000 more per patient for an urban hospital in Albuquerque than it does for the same procedure in rural Valencia, N.M. The company decided to convert the Valencia facility to primary care and transport acute cases to the city 35 miles away. But it will subsidize a more remote hospital facing the same Medicare payment disadvantage to ensure access for its rural constituency. By capping rural hospital payments, "Medicare is trying to force the hospitals to become more efficient," Mr. Snow says. "Let's at least acknowledge that it is {also} putting those hospitals at risk."

Figure 4.3 Section headings in the WSJ collection.

4.4.2 Using term frequency information

In addition to the evidence provided by the structural organisation of the documents, term frequency information about document terms is also exploited in order to determine the significance (or better, the extract-worthiness) of the sentences of the input text. The approach followed is influenced by the work reported by Luhn in 1958. Instead of merely assigning a weight to each term according to its ‘resolving power’, the method employed attempts to locate clusters of related terms with significant ‘resolving power’. In order to actually implement this method, three issues have to be clarified: How to determine a term’s discriminating power, how to define the clusters of related significant terms, and finally how to assign a weight to each sentence according to the compiled information.

- **Term significance**

  The issue of the identification of significant words within a document for the purposes of text summarisation was first investigated by Luhn in 1958. Utilising term frequency (TF) data for each index term, he generated a ranked list of the document terms, in order of decreasing frequency. He then concluded that useful index terms are those who possess a medium ranked TF value. Terms with high frequency of occurrence are excluded as ‘noise-words’ (stop words), and terms with low frequency are treated as rare words that convey no useful information for their containing document. The justification for this approach is based on the fact that the author of an article normally repeats certain words as he advances or varies his arguments, and as he elaborates on an aspect of a subject. This method can be further supported by a general ‘principle of least effort’, according to which it is easier for a writer to repeat certain words instead of coining new and different terms.[10]

  In the approach followed there is no need to define an upper limit for the frequency of significant words, since stop words have been removed from the document corpus. Still, a lower limit has yet to be defined. In order to define that limit, a fair amount of experimentation was performed on the sample collection of the 50 documents. However, the results obtained through this kind of experimentation can not be guaranteed to establish an optimum value for the lower limit. This can only be achieved by experimentation with appropriately large samples of documents [Luhn, 1958].

  The conclusions from the experimentation procedure were that a reasonable TF value for establishing the significance of a term is 7, and that this value should be adjusted according to the size of the document. The value of 7 should be applied to medium-sized documents of the collection. Such documents are those whose number of sentences is between 25 and 40. These numbers were obtained through the analysis of the 50 documents comprising the sample collection. For documents that contain more than 40 sentences, the TF value is augmented by 10% of the increase in document size. The increase is calculated in respect to the upper limit of the medium

[10] See [Salton & McGill, 1983], pp. 59-60.

document size, i.e. 40. For example, for a document that is 50 sentences long, the increase in size is 10, and therefore the TF limit is set to: 7 + [0.1 * (10)] = 8.

For documents smaller than 25 sentences, the same procedure should be applied, calculating the decrease in document size in respect to the lower limit of the medium document size (i.e. 25).

- **Clusters of significant words**

  The extension of the significance measure from single terms to clusters of related terms is based on the assumption that the more often certain words are found in each other's company within a sentence, the more significance may be attributed to each one of these words. Therefore, wherever the greatest number of frequently occurring different words are found in greatest physical proximity to each other, the probability is high that the information being conveyed is most representative of the article [Luhn, 1958]. The actual definition of the 'greatest physical proximity' is a major aspect of this method. Luhn suggested that a 'useful' limit is four or five non-significant words between significant words. This early observation seems to agree with more recent studies that show that in the English language 98% of the lexical relations occur between words within a span of 5 words in a sentence [Abracos & Lopes, 1997].

  Based on these observations, the method used in the work reported here defines two words as being significantly related if both of them are significant, and between them no more than 4 non-significant words intervene. If in that way a sentence is separated in two or more clusters, the one with the highest significance factors is taken as the measure for that sentence.

- **Sentence significance**

  A scheme for computing the significance factor for a sentence was given by [Luhn, 1958]. This scheme consists of defining the extent of a cluster of related words (i.e. the actual number of words in the cluster), and dividing the square of this number by the total number of words within this cluster. Consider the following example sentence:

‘I entered the [**place**, and **sat down** for a **little** while to **rest**] and get my breath again.’

For the purposes of this example, let us assume that words in bold are significant words. The portion of the sentence included in brackets is the actual cluster of significant words. The score for this sample sentence would be defined as the square of the number of bracketed significant words (25) divided by the total number of words in the bracket (10).

4.4.3 Integrating the information

Table 4.1 sums up the various sentence selection methods that were discussed in the previous paragraphs. A score for each sentence of the input text is calculated by summing up the partial scores assigned by the four selection methods.

Method	Evidence used	Sentence score
Leading text	The leading sentences of each document convey useful information.	Assign an ordinal weight to the first two sentences of each document.
Section headings	Headings of article sections reveal their principal content.	Assign a ‘heading score’ to sentences that comprise a section heading.
Title	The title of each article is indicative of its content.	Assign a ‘title score’ to each term occurring in the article’s title. Sentence ‘title score’ equals to the sum of the ‘title scores’ of its constituent terms.
Term frequency	Locate clusters of frequently occurring words within each sentence.	Calculate a score by dividing the square of the number of significant words within the cluster, by the total number of words within that cluster.

Table 4.1 Overview of the sentence selection methods employed.

The sentence extraction methods have been implemented in a way that ensures their inter-independence. Furthermore, the way that the selection methods are embodied in the summarisation system is modular. As a consequence, any number of further extraction methods can be added to the system (or any of the existing methods can be removed) without the need to modify the overall architecture. This feature of the summarisation system is extremely useful since it allows for experimentation with the various sentence selection methods, and the investigation of the effects they have on the automatically generated summaries.

4.5 Query customisation

In Chapter 3 of this thesis, the incoherent nature of the summary text was identified as the main problem that sentence extraction methods suffer from. This deficiency has usually negative effects in the acceptability of automatically generated summaries by users [Brandow et al., 1995]. However, when using a summarisation system in an operational, task-based environment, a possible way to alleviate the consequences of the incoherent summary text is by customising the summary to an information need expressed by a user through a query.

The rationale of this assumption is rather simplistic. When users are required to perform a specific task assisted by the auto-summaries (e.g. identify relevant documents to a specific query), their primary goal is to accomplish that task. They attempt to locate clues in the summaries that are relevant to the task, and that will support them on their judgement. In the specific example of deciding on the relevance of the documents to a specific query, users will attempt to comprehend the context in which the terms of the query are used in the documents. Legibility problems of the auto-summaries can then be of minor importance, since they do not relate to the user’s primary goal.

The purpose of this section is to provide an insight to the customisation of the summaries to a specific information need, since summaries generated by the system described in this chapter are required to be query-biased. Initially a brief description of the types of queries used by the system will be given, and subsequently the actual way by which information from the queries affects the process of summarisation is presented.

4.5.1 TREC queries

The queries used by the summarisation system are of a specific type, and it is essential that their characteristics are clearly defined before discussing their manipulation by the summarisation system. The specific type of queries used, is that employed in the TREC programme (Text REtrieval Conferences) [Sparck Jones, 1995]. Briefly, we can say that TREC can be viewed as an evaluation 'exercise' for existing IR systems, that attempts to measure specific aspects of their effectiveness. The effectiveness measures for the systems are obtained by ranking their performance in response to a set of queries, known as TREC topics.

In Figure 4.4 a sample TREC topic is depicted. As this example makes clear, TREC queries are long and detailed, and much more elaborate than normal information requests are. The TREC requests have been carefully formulated, and they possess a complex structure with several distinct fields. The 'Title' field for example, can be viewed as the exact query entered in an IR system by the user. The 'Narrative' field on the other hand, clearly indicates what properties documents must have in order to be deemed relevant, thus reflecting the user's needs [Sparck Jones, 1995].

The approach that was followed for generating the query terms for each topic, is based on the 'Title' section of the queries only. While it is possible to exploit information from the 'Description' and 'Narrative' sections as well, this is regarded as not being representative of the user's request as expressed through a query. As was explained in the previous paragraph the role of the 'Title' section can be identified as that of the actual query being input by a user in an IR system, whereas the other sections provide additional information that is not directly involved with the input of the query by the user. Thus, a representation of the query with terms only from the 'Title' section can be thought of providing a reasonable approximation to the user's information need.

<top>

<num> Number: 033

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description: Document must report on how the importation of foreign textiles or textile products has influenced or impacted on the U.S. textile industry.

<narr> Narrative: The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. "Textile industry" includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.

</top>

Figure 4.4 A sample TREC topic

4.5.2 Using evidence from the queries

The process of using the evidence provided by the queries consists of three phases: Indexing (or better, tokenising) the queries and making the tokenised form available to the system architecture, deciding how to utilise the query information, and finally calculating a 'query score' for each one of the sentences of the input text. We shall now examine each one of these steps.

- **Tokenising the queries**

  In section 4.3.2 of this chapter, a brief reference was made to the process of tokenising the queries that were used by the summarisation system. As was mentioned and justified previously, only the 'Title' section of each topic was indexed by SIRE. This was done using the appropriate tokenising module that is included with the IR system. This tokenising method yielded a small number of query terms per topic, the average number was 3.5 query terms per topic. There were 50 queries used by the system, and they were randomly selected from 250 available TREC topics. In order for the summarisation system to be able to access the tokenised form of the topics, a separate file for each query was generated. This file contains the tokenised form of the respective topic, and it facilitates the access to the tokenised data on a 'per query' basis.

- **Using the query information**

  The first step for actually customising the summaries to the queries, is to determine which of the documents of the collection are retrieved by which query. Let us consider a hypothetical document A of the collection. This document must have been retrieved by at least one of the 50 queries within the first 50 ranked documents, otherwise it would not have been included in the collection[11]. It is however often the case that a document is retrieved by more than one query[12]. For the purposes of our example, lets assume that document A has been retrieved by only one query, Q_1. Therefore, for document A, a summary must be produced in relation to query $Q_{1.}$ The information about the actual queries that each document is retrieved by, is stored in a text file that is readily accessible by the summarisation system.

  At this point, all the information needed by the summarisation system is available: for each document in the collection, the topics by which it is retrieved and the terms for these topics are known. The integration of this information is then required, in order to derive a score for each sentence indicative of the distribution of query terms in it. The algorithmic description of this procedure is presented in the subsequent paragraphs, continuing the example of document A cited above.

  - The file containing the information about the topics by which document A is retrieved is read. The topic is Q_1.
  - The file containing the terms for Q_1 is then read, and thus the query terms are made available to the summarisation system.
  - Subsequently, each constituent word of each one of the sentences of document A is looked up in the list of the query terms. If it matches one of the query terms, its position in its containing sentence is marked.

[11] See section 4.3.2 of this chapter.

[12] This is actually the reason why the collection comprises 2,220 documents instead of 2,500 (50 queries * first 50 ranked documents for each query).

  - The output of this procedure is a structure that provides information about which sentences of document A contain query terms, from which query these terms come from (in our case, only Q_1), how many query terms are included in each sentence, and finally their positions within these sentences.

- **Calculating a 'query score' for each sentence**

  Based on the information provided by that structure, a query score for each of the sentences of the input text can be calculated. For the computation of that score, the only information that is taken into account is the distribution of the query terms in each sentence. More specifically, the calculation of the score is based on the belief that the largest the number of query terms that occur in one sentence is, the more likely it is that this sentence conveys a significant amount of the query's content. Hence, an adequate indication of the context in which the query terms are used in a specific document can be provided by such sentences.

  The actual measure of significance of a sentence in relation to a specific query, can be derived by dividing the square of the number of query terms included in that sentence, by the total number of the terms of the specific query. Furthermore, for each sentence of the input text the system calculates as many 'query scores' as the number of the topics the document is retrieved by.

4.6 Generating the summaries

In section 4.4, a score for each input sentence was calculated according to information gathered from the structural organisation of its containing document, and from frequency data of its constituent terms. In the previous section, a score was assigned to each input sentence according to the distribution of query terms in it. The final 'significance' score for each sentence is defined as the sum of the two partial scores. A summary for that document is accordingly comprised of the top-scoring sentences of the input text.

In order to illustrate the process of the actual generation of the summaries, let us consider once more the example of document A, which is retrieved by topic Q_1. Let us assume that a sentence of document A has been assigned a partial score S_1 from the process of sentence selection, a score S_{21} from the process of query customisation in relation to topic Q_1. The final score for that sentence can therefore be defined as:

$SSi_1 = S_1 + S_{21}$ (score of sentence i in relation to topic 1)

The same calculation is repeated for each one of the sentences of document A, yielding a list of scores. This list is subsequently ordered in decreasing order of the sentence scores. The summary for document A is then generated by outputting the top-scoring sentences, until a desired summary length is reached. The summary length is automatically calculated, its value being a function of the document length.

4.6.1 Summary length

It was mentioned in section 4.2.1 of this chapter that purpose factors actually determine the form of the generated summaries. Clearly, the length of the automatically generated summaries is strongly related to their form, and it should therefore be affected by purpose factors. This is the case; taking into consideration the specific purpose for which the summaries are generated (to support a relevance decision), and the specific operational requirements (limited time to accomplish the relevance decisions), we can safely place an upper bound to the value of the summary length. This limit restricts the summary length to not becoming greater than six sentences. It is the author’s belief that given the specific operational circumstances, that value provides a reasonable amount of evidence to the user to accomplish his primary goal - the identification of relevant documents to a specific information need. The value for the summary length is then defined as the 15% of the number of sentences of the input text, unless that value exceeds the defined upper limit. In that case, the summary length is defined as equal to the upper limit (six sentences). Such a value seems to be in general agreement with suggestions made by [Edmundson, 1964], and [Brandow et al., 1995].

4.7 Rationale of the system design

The previous sections of this chapter have provided an insight to the architecture, and the principal design issues of the summarisation system. Having attained a comprehensive overview of the system, it is appropriate to provide the rationale of the major design decisions. These are discussed in the following paragraphs.

4.7.1 Choice of summarisation approach

The most important design decision taken pertains to the actual summarisation strategy that was followed: a sentence extraction approach was chosen instead of a language generation one. There are basically two main reasons for that decision. First, as was discussed in Chapter 3 of this thesis, language generation systems suffer from domain dependence; they are capable of processing information only within a domain whose characteristics are predictable. Such a characteristic was thought of as not being advantageous for the summarisation system used here. The second reason was that the actual implementation of a language generation system is complicated, and involves dealing with a number of research issues from various fields (e.g. text theory, discourse structure and coherence). Taking into consideration the time limits for the completion of the work reported in this thesis, it would be extremely difficult to produce a fully functioning summarisation system based on language generation methods.

4.7.2 Sentence selection methods

The actual decision about which sentence selection methods to use is another major decision for the design of the system. The reasoning for the choice of the employed methods was provided in the respective sections of the present chapter. However, it is obvious that a great number of combinations of these methods can be achieved, for example by weighting more heavily one method, or by adjusting parameters for another. Intensive experimentation was performed with the various parameters on the sample collection of the 50 documents that was used for ‘performance tuning’ purposes: Different weights were assigned to each one of the selection methods, certain criteria were excluded from the methods and the effect on the generated summaries was observed, and experimentation with the actual features of each selection made was also performed.

This experimentation process provides clues for an approximation of an effective setting of the various system parameters. It is not argued that this setting is the optimal one; it relies on subjective judgements performed by the author, and on intuitions derived from the study of the document corpus. Nevertheless, the evaluation of the system, that is discussed in the following two chapters, provides strong evidence that the derived settings were indeed effective, given the operational environment that the summaries were used in.

4.7.3 Term frequency considerations

A mere measure of the within-document frequency (TF) of each term was used in order to establish its significance. Other measures, such as the IDF measure discussed in Chapter 2 of this thesis, were not used.

The main reason for doing so is that the TF measure provides a reasonable indication of a term’s significance for the purposes of text summarisation. The TF measure is not extremely efficient in identifying index terms that distinguish the documents to which they are assigned from the remainder of the corpus, since it does not take into account information from the entire document collection [Salton & McGill, 1983]. However, this is of minor importance for the purposes of text summarisation. In such a case, what is wanted is not the identification of ‘useful’ index terms in relation to the entire document corpus, but the establishment of a term’s significance within its containing document. Hence, the use of the TF measure, as justified in section 4.4.2, is fully appropriate. Furthermore, by attempting to locate clusters of significant words for attributing significance to sentences instead of single words, the possibility that misleading terms will be over-valued by the summarisation system is reduced.

4.7.4 General comments

Throughout the discussion presented in this chapter, most of the problems of sentence extraction identified in Chapter 3 of this thesis were not addressed (e.g. coherency considerations, coverage and balance issues). This approach was based on the intuition that given the specific purpose factors that the system would be biased by, there was not a particular need to concentrate research effort on tackling these issues. Instead, emphasis should be placed on informing the user about the relevance of the retrieved documents to his request for information.

It was mentioned in section 4.5 that coherency problems can be tackled by the customisation of the summary to a specific information need. Moreover, it may often be the case that a document can be deemed as relevant to a query only by the presence in its body of two independent sentences that convey relevant to the query content. In that case, the summary for that document can possibly adequately inform the user on the relevance of the document, if it successfully identifies these two sentences. Issues of coverage and balance then become of secondary importance. The results of the evaluation that will be presented in the following chapters, provide supporting evidence for this intuition.

4.8 Sample output of the system

It was mentioned earlier in this chapter, that it is possible for a document of our collection to have been retrieved by more than one queries. In fact, the sample document presented in Appendix A is such an example: that document has been actually retrieved by two queries. For the purposes of the discussion, we shall give the ‘Title’ sections of these queries:

Q_1: *Welfare Reform*

Q_2: *Impact of the 1986 Immigration Law*

For that specific document, two summaries have been generated, one for each query. It would be interesting to examine these two summaries (query terms are in bold typeface).

S_1 (customised to query Q_1)

Just as countries compete in a world-wide market where goods and services are exchanged, they also compete in the market for people.
The U.S. accepted about 600,000 legal immigrants annually during the 1980s, not including the 2.5 million persons who applied for amnesty under the provisions of the 1986 Immigration **Reform** and Control Act.
Since the immigration **reforms** of 1965, U.S. immigration law has encouraged family reunification and discouraged the arrival of skilled immigrants: 75% of legal immigrants in 1987 were granted entry because they were related to an American citizen or resident, while only 4% were admitted because they possessed useful skills.
The differences in poverty and **welfare** recipiency rates among national origin groups are huge.
Before **welfare** benefits became widely available in the 1960s, prospective immigrants to the U.S. would make their decision based on a comparison of the economic opportunities available to them here and in their country of origin.

S_2 (customised to query Q_2)

Just as countries compete in a world-wide market where goods and services are exchanged, they also compete in the market for people.
The U.S. accepted about 600,000 legal **immigrants** annually during the 1980s, not including the 2.5 million persons who applied for amnesty under the provisions of the **1986 Immigration** Reform and Control Act.
Since the **immigration** reforms of 1965, U.S. **immigration law** has encouraged family reunification and discouraged the arrival of skilled **immigrants**: 75% of legal **immigrants** in 1987 were granted entry because they were related to an American citizen or resident, while only 4% were admitted because they possessed useful skills.
While the predicted lifetime earnings -- the best indicator of the value of the technical and intellectual skills that **immigrants** bring with them -- of the **immigrants** who entered the U.S.in the late 1970s were 30% lower than those of American natives, **immigrants** to Canada who arrived in Canada during that period are projected to earn only 13% less than Canadian natives.
For instance, the predicted lifetime earnings of **immigrants** from Canada, Germany or Britain are about 20% higher than those of natives; **immigrants** from India or Korea earn about 7% less than natives; and **immigrants** from the Dominican Republic, Jamaica or Mexico earn 30% to 40% less than natives.

The bias imposed on the summaries by the query terms is apparent, especially in the case of summary S_2. What is interesting to note is the fact that although the word *immigrant* was not actually included in the query, it was identified as a query term, and therefore affected the extract-worthiness of sentences containing it. The reason for this treatment of the word *immigrant*, is that it possesses the same stem with the actual query term *immigration.* Since stemming is applied on the queries, it is often the case that non-query terms are treated as query terms. However, it can be argued that this is in some cases beneficial, as in the example we are studying. The use of the word *immigrant* in the summaries gives us quite a clear view of the wider use of the notion of immigration in the context of the document. In some cases though, this treatment of stemmed query terms can be misleading. Consider the words *politics* and *polite.* No semantic relationship exists between them, and yet they possess the same stem: *polit.*

In both summaries the two leading sentences of the original text were included. It is quite evident that they provide some information about the overall content of the document. This is even more evident in the case of S_2, where the second sentence establishes rather clear the relevance of the document to query Q_2.

Another interesting point is the observation that summary S_1 establishes rather clearly the non-relevance of the document in relation to query Q_1. This is achieved by clarifying the context in which the query terms, and more specifically the term '*reform*', are used in the document. Hence, a user would be adequately informed on the fact that the document refers to *reform* in relation to immigration, and not in relation to welfare.

Through these two examples, the principal rationale of the designed summarisation system can be discerned: emphasis is placed on the establishment of the context in which the query terms are used in the documents. Moreover, the 'bonus-weighting' of the leading sentences of the text can provide users with an overall view of the generic document content. The other sentence extraction criteria are mostly used in 'tie-breaking' situations of equally extract-worthy sentences, or in cases where query terms are widely dispersed in the document text and provide no significant information.

4.9 Summary

The architecture of the summarisation system was outlined in this chapter, describing the method by which the summaries are generated. The actual reasoning of the sentence selection methods employed in the system, and the way that information from the document collection was utilised in order to identify the significance of sentences, was provided. Subsequently, the customisation of the summaries to specific information needs was described, presenting the way that the system incorporates query-biased information in the significance score for each input sentence.

Emphasis was placed on the fact that the process of summarisation is affected by both purpose factors, and special characteristics of the document corpus. Through this point of view, the main design decisions affecting the summarisation system were discussed. It should be noted that many of the decisions taken during the design of the system were based both on a small scale analysis of the characteristics of the document collection, and on empirical observations and intuitions derived through experimentation with a sample document collection. However, the experimental results presented in the following two chapters, obtained from the evaluation of the system, provide strong supporting evidence for the overall design of the summarisation system.

Chapter 5

Experimental design

5.1 Introduction

Until now the two major motivations for the research work reported in this thesis have been examined separately. Thus, in Chapter 2 it was suggested that the majority of IR systems do not present users with enough information about the relevance of retrieved documents to their information need. Motivated by this observation, the reported work proposes to present an automatically generated summary for each retrieved document, customised to the query input by the user. By this approach, an implicit assumption is made that the presence of the automatically generated, query-biased summaries, is expected to have an effect on the process of relevance judgement. However, the way that this assumption would be tested was deliberately not elucidated.

Subsequently in Chapter 3, while discussing the major issues of text summarisation, the need for a task-based evaluation scheme for summarisation systems became apparent. A brief outline for an evaluation scheme was then proposed, and its relation to the approach proposed in Chapter 2 was sketched: the testing of the implicit assumption made in that chapter can be embodied in the proposed evaluation scheme.

The primary aim of the present chapter is to provide the reader with an insight to the integration of the research work reported in this thesis. This is achieved through the description of the experimental procedure used to test the assumption that was stated previously. The experimental procedure on its own is rather simple. What needs to be clarified is that the function of the experimental method described here is more of a linking nature. It integrates the work reported in Chapter 4 (the query-biased summaries) into the task-based evaluation scheme for summarisation systems that was outlined in Chapter 3, in order to test the validity of the original assumption.

Initially, the main issues of experimental methodology that pertain to the testing procedure followed will be presented, in order to establish the terminology and the theoretical background for the discussion that is subsequently evolved. Following this brief introduction to experimental design, the specific methodology adopted shall be presented: the treatment of the issues mentioned in the introductory section will be described, and the actual experimental scenario shall be detailed. What is more important is that during this discussion the connection of the experimental procedure with the proposed evaluation scheme will be established. Finally, the user interface that was developed for the purposes of the experiment is presented.

5.2 Fundamental experimental design concepts

A science is built upon a large body of reliable facts and information. These facts are not easy to come by; they are established through observation, recording, and analysis of data generated during the observation periods [Keppel, 1973]. A common method for establishing facts is the *experimental method*. Little is known about the cognitive aspects of the process that leads to the formulation of a *hypothesis* through the observation of facts. Generally, such statements usually possess an intuitive nature; in fact very few hypotheses are actually formulated by direct deductions from some more general theory [Miller, 1984]. They mostly arise from some kind of interaction between the experimenter's intuitions, theoretical ideas and factual knowledge.

The best way to elaborate on the basic concepts of the experimental method is by means of an example. It would be convenient for the purposes of this chapter to examine the way that the implicit assumption of Chapter 2 would be tested by means of an experimental method. Let us state the assumption adopted: 'The presence of automatically generated summaries, customised to a specific query, for each retrieved document in a ranked list, is expected to have a positive effect on the process of relevance judgement by the users'. This is the observation we wish to establish, this is the *research hypothesis* we wish to examine through experimentation.

5.2.1 Experimental variables

The experimental method consists of the contrast between two treatment conditions. The subjects in these two conditions are treated identically, except for one feature that is different. This difference is termed as the *independent variable.* Some aspect of the performance of the subjects in the two treatment conditions is measured and recorded after the independent variable has been administered. This feature of the behaviour of the subjects is referred to as the *dependent variable.* In our case, the independent variable is the presence of the auto-summaries in the ranked list of the retrieved documents, while the dependent variable is the user's performance in the process of relevance judgement. Any difference we observe on the dependent variable is called the *treatment effect*, and is usually assumed to have been caused by the independent variable. Moreover, we are only interested in examining cases where the independent variable takes two values - in our example, the presence or the absence of the summaries in the ranked list of the retrieved documents. In this way, a group of subjects can be assigned to each one of the levels of the independent variable.

The assumption that the treatment effect is caused only by variation of the independent variable may lead to the erroneous conclusion that there are no unwanted variables in an experiment, and that all variation in the subject's performance will be caused by changes in the independent variable. Unfortunately, this is not the case. In fact, many experiments have been rendered useless through a breakdown in the chain of logic between the initial assumptions and the final conclusions due to factors introduced which vary systematically with the different experimental treatments [Keppel, 1973]. These factors, or variables, are called *irrelevant* or *confounding* variables.

Consider the hypothesis we are testing: we will eventually have to use a number of subjects to test the validity of our hypothesis. It would be rather difficult to ensure that all subjects will have the same level of experience with IR systems, posses the same educational background, or even that they are of the same intellectual level. Furthermore, external to the subjects factors may as well uncontrollably vary: the equipment used for the experiment could have different settings for different users, or the room conditions may be favourable for some users (e.g. quiet environment) and not favourable for some other users (e.g. external noise, even extreme change in the room temperature). It would be naïve to underestimate the effect that these variables can have on the experimental procedure. The former category of irrelevant variables, the one pertaining to the characteristics of the subject population, is termed *subject variables*, while the latter is known as *situational variables.* Methods for controlling these two categories are briefly presented in the following paragraphs.

5.2.2 Control of irrelevant variables

The control of subject variables pertains to ensuring that the groups of subjects tested under each experimental condition are as similar as possible on all the dimensions along which people can differ. This can be achieved by controlling the way in which subjects are allocated to the experimental conditions. Three methods are primarily used for this kind of control[13]: the *repeated measures design*, which requires the assignment of each subject to both experimental conditions; the *matched subjects design*, according to which we must select pairs of similar subjects[14] and assign each member of the pair to a different experimental condition; and finally the *independent groups design*, which requires the division of the subjects into two entirely separate, randomly selected groups. It is obvious that these methods attempt to counterbalance the effects of the subject variables using a different approach. However, in deciding which method to use in an actual experiment, other considerations become significant. For example, the

[13] For a comprehensive discussion on these methods, refer to [Miller, 1984], pp. 10-15.

[14] ‘Similar’ on the variables that influence the condition under study.

repeated measures design requires that the subjects are employed for a long period of time, in order to participate in both experimental conditions; that may often be difficult to arrange. The matched pairs design presents difficulties in the sense that it is always difficult to 'measure' the similarity between two subjects in a number of variables. Finally, the independent groups design, although less sensitive to the effects of the independent variable than the other two methods, can be easily applied to almost any experiment.

Situational variables on the other hand, are associated with the experimental situation itself (e.g. background noise, equipment settings, experimenter's behaviour, etc.). Such factors could easily confound the effects of the independent variable if they changed systematically from one condition to another. The most effective way of avoiding this is by holding the variables in question constant throughout the experimental procedure. Clearly once we have held these variables constant they can not interfere in any way with the effects of the independent variable. For example, in the case we are examining, we could assure that all the equipment used for the experiment would have the same settings, or that there will be only one experimenter holding a invariant attitude towards the subjects. However, there are some variables that can not be held constant, as for example the fact that it may be necessary to test the subjects on different days of the week. The only way to deal with such factors is by balancing their effects across the two conditions of the experiment.

5.2.3 Other aspects of experimental design

Before proceeding to examine the specific design followed for the purposes of this thesis, it should be beneficial to clarify some other aspects of experimental design. These aspects mainly pertain to translating the basic experimental design, as described previously, into a concrete specification of what has to be done to actually run the experiment. For example issues like the methodology we will use to measure the independent variable, the kind of instructions that will be given to subjects, or the time limit they will have to perform the tasks within, have to be decided beforehand. These procedural matters, however trivial they may seem, make us think about the inferences we wish to draw from the experiment. Therefore, settling these procedural matters, or else *operationalising the experiment*, is a significant aspect of experimental design.

Finally, no mention has been made to the actual measurement of the dependent variable, e.g. of the performance of the users in judging the relevance of the retrieved documents. This issue will be the main subject of the next chapter (the analysis of the experimental results), and hence has deliberately not been addressed in the context of the present discussion.

5.3 The experimental scenario

Through the discussion presented in the previous section the basic issues related to the design of an experiment were outlined. It is the purpose of this section to describe the way that these issues are addressed for the purposes of the experiment performed in the context of this thesis. The discussion in this section will follow the pattern established in the introductory presentation of design issues: the research hypothesis will be stated, the basic experimental design will be outlined, the treatment of the irrelevant variables will be elaborated, and finally the process of operationalising the experiment shall be analysed by detailing the actual experimental scenario.

5.3.1 Research hypothesis

It was clarified previously that the aim of the specific experiment is to establish the assumption that the presence of automatically generated summaries, customised to a specific query, for each retrieved document in a ranked list, is expected to have a positive effect on the process of relevance judgement by the users. However, throughout the discussion that is to follow the reader should bear in mind that this specific hypothesis is closely related to the evaluation scheme for summarisation systems proposed earlier in this thesis. It should be appropriate to establish this relation at this point.

The aim of the proposed evaluation scheme is to judge the utility of a summarisation system in the context in which it will eventually be used, and for the purposes for which it has been built. According to this rationale, the *indicative* function [Rush et al., 1971] of a summary is the one which should be primarily evaluated. By integrating the summarisation system with an existing IR system, we both define its operational context, and its primary function: the summaries are used as a preview format in order to support a relevance decision by the users. Therefore, the proposed evaluation scheme aims at measuring the effectiveness of the automatically generated summaries in supporting the user's relevance decisions. This principal aim of the evaluation process can now be clearly mapped to the research hypothesis that we propose to establish: by proving this hypothesis we have a positive indication for the effectiveness of the auto-summaries, thus a positive evaluation measure; by refuting it we obtain the contrary conclusions.

5.3.2 Design considerations

Having established the actual hypothesis to be examined, we can introduce the basic design settings upon which the actual testing of the hypothesis will be conducted.

- **Experimental conditions**
  We are interested in two levels of the independent variable in our experimental design: the presence or absence of the query-biased summaries in the ranked list of the retrieved documents. In this way, the design will comprise of two tasks that the groups of subjects will have to perform: judge the relevance of the documents in the ranked list, with and without the assistance of the summaries. The performance of the users in these tasks constitutes the dependent variable of the experiment, and we shall attempt to prove that any variation of the performance between the two groups is attributed only to the change in the level of the independent variable.

- **Groups of subjects**

  The number of subject groups is analogous to the number of levels of the independent variable (the experimental conditions), and thus of the number of tasks to be performed. Therefore, two groups of subjects will be employed: one group will perform relevance judgements assisted by the customised summaries, while the other group will perform the same task presented with the typical output of an IR system[15] (the title, and first few sentences of each document).

  The actual number of subjects to be used in each group is a principal design issue. Clearly, in order for any experimental results to be deemed significant, a reasonably large number of subjects must be employed. The exact number that constitutes an optimal lower limit depends on the type of the experiment. However, it is generally accepted that the prospects for obtaining significant results through experimentation grow better as the sample is increased [Miller, 1984]. In practical terms this means that we should employ as many subjects as our time limits and resources allow. Taking these limitations into consideration, in the experiment described in this chapter two groups consisting of 10 subjects each were employed. It is believed that this number is capable of attributing significance to any experimental results obtained.

  Finally, another issue pertaining to the groups of subjects used is the actual way that these groups are selected. In theory, we should select the subjects randomly from the population to which we intend to generalise our experimental findings. In practice, the choice of subjects is determined by the availability of particular individuals. In the specific experimental design the population of subjects was, in majority, comprised of postgraduate students doing a conversion course in computer science. Clearly, the subject population is not representative of the population to which we wish to generalise the conclusions. This crucial decision can influence the ability to extend any experimental results beyond the bounds of the experiment itself, because statistical inconsistencies may arise [Keppel, 1973; Miller, 1984]. However, relevant studies have shown that although there are

[15] See section 2.4 of this thesis for a discussion about the typical output of IR systems.

‘risks’ in generalising the experimental results in such cases, an investigator may feel safe in doing so since the statistical differences introduced are generally of a small scale [Keppel, 1973].

5.3.3 Treatment of irrelevant variables

In section 5.2.2 a discussion about the methods to control the irrelevant variables introduced in an experimental design was presented. It is appropriate at this point to outline the actual control methods adopted in the specific design described in this chapter.

- **Subject variables**

  The method of independent groups design was adopted for the control of subject variables. The total subject sample, consisting of 20 people, was divided into two groups of 10 subjects each. The allocation of subjects to groups was performed randomly, by means of a draw. The main reason for choosing this approach was that for practical reasons it was impossible to commit the subjects for long enough time to complete both tasks (related measures design), and that it was judged inappropriate to follow the matched subjects design due to its subjective nature of matching ‘similar’ subjects.

- **Situational variables**

  There was an effort to hold the situational variables of the experiment constant throughout the whole experimental procedure. More specifically:

  – The experiments were carried out on two identically set up computers (both from hardware and software points of view). It was assured before the employment of each subject that the machines would have exactly the same settings. The same two computers were used for all the experimental sessions.

  – There was only one experimenter present throughout the experimental procedure, and there was an attempt to show the same attitude towards all the subjects. We can therefore safely enough assume that the experimenter's attitude was not an uncontrollably varying factor.

- The experimental sessions took place over a period of two days. The actual times of the sessions over these two days were as identical as possible. This was arranged in order to ensure that the external conditions of the room (especially the external noise) would be identically shared over the various sessions.

5.3.3 Operationalising the experiment

The exact procedure that was followed in each experimental session will subsequently be presented. This should provide the reader with an insight to the principal aim of the procedure; nevertheless the actual quantities to be measured from this procedure will deliberately not be mentioned. It is the purpose of the next chapter to deal with such matters.

The actual steps of the experimental procedure are as follows:

- Each subject was assigned to one of the two levels of the independent variable in the way that was previously explained. In that way the task that each subject should perform was defined.
- In order to perform the relevance judgements, each subject was presented with 5 queries. Queries were randomly assigned to subjects (by means of a draw). Moreover the 50 queries were randomly selected from a total of 250 TREC topics[16]. Finally, the same 50 queries were used for the two experimental conditions.

  It should be noted at this point that a major design issue was whether users should be able to actually enter on-line the queries to the IR system. It was decided that it would be more advantageous for the purposes of the research work, if a list of

[16] For a description of the type of queries used for the purposes of the experiment, see section 4.5.1 of this thesis.

pre-prepared queries would be presented to the users instead of allowing them to interactively input the queries. The rationale of this decision shall become apparent in the next chapter, where the results of the experiment are presented.

- Based on the task that each subject would perform, and the queries he/she would deal with, a simple user interface was prepared. The user interface will be presented in detail in the next section of the present chapter.
- As soon as the subject was placed in front of his assigned computer, the instructions about the experiment were handed to him/her. Each subject could then go through the instructions in his/her own pace. Any questions about the instructions were answered by the experimenter. Subjects were otherwise not told of the aims of the experiment, i.e. they did not know the hypothesis under testing.
- Subsequently, the user interface was presented to each subject by the experimenter, briefly describing its functionality.
- Each subject had then a time limit of 5 minutes to identify the relevant documents to each query that was prepared for him/her. The timing was performed by the experimenter. The relevant documents were marked by the subjects on a separate piece of paper (‘answer paper’) prepared for each query. On that paper, subjects were also asked to mark the document they were last examining when the 5 minutes period expired. If a subject would manage to examine all the retrieved documents for a query before the specified time ended, the experimenter was notified and the fact was noted on the answer paper. That paper was handed to the experimenter after the 5 minutes time expired.
- Once the subject had completed the assigned task, a questionnaire was presented to him/her. The completed questionnaire was then returned to the experimenter.
- A brief discussion was subsequently held with each one of the subjects that were further interested. At that point, the nature of the experiment was presented to them. They were encouraged to express their opinions about the experimental procedure, and about the overall reasoning of the experiment.

Through the previous description of the experimental procedure, its connection to the proposed evaluation scheme has become more apparent. The procedure fits the outline given in Chapter 3 of this thesis for a novel evaluation scenario of summarisation systems.

The data that was so collected from each subject comprised the completed ‘answer papers’ for each query, and the completed questionnaire. A sample of each one of the two sources of data, as well as of the instructions given to the subjects, is presented in Appendix C. The next section will deal with the presentation of the user interface that was prepared for the purposes of the experiment.

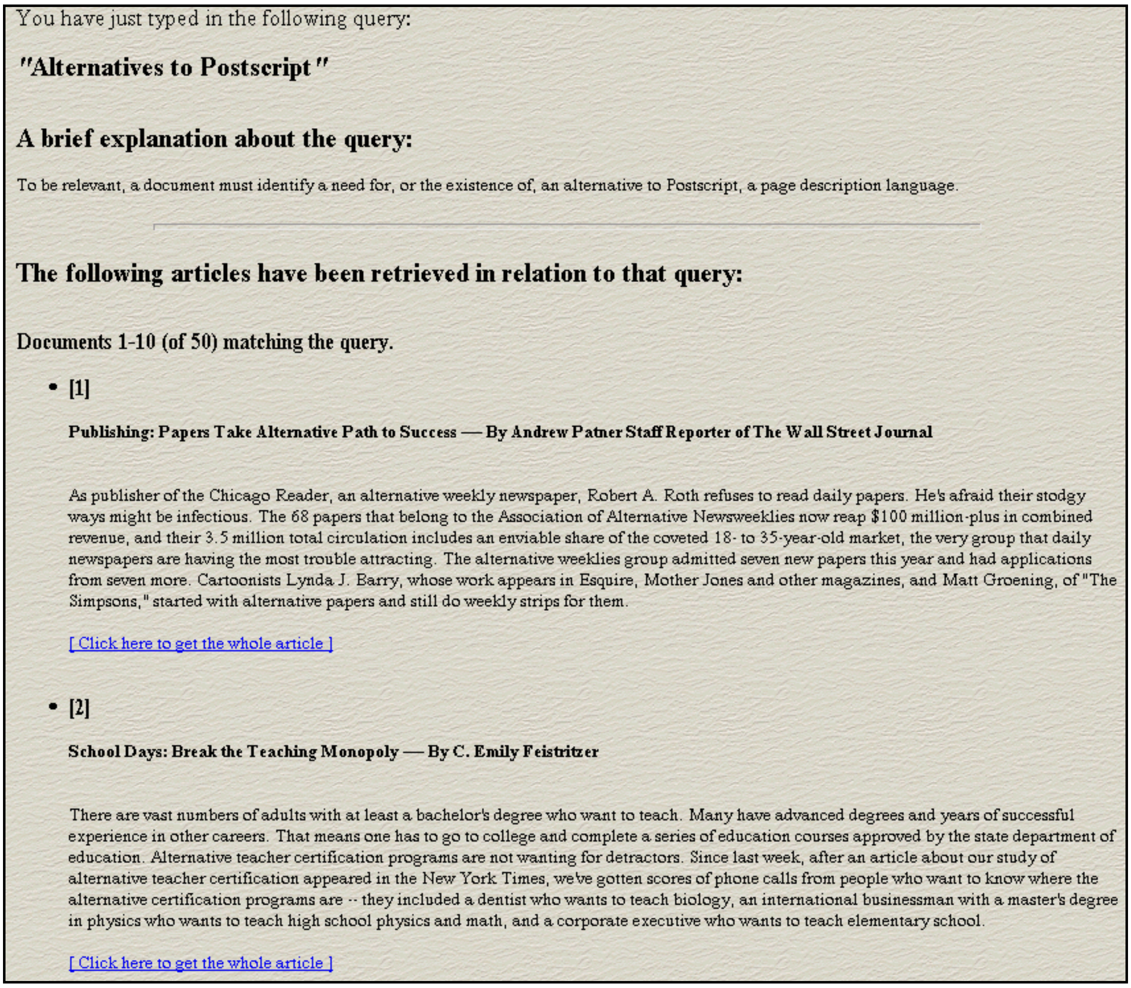
5.4 The user interface for the experiment

In Figure 5.1, a part of the user interface developed for the purposes of the experiment is depicted. Clearly, the interface can be divided in two major parts: a section pertaining to the query, and a section presenting the retrieved documents in relation to that query.

5.4.1 The parts of the interface

The upper part of the interface relates to the presentation of the query to the user. The actual query as it would have been input to the IR system had the user had that specific information need, is presented in the top of this section. This part of the query is in essence the ‘Title’ section of the TREC queries that was discussed in section 4.5.1 of the previous chapter. Following this part, the ‘Narrative’ section of the query is given as a form of a brief explanation about the query. This kind of information can be thought of as conveying the information need that is expressed by the user through the actual query. This part of the interface is held constant in both experimental conditions. The varying part is the one presenting the retrieved documents in relation to the query.

The ranked list of the retrieved documents is presented in the lower part of the interface. The actual information it provides to the user in relation to the documents’ relevance to the query varies according to the experimental condition that the user is assigned to: an automatically generated summary of each document customised to the specific query, or the leading sentences (up to three) of each document are shown. In both cases, the title of the document precedes the accompanying information. Moreover, the users are allowed to access the full text of the documents. Information about the actual number of accesses to the full text of the documents is automatically recorded on a per-user basis. Finally, it should be noted that for each query the first 50 retrieved documents are presented in 5 similarly structured pages, each of which contains 10 documents. Facilities for navigating between the pages for each query are also provided to the users.



You have just typed in the following query:

"Alternatives to Postscript "

A brief explanation about the query:

To be relevant, a document must identify a need for, or the existence of, an alternative to Postscript, a page description language.

The following articles have been retrieved in relation to that query:

Documents 1-10 (of 50) matching the query.

- **[1]**

  Publishing: Papers Take Alternative Path to Success — By Andrew Patner Staff Reporter of The Wall Street Journal

  As publisher of the Chicago Reader, an alternative weekly newspaper, Robert A. Roth refuses to read daily papers. He's afraid their stodgy ways might be infectious. The 68 papers that belong to the Association of Alternative Newsweeklies now reap $100 million-plus in combined revenue, and their 3.5 million total circulation includes an enviable share of the coveted 18- to 35-year-old market, the very group that daily newspapers are having the most trouble attracting. The alternative weeklies group admitted seven new papers this year and had applications from seven more. Cartoonists Lynda J. Barry, whose work appears in Esquire, Mother Jones and other magazines, and Matt Groening, of "The Simpsons," started with alternative papers and still do weekly strips for them.

  [Click here to get the whole article]

- **[2]**

  School Days: Break the Teaching Monopoly — By C. Emily Feistritzer

  There are vast numbers of adults with at least a bachelor's degree who want to teach. Many have advanced degrees and years of successful experience in other careers. That means one has to go to college and complete a series of education courses approved by the state department of education. Alternative teacher certification programs are not wanting for detractors. Since last week, after an article about our study of alternative teacher certification appeared in the New York Times, we've gotten scores of phone calls from people who want to know where the alternative certification programs are -- they included a dentist who wants to teach biology, an international businessman with a master's degree in physics who wants to teach high school physics and math, and a corporate executive who wants to teach elementary school.

  [Click here to get the whole article]

Figure 5.1 The user interface

During the design of the interface an attempt was made to provide the users with a familiar 'system image', e.g. to present them with a typical IR system outlook. It is believed that in this way subjects can more easily concentrate their efforts on accomplishing their assigned tasks, since they will be familiar with the computational environment. Finally, it should be noted that the interface was automatically generated, using a generic template that was then adjusted according to the specific purposes of the two experimental conditions.

5.5 Summary

The design adopted for the investigation of the effects of the automatically generated summaries on the process of relevance judgements was presented in this chapter. Initially the theoretical basis of experimental design was established, through a description of the main issues relating to it. Subsequently, the way that these issues were addressed for the purposes of the described experiment was presented. During the course of that discussion the relation of the experimental procedure with the evaluation scheme proposed earlier in this thesis was established, clarifying the fact that any experimental results from this procedure constitute an indication of the effectiveness of the summarisation system. Finally, a brief description of the user interface that was designed for the purposes of the evaluation was presented. The results that were obtained through the described experimental procedure will be presented in the next chapter of this thesis.

Chapter 6

Presentation and analysis of the experimental results

6.1 Introduction

The aim of the experimental procedure reported in this thesis is to examine the validity of the research hypothesis stated in the previous chapter, and through this investigation to provide a measure of effectiveness for the automatically generated, query-biased summaries. The design issues of this experimental procedure were addressed in the previous chapter, but neither the quantities we are interested in measuring nor the means by which these quantities are to be acquired were mentioned in the context of that chapter.

It is the purpose of the present chapter to present and analyse the data that were collected through the experimental procedure. Initially the quantities that were measured will be established, providing linkage to issues pertaining to the evaluation of summarisation systems. Subsequently the way by which these measures were obtained shall be reported, and then the actual presentation of the experimental results will take place. A discussion based on the analysis of these results will finally be presented.

6.2 Experimental measures

Clearly a research hypothesis must be testable, that is, we must be able to obtain some factual evidence which has the potential to support, or refute, the theory being tested. Otherwise, it is impossible to test the validity of the hypothesis, and therefore the problem falls outside the realm of scientific inquiry [Miller, 1984]. The factual evidence must be quantified by some specific measures that should be acquired through the experimental procedure. The actual choice of measures should facilitate the inference of conclusions: the proof or the refusal of the research hypothesis.

The decision about the quantities to measure through an experiment is an issue that is addressed even prior to the design of the procedure. In fact, the design is based on, and influenced by, the quantities we wish to measure. However, for purposes of maintaining the semantic coherence of this thesis, we shall establish the measures used to direct the inference in our experimental situation in the context of the present chapter.

6.2.1 Measuring user performance

The variable we wish to examine through experimentation (the dependent variable), is the performance of the users in the process of relevance judgements on documents retrieved by specific queries. In order to do so, a set of criteria that shall provide a satisfactory coverage of the aspects of the examined variable has to be defined. Such criteria for the experiment conducted are:

- The effectiveness of the relevance judgements performed by the subjects.
- The speed with which these judgements were performed.
- The need of the subjects to seek assistance from the full text of the retrieved documents.
- The subjective opinion of the users about the assistance provided by the information that was accompanying each retrieved document.

Based on the above set of criteria, we aim not only to establish the research hypothesis, but through that establishment to draw conclusions about the effectiveness of the summarisation system that was developed. We are justified to do so, since the evaluation of the summarisation system is performed in an end-user, task-based environment. Thus, the performance of the users in that environment can be considered as an indication of the effectiveness of the summarisation system. Clearly, if the users were required to perform a different set of tasks then a different set of measures would have to be established, possibly affecting the outcome of the evaluation. But again, in such a case the summaries that the system would generate would be influenced by different purpose factors, and would therefore possess different characteristics.

6.3 Quantifying the measures

Having defined a set of measures, we need to define ways by which values will be attributed to them: we need to quantify the measures. This is achieved by utilising the data that was collected through the experiment: the ‘answer papers’ for each query, the completed by each subject questionnaire, the information about the accesses to the full text of the documents, and finally the post-experimental discussions with the subjects. The manipulation of this data in order to quantify the set of measures is subsequently described.

6.3.1 Accuracy

There must be a way defined, by means of which the actual performance of the users in the task of relevance judgements can be measured. An obvious way is to take into account the ‘correct answers’ that each ‘answer paper’ contains, that is, to examine how many of the documents that each subject has indicated as relevant are indeed relevant. Obviously, in order to do so the actual relevant documents for each query must be known beforehand, i.e. we must have *relevance assessments* for each one of the queries used in the experiment. It is hence necessary at this point to present the way that relevance assessments for the queries used in the specific experiment were acquired.

TREC relevance assessments

In Chapter 4, the type of queries used for the purposes of the experiment was presented, and their main characteristics were discussed. It was then mentioned in Chapter 5, that as a design decision of the experimental procedure users would not interactively input queries to the IR system; instead a list of pre-prepared queries would be presented to them. The principal reason for this decision was that there are *relevance assessments* available for the TREC topics, thus facilitating the measurement of user performance.

With respect to the relevance assessments for the TREC queries, it should be noted that documents do not have to be wholly, or even primarily, about a request topic in order to be deemed relevant [Sparck Jones, 1995]. A document may be characterised as relevant given only the facts conveyed in two independent sentences. The implications of the fact that these relevance assessments are considered as the basis on which user performance will be measured, shall be discussed during the analysis of the experimental results later in this chapter.

Having established the notion of relevance assessments for the queries used, we can, by means of an example, describe the measure used to quantify the accuracy of the relevance judgements performed by the subjects. Lets assume that subject A was presented with the query Q_1, and with the first 50 documents retrieved in relation to that query. Lets also assume that within the 5 minutes period subject A marked documents 1, 4, 6, 7, 13, and 17 (6 in total) as relevant to the query he was examining, and that he also indicated that when the time limit expired he was examining document 22. According to the relevance assessments for query Q_1, taking into account only the first 22 retrieved documents[17], the relevant ones are 1, 2, 3, 4, 6, 9, 13, 15, and 19 (9 in total). Hence, subject A had successfully identified 4 out of the 9 in total relevant

[17] It would be of no use to examine relevant documents beyond the last document that the subject had examined.

documents (1, 4, 6, and 13), showing a 44.4% *success rate* in identifying the 'correct answers'. It should be noted that there is a case in which the *success rate* can not be defined: when no relevant documents exist up to the last document that the subject was examining. In our example if no relevant documents have been retrieved within the first 22 documents it would be impossible to define the success rate of subject A.

The notion of success rate alone can not allow us to decide on the accuracy of the relevance judgements. If subject A had indicated all 22 documents as relevant, he would obviously had a success rate of 100% since he would have identified all the relevant items. On the other hand, the subject would have also erroneously pronounced relevant 13 documents (22 indicated relevant - 9 actually relevant). In our example, 2 documents (7, and 17) have erroneously been indicated as relevant by subject A. A useful measure of user performance can then be defined by the number of 'correctly' judged relevant documents divided by the total number of the indicated relevant documents. In the specific example, subject A would show a 66.6% (4 / 6) *utilisation* in the performed relevance judgements.

It is here proposed that these two measures (success rate, and utilisation) should be both examined in order to define the accuracy of the relevance judgements. The former measure pertains to the ability of the subject to correctly identify the relevant documents, while the latter measure indicates the 'correctness rate' of the subject's judgements; neither of these measures alone can allow us to make inferences about the accuracy of the relevance judgements.

6.3.2 Rapidity

Subjects were asked to indicate on the 'answer paper' for each query the last document they were examining when the 5 minutes period expired. Hence, for each user the actual number of documents he/she managed to examine within the specified time limit is known. This number is used as an indication of the speed with which the relevance judgements were performed by each subject. This indication can provide evidence about which experimental condition facilitates the faster execution of the user-performed task. Obviously other factors may as well become determining (e.g. the browsing habits of each subject), and they will be discussed during the analysis of the results.

6.3.3 Reference to the full text of the documents

Information about the actual number of times that each subject had to refer to the full text of the documents is automatically stored by the user interface used for the experiment. This data is indicative of the clues that the user is presented with in relation to the relevance of the retrieved documents: frequent reference to the full text can be interpreted as inadequacy of the accompanying document information to provide enough evidence, whereas rare reference indicates the opposite.

6.3.4 Subjective opinions of the users

Subjects were asked to state their opinion about the assistance offered to their judgements by the accompanying for each document information. This information was collected from the questionnaire that each subject completed: they were asked to rate the helpfulness of the accompanying information on a scale from 1 (most helpful) to 5 (least helpful). Furthermore, through the post-experimental discussion sessions with the subjects, their opinion on a number of relevant issues was also stated. Although this opinion can not be quantified, it can provide supporting evidence to the overall rating of the utility of the generated summaries.

6.4 Presentation of the results

The results that were collected on the four previous categories from the experimental procedure will subsequently be presented. A discussion about the results in each distinct category will also be made. Conclusions drawn from the overall consideration of the results will be reported at the end of this section.

6.4.1 Accuracy

The success rate for the group of subjects using the summaries is considerably larger than that of the group using a typical IR output: the difference in performance is 15.84%. The interpretation of this result is that users in the 'summary group' managed to successfully identify a larger number of relevant documents than the other group.

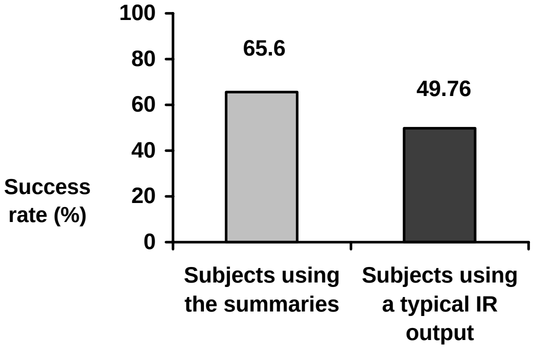


Figure 6.1 Success rate values for the two groups

Nevertheless, in order to have an overall view of the accuracy of the relevance judgements, we need to examine the performance of the two groups in the utilisation of the judgements. The utilisation measures obtained for the two experimental groups are presented in Figure 6.2.

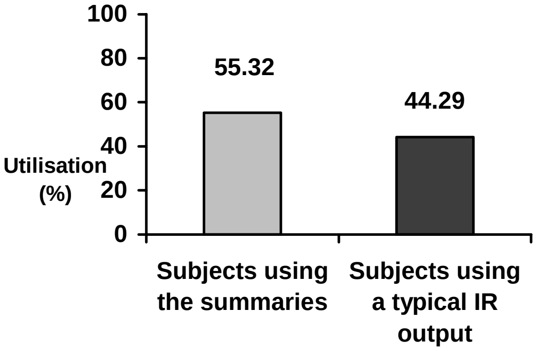


Figure 6.2 Utilisation measures for the two groups

The data presented in Figures 6.1 and 6.2, have been acquired by averaging the results for each query over the total number of queries, thus producing the average success and utilisation values per query. In Appendix D the results obtained from each individual user, for each one of the two experimental groups, are presented. In order to establish the statistical significance of these results, *t-tests*[18] were performed on both these measures, indicating that, with probability of error 0.05, the results are attributed to the change of level of the independent variable and not to chance factors.

The significance of the TREC relevance assessments in obtaining these results is a factor that has to be examined. Initially, it could be argued that the acceptance of the TREC assessments as the 'correct answers' for the relevant to each query documents is not fully justified: different users may judge different documents as relevant, and furthermore it is possible that the TREC assessments are not totally accurate. It may be the case for example, that a document not deemed as relevant by the TREC assessments is justifiably considered as relevant by one of the subjects. This possibly is a significant factor for the low values obtained in the *utilisation* measure of the relevance judgements: users tend to have different (than the TREC assessments) opinions about the relevance of a large number of documents. In order to balance the effects of this issue in the experimental procedure, the same assessments apply for all the queries, and for all subjects in both experimental conditions. Therefore, whatever the effects of the 'subjective' TREC assessments may be, they are evenly distributed over the whole subject population.

In addition to that issue, it must also be noted that the rationale of the TREC assessments fits to the logic of the summarisation system: documents do not have to be wholly, or even primarily, about a request topic in order to be deemed relevant. A couple of sentences with relevant to the query content are an adequate criterion for the relevance of the whole document. The summarisation system selects information from each document based on, among other criteria, the distribution of query terms in its constituent sentences. In this way, the generated summaries aim to help users to more easily identify the relevant to the query pieces of information that are contained in each document. Thus, the rationale of the summarisation system approximates the rationale

[18] For a discussion on statistical methods to establish the significance of experimental results see [Miller, 1984], pp. 48-65.

of the TREC assessments: high distribution of query terms in a sentence, can possibly be an evidence of its relevance to the query. The results reported in this section strengthen this belief: subjects using the summaries have been significantly assisted by the conveyed information in relation to the specific query.

The queries employed in the experimental procedure were artificial, i.e. they were not actually expressing the subjects' information needs. This fact is possibly affecting the overall process of relevance judgements. This observation can possibly explain the lower values in the *utilisation* measure: users find it difficult to map their information need to that expressed through the query they are presented with. However, all subjects, in both experimental conditions, were presented with artificial queries. Therefore, the counterbalance of any effects that the artificial requests for information may introduce, is addressed in the context of controlling the irrelevant variables of the experimental procedure (randomly assigning queries to subjects and subjects to experimental conditions). Nevertheless, if the issue of measuring user performance without available assessments for the queries employed can be addressed, it would definitely be interesting to replicate the same experimental situation with the interactive input of queries by users.

A final point of discussion is the effectiveness of the summaries on 'warning' users about the non-relevance of the documents. A first indication is provided by the highest utilisation measure that the group of subjects using the summaries showed. They tend to erroneously judge less documents than the other group, so we could say that summaries help them to distinguish the non-relevant documents more effectively. A strengthening point for this observation is provided from the cases where the success rate can not be defined. As we explained earlier, in such cases no relevant documents exist up to the document that the user had last examined. There were 8 such cases in the group using the summaries, and 9 in the other group. The former group, managed to avoid ticking any irrelevant documents 4 out of 8 times, while the latter group succeeded in doing so only 1 out of 9 times.

At this point, having attained the overall view of the presented results, we can conclude that subjects using the automatically generated summaries perform their relevance judgements significantly more accurately than users using a typical IR output.

In essence this means that they identify more relevant documents, while at the same time they erroneously judge as relevant less documents. This difference in performance can be attributed to the corresponding change of level in the value of the independent variable of the experiment: the presence of the automatically generated summaries as accompanying information in the ranked list of the retrieved documents.

6.4.2 Rapidity

In Figure 6.3, the results obtained in the category of the rapidity of the relevance judgements are presented. These results have been obtained by averaging the number of examined documents for the two experimental conditions over the total number of queries used (50). Thus, subjects using the summaries examined on average 22.62 documents per query, while subjects using a typical IR output examined on average 20 documents.

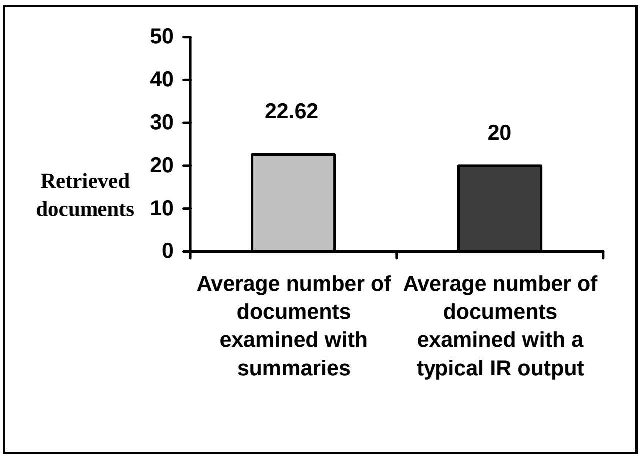


Figure 6.3 Rapidity results

However not great the difference in performance between the two groups may seem, there is a 13% increase in the average number of documents examined by the group of subjects using the auto-summaries. Taking also in consideration the specific time limits of the experiment, we can conclude that there is a definite tendency that users presented with the query-biased summaries perform the relevance judgements more rapidly than users presented with a standard system output.

The most important issue that one should consider when viewing these results, is that certain users possess certain browsing skills, and these skills affect the rate at which they examine the documents. Users were not clearly instructed to be motivated towards a specific goal, e.g. examining as many documents as possible, or thoroughly examining each one of the retrieved documents. Therefore each subject utilised his/her own personal browsing habits. This led to the categorisation of subjects in two groups: subjects that browsed quickly through the list of the retrieved documents, examining in detail only those documents that could possibly bear relevance, and subjects that exhaustively examined each document in the list.

With use of summaries		**With a typical IR output**	
Subjects	**Examined documents**	**Subjects**	**Examined documents**
User 1	25, 23,17, 29, 20	User 11	20, 26, 20, 31, 39
User 2	25, 22, 33, 37, 30	User 12	16, 19, 17, 36, 50
User 3	19, 33, 20, 20, 20	User 13	4, 10, 13, 13, 12
User 4	24, 29, 38, 50, 50	User 14	7, 12, 10, 12, 18
User 5	7, 13, 16, 16, 14	User 15	20, 28, 46, 29, 32
User 6	21, 20, 14, 9, 13	User 16	11, 30, 24, 27, 21
User 7	11, 10, 7, 10, 11	User 17	5, 8, 10, 5, 12
User 8	16, 7, 24, 41, 50	User 18	13, 33, 26, 18, 23
User 9	28, 27, 45, 25, 25	User 19	18, 16, 31, 18, 31
User 10	11, 9, 37, 16, 14	User 20	20, 12, 14, 19, 40

Table 6.1 Results grouped on a per-user basis

The results presented in Table 6.1 strengthen this belief. In this presentation, the actual performance of each subject is clearly depicted through the number of documents he/she examined for each one of the queries he/she was presented with. Given the variability in the difficulty of the queries that each subject was assigned to, the results of Table 1 show a considerable concentration of values on a per-user basis.

Therefore, in order to be able to conclude more safely about the effect of the auto-summaries on the speed of the process of relevance judgements, we should employ the same subject in both experimental conditions, and with queries of approximately the same level of difficulty. In that case, we could examine the actual effect of the summaries on the speed of his/her judgements. Nevertheless, considering the fact that queries were randomly assigned to subjects, and that subjects were randomly assigned to experimental conditions, hence adequately counterbalancing the *subject variable*[19], we can infer that there is an effect caused by the presence of the auto-summaries on the speed of the process of relevance judgements, and that this effect is positive.

6.4.3 Reference to the full text of the documents

The data collected regarding the accesses of the full text of the documents, showed that subjects using the summaries had to refer to 0.3 full texts of the retrieved documents for each query, whereas subjects from the other experimental group had to refer to 4.74 full texts. If we normalise these values to the average number of documents that each experimental group examined for each query, we obtain the results shown in Figure 6.4. This figure shows that each subject using the summaries had to refer to the full text of 1.32% of the documents he examined for each query, while subjects in the other experimental condition had to refer to the 23.7% of the examined documents.

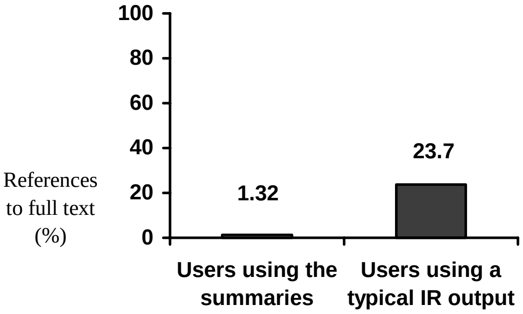


Figure 6.4 Average number of references to the full text of the documents (per query)

[19] The reader may refer to the previous chapter of this thesis for an introduction to the notion of subject variables.

Obviously, the difference between the two measured values is large. This difference can be clearly attributed to the accompanying information that the subjects were presented with for each retrieved document. This result verifies the initial assumption that the approach adopted by the majority of IR systems for presenting the user with evidence about the relevance of retrieved documents to queries is inadequate. Users need more clues to establish the relevance of documents, and especially they need clues about the context in which the query terms are used in these documents. If these clues are not provided from the accompanying information, users resolve to the full text of the documents, which is usually large and difficult to manage. It is the case in the specific experimental situation, that the automatically generated, query customised summaries, provided the subjects with enough evidence to support their relevance judgements. Furthermore, bearing in mind the results pertaining to the accuracy of the relevance judgements, we can conclude that the summaries also provided the users with the necessary information to adequately decide on the relevance of the documents.

Had the difference in performance between the two experimental groups been not so great, we would have to prove that the individual browsing habits of the subjects did not significantly affect the result. We would in that case have to relate each subject's speed of relevance judgements with the accuracy of the judgements, and with the number of references to the full text of the documents. For example, we should relate a subject's accuracy with the number of references to the full text, in order to examine if this number affects his/her performance. However, having established both the superior accuracy of the group using the summaries, and its considerably faster performance in the assigned task, there is no need to do so.

6.4.4 Subjective opinions of the users

As a form of confirmation of the results obtained in the previous categories, the subjective opinions of the users rate the utility of the auto-summaries higher than that of the typical IR output. This result is clearly depicted in Figure 6.5. It is reminded that users were asked to rate the utility of the accompanying for each document information on a scale from 1 (most helpful) to 5 (least helpful). The data shown in this figure indicate that subjects using summaries rated on average the utility of the accompanying information with 1.5, while subjects assigned in the other experimental condition rated the utility of the information they were presented with, with a ‘mark’ of 2.5.

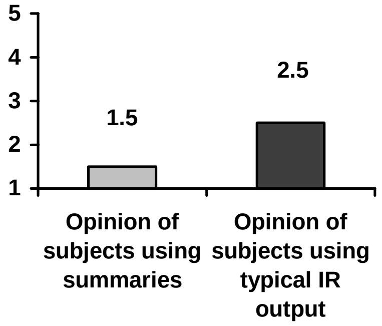


Figure 6.5 Subjective opinions of the users

Furthermore, during the post-experimental discussions, users that were presented with a typical IR output expressed their dissatisfaction regarding the information they were presented with. More specifically, they emphasised on the fact that they had to refer to the full text for almost every document they were examining. In fact some of those subjects mentioned that it would be helpful for their relevance judgements if the could somehow see how the query terms are used within each retrieved document. Hence, the outcome of the post-experimental discussions is yet another indication of the assumption made, that users require more clues about the relevance of the retrieved documents than they are usually presented by typical IR systems. The automatically generated, query-customised summaries, have focused on capturing that requirement.

6.4.5 Overall view of the results

Having examined the results obtained in the effectiveness measures defined for the specific experimental situation, we can draw the following conclusions on the effect of the summaries on the process of relevance judgements:

– They assist users in performing relevance judgements more accurately. Users can identify more relevant documents for each query, while at the same time judge erroneously less documents. This effect of the summaries is attributed mainly to their indicative nature, and especially on the fact that they adequately indicate the context within which the query terms are used in the retrieved documents.

- They almost alleviate the need to refer to the full text of the documents. Users rely almost solely in the information conveyed in the query-biased summaries in order to perform their relevance judgements. If we examine this result in relation to the increase in the accuracy of the relevance judgements, we can conclude that the automatically generated summaries successfully provide users with clues about the relevance of the retrieved documents.
- They seem to allow users to decide on the relevance of the documents more rapidly. Having established the increased accuracy of the judgements, we can extend the experimental results, and state that the automatically generated summaries allow users to perform their judgements more accurately and more rapidly.

An interesting implication of the results can be presented in IR systems, in cases where users are not willing to spend much time for identifying the relevant documents, or where it can be ineffective to access the full text of the documents. Typical examples of such systems are the internet search engines. In a typical interaction with a search engine it is rather ‘costly’ to access the full text of the documents, since they may be located in a physically distant storage device. It therefore becomes apparent that the approach proposed in this thesis could easily be applied is such cases, in order to allow users to minimise the accesses to the full texts of irrelevant documents, while at the same time to effectively access documents of interest.

6.5 Summary

In this chapter, the results that were obtained from the experimental procedure were presented and analysed. Initially, the actual quantities (user-performance measures) that were acquired through the experiment were established, and the way by which these quantities were measured was described. The experimental results were then presented and analysed for each one of the defined performance measures. The conclusion drawn from the analysis of the experimental results presented in this chapter, is that the automatically generated summaries effectively reflect the user’s information need, and therefore have a positive effect on the process of relevance judgements.

Chapter 7

Future work and conclusions

7.1 Future work

The results presented in the previous chapter proved the effectiveness of the summarisation system developed for the purposes of this thesis. More specifically, the results showed that the presence of automatically generated summaries as accompanying information to documents retrieved in relation to a specific query, has a positive effect on the process of relevance judgements performed by users. However, there are a number of points on which future research work can concentrate in order to further improve these results. These points pertain both to design and evaluation issues of summarisation systems, and to issues of effectively representing the user's information need.

Experimentation and evaluation

- During the design of the summarisation system used in this thesis, various combinations of sentence extraction methods and of system settings (e.g. summary length) were tried, but due to time and resource limitations the effects of these combinations on the generated summaries were not evaluated. It would therefore seem appropriate to apply a task-based evaluation scheme, in order to estimate the variation of the effectiveness of the summarisation system according to the various combinations. In such a scenario, users could interactively adjust the various parameters: they could, for example, set a series of values for the summary length until the presented information would fit their requirements.

- Another interesting aspect would be to evaluate the effectiveness of a summarisation system that would concentrate on coherency aspects of the generated summaries. If the improvement in the performance of the subjects is significant, that would be an indication that coherency of the output text is a significant effectiveness factor that is irrelevant of any specific purposes applied on the design of the system.
- The comparative evaluation of the summarisation system with other systems that attempt to improve the presentation of retrieved documents to users would be an interesting research point. For example, the summarisation system developed here could be evaluated against passage retrieval systems that present the user with paragraphs matching his query [Knaus et al., 1995].
- Finally, different evaluation scenarios that would measure different aspects of summarisation systems should be developed. For example, it would be interesting to apply different purpose factors on the summarisation system, e.g. to assist users in comprehending the content of the document in order to answer specific questions pertaining to the specific document. Then the actual effects of purpose factors on the form of the summaries, as well as on other aspects of the summarisation process, could be quantified.

Query customisation

In order to present the user with enough evidence about the relevance of retrieved documents to his information need, there are a number of potential enhancements that could be accommodated in the design of the summarisation system developed for the purposes of this thesis.

- It is believed that phrases can provide significant information about the content conveyed by documents and by queries [Fagan, 1987]. The use of phrases for content identification should therefore be investigated in the context of a summarisation system. In that way clues for the significance of the input text can be provided, both from the recognition of query phrases in the documents, and from the identification of phrases within the sentences of the input text. Both statistical and syntactic methods should be explored.

– The use of synonyms for query terms should also be examined. It has been suggested [Stairmand & Black, 1996], that there are limitations imposed on the process of retrieval of relevant documents by the fact that indexing of documents (and of queries) is performed by *key words* as opposed to *key concepts*. In this way concepts such as polysemy[20] and synonymy are ignored. For example, when a user inputs a query about 'commercial aircraft manufacturers', he is possibly also interested in locating synonyms of the term 'aircraft' (e.g. airplane, aeroplane, plane) in the context of the retrieved documents. The effect of such an approach to the effectiveness of the summarisation system should be investigated.

– Finally, user interface issues relating to the presentation of relevant to the query information to users should be researched. For example, various ways to focus the user's attention on the context in which the query terms are used within each retrieved document would be an interesting point of further research.

7.2 Conclusions

The aim of this thesis was to investigate the effectiveness of automatically generated summaries customised to a specific query, in assisting users to judge the relevance of documents retrieved in response to that query. This aim was motivated by the observation that users are usually not presented with enough evidence about the relevance of retrieved documents to their request for information. In order to actually test the effectiveness of summaries, the research work reported in this thesis concentrated on the design of a summarisation system, and on the development of an evaluation scheme that would examine its effectiveness.

[20] It means a word that can refer to more than one concepts.

To derive an effective design for the summarisation system, the work carried out in the field of automatic text summarisation was initially reviewed. This review was given in Chapter 3 of this thesis, and it presented the two main approaches followed for the summarisation of textual documents: the extraction of ‘significant’ sentences from the input text, and the generation of summary text based on artificial intelligence and knowledge-based techniques. The main characteristics of these approaches were discussed: sentence extraction methods rely on the identification of significance clues in order to assign scores to the sentences of the input text, whereas language generation approaches attempt to exploit semantic and syntactic information from the original documents. The conclusion drawn from this review was that although sentence extraction methods suffer from a number of problems (mainly the lack of coherence of the output text), they are capable of generating summaries that are domain independent. On the other hand, language generation systems rely on specific domain characteristics, and therefore systems using such methods are restricted to a limited application area.

In the same chapter of this thesis, the work carried out in the evaluation of summarisation systems was also reviewed. The most common evaluation approaches measure quantitative characteristics of the automatically generated summaries, as for example the number of sentences that the summarisation system and a human expert both select for inclusion in a summary. The fact that there have not been many attempts to evaluate qualitative aspects of the summaries was emphasised. Based on that observation, a scheme that would evaluate summaries in an operational, task-based environment was outlined. The essence of the proposed scheme is that the utility of a summarisation system should be primarily evaluated in the operational context in which it will eventually be used, and for the purposes for which it has been designed.

Based on the study of the research work on the field of text summarisation, the architecture for the summarisation system used for the purposes of this thesis was designed, and presented in Chapter 4. A design goal for the automatically generated summaries was that they should be customised to a specific query, so as to provide users with clues about the relevance of the retrieved documents to the query. The summarisation system was based on the computation of a score for each sentence of the input text that would be indicative of its extract-worthiness. In order to compute a significance score for each sentence, evidence from the structural organisation of the documents to be summarised, from the frequency of terms within these documents, and from the distribution of query terms in each input sentence were utilised. The summaries would then comprise a predetermined number (summary length) of top-scoring sentences.

The point that was emphasised throughout this chapter, was that the form of the automatically generated summaries is significantly determined by the purpose for which they are actually generated. Therefore, many of the design decisions taken during the development of the summarisation system, were affected by the specific purpose for which the summaries were needed: to effectively inform users about the relevance of each summarised document to a specific query.

Having designed the summarisation system, the procedure that would test its effectiveness had to be established. Therefore, Chapter 5 elaborated on the task-based evaluation scenario that was outlined earlier in this thesis. In summary, two groups of subjects were employed, with each group assigned to a different experimental condition. These conditions were determined by the presence or the absence of the automatically generated summaries as accompanying information to each document retrieved in response to a specific query. Subjects were then asked to identify the relevant documents in response to that query. The rationale of the experimental design adopted, was to examine the effectiveness of the summarisation system in a specific operational context (interaction with users), and for a specific purpose (to assist users in judging the relevance of documents).

The procedure by which the data collected from the experimental procedure were used as an indication of the effectiveness of the summarisation system, was then presented in Chapter 6. The set of measures employed covered aspects such as the accuracy of the relevance judgements, the speed by which they were performed, the need of subjects to refer to the full text of the retrieved documents, and the subjective opinion of the users about the general utility of the summaries.

The conclusions drawn from the experimental results were that subjects presented with the automatically generated summaries, judged more accurately and more rapidly the relevance of the retrieved documents, without having to refer to the full text of the documents to seek for complementary relevance evidence. Through these results, the effectiveness of the summarisation system in the specific combination of operational circumstances and purpose factors was proved.

To sum up, the results of the work reported in this thesis should provide motivation for the concentration of future research work on the effective representation of user information needs in the context of information retrieval systems.

REFERENCES

[Abracos & Lopes, 1997] Abracos, J.; Lopes, G.P. Statistical methods for retrieving most significant paragraphs in newspaper articles. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarisation (ISTS '97)*, pp. 51-57. Madrid, Spain, July 11 1997.

[Aretoulaki, 1997] Aretoulaki, M. COSY-MATS: An Intelligent and Scalable Summarisation Shell. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization (ISTS '97)*. Madrid, Spain, July 11 1997.

[Brandow et al., 1995] Brandow, R.; Mitze, K.; Rau, L.F. (1995) Automatic condensation of electronic publications by sentence selection. In *Information Processing & Management*, 31(5), pp. 675-685, September 1995.

[Callan, 1994] Callan, J. P. Passage-level evidence in document retrieval. In Croft, W.B. and van Rijbergen, C.J. (editors) *Proceedings of the Seventeenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 302-310. ACM Press, July 1994.

[DeJong, 1982] DeJong, G. An overview of the FRUMP system. In Lehnert, W.G. and Ringle, M.H. (editors) *Strategies for Natural Language Processing*, pp. 149-172. London: Lawrence Erlbaum, 1982.

[Edmundson, 1964] Edmundson, H.P. Problems in automatic abstracting. In *Communications of the ACM*, 7(4), pp. 259-263, April 1964.

[Edmundson, 1969] Edmundson, H.P. New methods in automatic abstracting. In *Journal of the ACM*, 16(2), pp. 264-285 April 1969.

[Enders-Niggemeyer et al., 1993] Enders-Niggemeyer, B.; Hobbs, J.; Sparck Jones, K. Summarizing Text for Intelligent Communication, Dagstuhl Seminar 9350.
URL : http://www.dib.fh-hannover.de/SimSum/Abstract/, December 1993.

[Fagan, 1987] Fagan, J. L. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. *Ph.D. Thesis*, Department of Computing Science, Cornell University, Ithaca, New York. 1987.

[Frakes & Baeza-Yates, 1992] Frakes, W.B.; Baeza-Yates, R. (editors). Information retrieval: Data structures and algorithms. Prentice Hall, London.

[Goffman & Newill, 1967] Goffman, W.; Newill, V. A. Communication and epidemic processes. In *Proceedings of the Royal Society*, Series A, 298(1454), pp. 316-334. May 1967.

[Hand, 1997] Hand, T.F. A proposal for a task-based evaluation of text summarisation systems. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarisation (ISTS '97)*, pp. 31-38. Madrid, Spain, July 11 1997.

[Jacobs & Rau, 1990] Jacobs, P.S.; Rau, L.F. Scisor: Extracting information from on-line news. In *Communications of the ACM*, 33(11), pp. 88-97, November 1990.

[Keppel, 1973] Keppel, G. Design and analysis: A researcher's handbook. Prentice Hall, New Jersey.

[Knaus et al., 1995] Knaus, D.; Mittendorf, E.; Schäuble, P.; Páraic, S. Highlighting relevant passages for users of the interactive SPIDER retrieval system. In *Proceedings of the TREC-4 Conference*.

[Kupiec et al., 1995] Kupiec, J.; Pedersen, J.; Chen, F. A trainable document summarizer. In Fox, E.A.; Ingwersen, P.; Fidel, R. (editors) *Proceedings of the Eighteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68-73. ACM Press, July 1995.

[Lancaster, 1968] Lancaster, F.W. Information retrieval systems: characteristics, testing and evaluation. Wiley, New York.

[Luhn, 1958] Luhn, H.P. The automatic creation of literature abstracts. In *IBM Journal of Research and Development*, 2(2), pp. 159-165, April 1958.

[Maizell et al., 1971] Maizell, R.E.; Smith, J.F.; Singer, T.E.R. Abstracting scientific and technical literature: An introductory guide and text for scientists, abstractors and management. Willey-Interscience. John Willey & Sons Inc., New York, 1971.

[Mani & Bloedorn, 1997] Mani, I.; Bloedorn, E. Multi-document summarisation by graph search and matching. In *Proceedings of AAAI-97*, Providence Rhode Island, 1997.

[Maybury, 1995] Maybury, M.T. Generating summaries from event data. In *Information Processing & Management*, 31(5), pp. 735-751, September 1995.

[McKeown, 1985] McKeown, K.R. Text Generation: Using discourse strategies and focus constraints to generate natural language text. Cambridge University Press, 1985.

[McKeown et al., 1995] McKeown, K.; Robin, J.; Kukich, K. Generating concise natural language summaries. In *Information Processing & Management*, 31(5), pp. 703-733, September 1995.

[McKeown & Radev, 1995] McKeown, K.; Radev, D.R. Generating summaries from multiple news articles. In Fox, E.A.; Ingwersen, P.; Fidel, R. (editors) *Proceedings of the Eighteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-82. ACM Press, July 1995.

[Miike et al., 1994] Miike, S.; Itoh, E.; Ono, K.; Sumita, K. A full-text retrieval system with a dynamic abstract generation function. In Croft, W.B. and van Rijbergen, C.J. (editors) *Proceedings of the Seventeenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 152-161. ACM Press, July 1994.

[Miller, 1984] Miller, S. Experimental design and statistics (2nd edition). NY, Routledge 1984.

[Paice, 1981] Paice, C.D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R.N.; Robertson, S.E.; van Rijbergen, C.J.; Williams, P.W. (editors) *Information retrieval research*, pp. 172-191. Butterworths, 1981.

[Paice, 1990] Paice, C.D. Constructing literature abstracts by computer: Techniques and prospects. In *Information Processing & Management*, 26(1), pp. 171-186, 1990.

[Paice, 1993] Paice, C.D. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research & Development in IR*, 1993.

[Rush et al., 1971] Rush, J.E.; Salvador, R.; Zamora, A. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. In *Journal of the American Society for Information Science*, 22(4), pp. 260-274, 1971.

[Salton & McGill, 1983] Salton, G.; McGill, M.J. Introduction to modern information retrieval. McGraw-Hill, New York, 1983.

[Salton & Buckley, 1988] Salton, G.; Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. In *Information Processing and Management*, 24(5), pp. 513-523, 1988.

[Salton et al., 1996] Salton, G.; Singhal, A.; Buckley, C.; Mitra, M. Automatic text decomposition using text segments and text themes. In *Hypertext '96, The 7th ACM Conference on Hypertext*, pp. 53-65. New York: ACM Press 1996.

[Salton et al., 1997] Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. Automatic text structuring and summarisation. In *Information Processing & Management*, 33(2), pp. 193-207, 1997.

[Saracevic, 1969] Saracevic, T. Comparative effects of titles, abstracts and full texts on relevance judgements. In *Proceedings of the American Society for Information Science*, (6), pp. 293-299.

[Skorokhod'ko, 1971] Skorokhod'ko, E.F. Adaptive method of automatic abstracting and indexing. In *Information Processing*, volume 2, p.p. 1179-1182. North-Holland Publishing Company.

[Sparck Jones & Kay, 1973] Sparck Jones, K.; Kay, M. Linguistics and information science. Academic Press, New York and London.

[Sparck Jones & Enders-Niggemeyer, 1995] Sparck Jones, K.; Enders-Niggemeyer, B. Introduction: Automatic summarising. In I*nformation Processing & Management*, 31(5), pp. 625-630, September 1995.

[Sparck Jones, 1995] Sparck Jones, K. Reflections on TREC. In *Information Processing & Management*, 31(3), pp. 291-314, 1995.

[Stairmand & Black, 1996] Stairmand, M. A.; Black, W.J. Conceptual and contextual indexing using WordNet-derived lexical chains. In *Proceedings of the 18th BCS Colloquium on information retrieval research*, pp. 47-65. Manchester Metropolitan University, March 1996.

[Van Rijsbergen, 1979] van Rijsbergen, C.J. Information retrieval (2nd edition). Butterworths, London.

Appendix A

Sample document from the WSJ collection

The following is a sample document of the collection that was used for the purposes of the summarisation system.

```
<DOC>
<DOCNO>
WSJ900405-0113
</DOCNO>
<DOCID>
900405-0113.
</DOCID>
<HL>
   The U.S. Takes the Wrong Immigrants
   ----
   By George J. Borjas
</HL>
<DATE>
04/05/90
</DATE>
<SO>
WALL STREET JOURNAL (J), PAGE A18
</SO>
<LP>
   Just as countries compete in a world-wide market where goods and
services are exchanged, they also compete in the market for people.
The U.S. accepted about 600,000 legal immigrants annually during the
1980s, not including the 2.5 million persons who applied for amnesty
under the provisions of the 1986 Immigration Reform and Control Act.
Including illegals, more immigrants were admitted during the 1980s
than in any other decade in American history.
   By presenting a specific set of economic opportunities, and by
pursuing an immigration policy that prevents the entry of some persons
but encourages the entry of others, the U.S. makes a particular type
of offer in the "immigration market." The evidence is clear that the
American offer is becoming progressively less attractive to the
world's most talented people.
</LP>
```

<TEXT>

Since the immigration reforms of 1965, U.S. immigration law has encouraged family reunification and discouraged the arrival of skilled immigrants: 75% of legal immigrants in 1987 were granted entry because they were related to an American citizen or resident, while only 4% were admitted because they possessed useful skills. As a result, the skills of immigrants entering the U.S. have declined sharply over the past few decades relative to the skills of natives. Newly arrived immigrants admitted in the late 1950s had about half a year more schooling than natives did and earned about 8% less per hour. Immigrants admitted in the late 1970s had 0.7 fewer years of schooling and earned about 17% less than natives.

The U.S. competes for immigrants with several other countries, but its main rivals are Australia and Canada. Among them, these three countries account for two-thirds of the world’s legal immigration. But since 1965, Australia and Canada have succeeded better in drawing the most talented immigrants. While the predicted lifetime earnings -- the best indicator of the value of the technical and intellectual skills that immigrants bring with them -- of the immigrants who entered the U.S.in the late 1970s were 30% lower than those of American natives, immigrants to Canada who arrived in Canada during that period are projected to earn only 13% less than Canadian natives. Late 1970s immigrants to Australia will likely earn about the same as Australian natives.

By contrast, those immigrants who arrived in the U.S. in the early 1960s will have lifetime earnings just 7% below those of natives -- a figure exactly to the performance of contemporary immigrants to Australia and only very slightly worse than the 3% less that early 1960s immigrants to Canada will earn.

The flagging performance of immigrants to the U.S. is due in part to the changes in their national origins. The national origin groups that dominate the immigrant flow today do relatively less well in the labor market than the groups that dominated earlier flows. For instance, the predicted lifetime earnings of immigrants from Canada, Germany or Britain are about 20% higher than those of natives; immigrants from India or Korea earn about 7% less than natives; and immigrants from the Dominican Republic, Jamaica or Mexico earn 30% to 40% less than natives.

One ready explanation for these disparities is that the skills of workers originating in advanced, industrialized economies are more easily transferable to the U.S. labor market than are the skills of persons from less developed countries. But a subtler cause is at work too: It is the most skilled workers who wish to leave such countries as Sweden and Britain.

Because of the wage structure and redistributive income policies in the European social democracies, the income gap between the highly skilled and the less skilled is small. Highly skilled workers are not well rewarded and the less skilled are protected from poor labor market outcomes. This creates an incentive for those highly skilled people to emigrate. In poor countries, on the other hand, the wage gap tends to be very large, and it is the less skilled who have the most incentive to leave.

The problem with unskilled immigrants is not, as is often supposed, that they reduce the living standards of native workers. A 10% increase in the number of immigrants decreases the average native wage by at most two-tenths of 1%, and has little effect on the unemployment rate of practically all native groups, including the black poor.

In fact, even a large and unexpected increase in the supply of immigrants -- such as the arrival of 125,000 Cubans into the Miami area during the 1980 Mariel boatlift – does little harm to native earnings and employment. The trend in the wages and unemployment rates of both black and white natives in Miami over the 1979-1985 period differs little from the trends observed in a number of comparable cities, none of which received anything like the Miami influx.

But though unskilled immigration doesn’t deprive natives of jobs or wages, it does have other costs. The poverty rate of newly arrived immigrants admitted in the late 1960s was five percentage points higher than that of natives, while the poverty rate of immigrants admitted in the late 1970s was 18 percentage points higher. Similarly, immigrant households admitted in the late 1970s are about two percentage points likelier to receive public assistance than immigrants who arrived in the early 1960s.

The differences in poverty and welfare recipiency rates among national origin groups are huge. Only 7% or 8% of the immigrants from Britain or Germany are living below the poverty line, but 14% of Koreans, 26% of Mexicans, and 34% of immigrants from the Dominican Republic are. The native poverty rate is 12%. Similarly, only 5% of German and British households are on welfare, but 10% of Filipino households, 17% of Cuban households and 26% of Dominican households receive public assistance. Eight percent of native households live on welfare.

The shift toward a more unskilled immigrant flow increases the burden of income transfer programs. There really is a fundamental conflict between the welfare state and immigration. Before welfare benefits became widely available in the 1960s, prospective immigrants to the U.S. would make their decision based on a comparison of the economic opportunities available to them here and in their country of origin. Now they compare welfare opportunities too. Without welfare, the U.S. could open its borders entirely, knowing that those who were needed would stay, while those who weren’t would probably leave. National income and tax revenues are substantially lower than they would have been if the U.S. had attracted more skilled immigrants. If the people who immigrated in the late 1970s had been as skilled as those who came in the early 1960s, national income would be at least $6 billion higher and tax revenues would have increased by $1.5 billion per year.

In both Canada and Australia, visas are now allocated through a point system, which grades visa applicants in terms of educational attainment, age and occupational background. The presence of relatives in the country is only one factor among many. Canada and Australia also "sell" visas to persons who have sufficient financial resources to open businesses and create employment opportunities for natives. But it may well be, for instance, that the skilled persons who choose Australia or Canada would have preferred the more favorable tax policies or labor market conditions provided by the U.S. If American law were different, they might have come here instead.

Of course, a visa allocation system based on the applicant’s ability to pay discriminates against people who lack these financial resources. Similarly, a point system discriminates against people who lack the favored skill characteristics. Any visa allocation system, however, is bound to lead to inequities, and the inequities that would be created by a merit-oriented immigration policy would be no more egregious than those associated with present or previous immigration policies. Throughout the first half of the century, the nationalorigins quota system flatly prohibited the entry of Asians; current policy prevents the entry of most persons who do not have relatives in the U.S.

Immigration policy is inherently discriminatory. It selectively picks and chooses from the many applicants. It may stress national origin, or skills, or financial resources, or family ties, or any combination of characteristics that Americans deem economically and politically desirable, and consistent with the country’s values and beliefs. Because there are only a limited number of visas, the policy

```
has to restrict or prohibit the entry of many classes of persons.
Inevitably, difficult choices must be made.
   ---
   Mr. Borjas, a professor of economics at the University of
California, Santa Barbara, is the author of "Friends or Strangers: The
Impact of Immigrants on the U.S. Economy,"
(Basic, 1990).
</TEXT>
</DOC>
```

Appendix B

The *information file*

A sample part of the *information file* is given.

```
a 14 WSJ900619-0004 12517   ....
h 14 WSJ900619-0004 14641   ....
b 1  1  1  ......................
b 1  2  1  ......................
b 3  465  1  ....................
b 3  500  1  ....................
b 1  7  1  ......................
b 6  aboard 1  .................
b 6  accord 1  .................
b 2  ad 1  .....................
b 4  agre 1  ...................
b 7  aground 2  ................
b 6  appear 1  .................
b 4  area 1  ...................
b 4  awai 1  ...................
b 3  bai  2  ....................
b 4  barg 5  ...................
b 5  beach  1  ..................
b 7  bermuda 1  ................
b 4  boom 1  ...................
b 8  bouchard  1  ...............
b 7  brannan 1  ................
b 7  buzzard 2  ................
b 5  capac  1  ..................
b 5  cargo  1  ..................
b 5  carri  2  ..................
b 4  caus 1  ...................
b 5  clean  2  ..................
b 7  cleanup 2  ................
b 4  clhb 1  ...................
b 5  coast  4  ..................
b 10 contractor  1  ............
b 4  corp 1  ...................
b 5  cruis  1  ..................
b 5  damag  2  ..................
```

```
b 6  depart 1  .................
b 8  discharg  1  ...............
b 5  doesn   1  ..................
b 3  don  1  ....................
b 8  downwind  1  ...............
b 6  easili 1  .................
b 9  elizabeth 1  ..............
b 3  env  1  ....................
b 7  environ1  ................
b 3  fog  1  ....................
b 4  foot 1  ...................
b 4  forc 1  ...................
b 7  founder1  ................
b 4  fuel 2  ...................
b 6  gallon 3  .................
b 6  ground 2  .................
b 5  guard   4  ..................
b 4  gulf 1  ...................
b 6  harbor 2  .................
b 5  haven   1  ..................
b 4  heat 1  ...................
b 7  heavier1  ................
b 4  hour 1  ...................
b 6  inform 1  .................
b 6  integr 1  .................
b 4  know 1  ...................
b 4  late 1  ...................
b 4  leak 1  ...................
b 4  life 1  ...................
b 5  major   2  ..................
b 12 massachusett   1  ..........
b 6  mexico 1  .................
b 4  mile 1  ...................
b 8  minuscul  1  ...............
b 3  mob  1  ....................
b 5  mobil   1  ..................
b 4  near 1  ...................
b 3  new  1  ....................
b 5  offic   1  ..................
b 6  offici 5  .................
b 3  oil  13 ...................
b 2  on 2  .....................
b 4  oper 1  ...................
b 8  optimist  1  ...............
b 5  owner   2  ..................
b 3  pai  1  ....................
b 6  partli 1  .................
b 4  past 1  ...................
b 3  pet  1  ....................
b 9  petroleum 1  ..............
b 5  petti   1  ..................
b 4  plai 1  ...................
b 4  poor 1  ...................
b 3  ran  2  ....................
b 6  residu 1  .................
b 4  role 1  ...................
b 6  ruptur 2  .................
b 4  rush 1  ...................
b 4  said 6  ...................
b 3  sea  1  ....................
b 6  seawat 1  .................
b 6  second 1  .................
```

```
b 4  ship 1 ...................
b 8  signific  1 ...............
b 5  spill  6 ..................
b 4  star 1 ...................
b 1  t  1 ......................
b 4  tank 2 ...................
b 6  tanker 1 .................
b 5  total  1 ..................
b 7  tourist1 ................
b 9  transport 2 ..............
b 3  trn  1 ....................
b 7  tuchman1 ................
b 6  unload 1 .................
b 6  vessel 3 .................
b 6  visibl 1 .................
b 3  wai  1 ....................
b 5  water  1 ..................
b 4  week 3 ...................
b 9  yesterdai 2 ..............
b 4  york 1 ...................
j 7  aground1 ................
j 7  allanna1 ................
j 4  barg 1 ...................
j 5  coast  1 ..................
j 7  journal1 ................
j 12 massachusett   1 ..........
j 3  oil  1 ....................
j 6  report 1 .................
j 3  run  1 ....................
j 5  staff  1 ..................
j 6  street 1 .................
j 8  sullivan  1 ...............
j 4  wall 1 ...................
```

Appendix C

Documents used for the experiment

The documents that were handed to each one of the subjects in the experimental procedure are given in this section.

Instructions

- You will be given a set of 5 queries
  - For each query you can only spend a maximum of 5 minutes
  - You are presented with the 50 first matching documents for each query (broken into 5 pages, of 10 documents each)
  - During the 5 minute period you are asked to find documents that you consider as being relevant to the query
  - For each document in the ranking list that you think it is relevant to the query, check in the respective box on the query's answer paper
  - Please try to examine the 50 documents in the order they are presented in the list, without skipping any
  - If you finish with all of the 50 documents before the 5 minutes period expires, please let me know
- Do not immediately start working on the next query. You will be asked to do so after the 5 minutes time for the previous query has expired.
- Finally, after you have finished with the five queries, please take some time to fill in the simple questionnaire that has been handed to you.

QUERY NO. 41

Your query is:" Military Coups D’etat"

A relevant document will identify the country involved, the group responsible for the coup or coup attempt, the target of the coup, and the motivation of the coup plotters. It should NOT be about civilian government shake ups.

For each document you judge as relevant, check in the respective box.

□ 1
□ 2
□ 3
□ 4
□ 5
□ 6
□ 7
□ 8
□ 9
□ 10
□ 11
□ 12
□ 13
□ 14
□ 15
□ 16
□ 17
□ 18
□ 19
□ 20
□ 21
□ 22
□ 23
□ 24
□ 25
□ 26
□ 27
□ 28
□ 29
□ 30
□ 31
□ 32
□ 33
□ 34
□ 35
□ 36
□ 37
□ 38
□ 39
□ 40
□ 41
□ 42
□ 43
□ 44
□ 45
□ 46
□ 47
□ 48
□ 49
□ 50

QUESTIONNAIRE

Name:

1. How would you rate the complexity of the queries in general:

1 2 3 4 5

[Not complex at all] [Very complex]

2. How much do you think the accompanying text shown for each document (not the full document text), helped you in your judgments:

1 2 3 4 5

[Very much] [Not really]

Any further comments, opinions

Appendix D

Experimental results per subject

In this section, the experimental results are presented, as obtained from each individual user. First the ‘summary’ group of subjects is presented, and then the group that used a typical IR system output. It should be noted that when the value for the *success rate* can not be defined (when no relevant documents exist up to the last document that the subject examined), this is indicated by the presence of a ‘-‘ in the respective table cell.

‘Summary’ group

	Success rate (%)	**Utilisation (%)**	**Number of documents examined**	**Accesses to full text**
User 1				1
Query 02	50	6.66	22	
Query 07	100	27.3	25	
Query 08	100	50	33	
Query 32	100	100	37	
Query 36	77.7	63.7	30	
User 2				1
Query 03	60	100	19	
Query 25	42.8	100	20	
Query 40	100	100	20	
Query 43	-	0	33	
Query 46	50	60	20	
User 3				3
Query 04	71.4	100	24	
Query 05	75	70.6	38	

Query 37	100	36.4	50	
Query 41	66.6	16.7	29	
Query 47	0	0	50	
User 4				3
Query 06	0	0	7	
Query 09	75	37.5	13	
Query 13	100	50	16	
Query 23	50	14.3	16	
Query 33	14.2	50	14	
User 5				0
Query 10	25	100	21	
Query 11	29.4	100	20	
Query 22	66.6	50	14	
Query 28	33.3	66.6	9	
Query 42	66.6	100	13	
User 6				1
Query 12	-	0	11	
Query 20	100	42.9	10	
Query 24	100	50	7	
Query 31	80	57.2	10	
Query 44	88.8	88.9	11	
User 7				2
Query 15	100	21.2	16	
Query 29	100	66.6	7	
Query 38	100	17.4	24	
Query 39	50	100	41	
Query 48	0	0	50	
User 8				1
Query 16	100	16.7	28	
Query 27	57.8	91.7	27	
Query 30	-	0	45	
Query 34	0	0	25	
Query 45	60	90	25	
User 9				0
Query 17	-	0	11	
Query 19	50	50	9	
Query 21	-	100	37	
Query 35	-	100	16	
Query 50	100	40	14	
User 10				3
Query 01	62.5	71.4	25	
Query 14	75	75	16	
Query 18	77.7	87.5	20	
Query 26	-	100	23	
Query 49	-	100	29	

‘Typical IR system’ group

	Success rate (%)	**Utilisation (%)**	**Number of documents examined**	**Accesses to full text**
User 1				13
Query 05	66.6	88.9	20	
Query 07	75	33.3	26	
Query 29	16.6	50	20	
Query 44	27.2	60	31	
Query 45	55.5	83.4	39	
User 2				12
Query 02	0	100	16	
Query 04	66.6	33.3	19	
Query 15	66.6	33.3	17	
Query 30	-	0	36	
Query 47	0	0	50	
User 3				35
Query 17	-	0	4	
Query 25	100	83.4	10	
Query 40	100	25	13	
Query 43	-	0	13	
Query 46	50	50	12	
User 4				27
Query 14	100	33.3	7	
Query 22	100	33.3	12	
Query 28	85.7	85.8	10	
Query 39	-	0	12	
Query 41	50	7.2	18	
User 5				6
Query 01	33	28.6	20	
Query 20	28.5	25	28	
Query 36	30	60	46	
Query 37	33.3	33.3	32	
Query 49	-	0	29	
User 6				19
Query 11	33.3	100	11	
Query 19	33.3	50	30	
Query 23	0	0	24	
Query 31	33.3	50	27	
Query 38	33.3	33.3	21	
User 7				27
Query 06	100	100	5	
Query 13	0	0	6	
Query 18	80	100	10	

Query 24	100	50	5	
Query 26	-	0	12	
User 8				22
Query 08	100	50	13	
Query 10	0	100	33	
Query 12	-	100	29	
Query 34	50	66.6	18	
Query 50	100	66.6	23	
User 9				22
Query 03	60	100	18	
Query 33	22.2	40	16	
Query 35	-	0	31	
Query 42	7.6	100	18	
Query 48	0	0	31	
User 10				54
Query 09	60	42.9	20	
Query 16	100	20	12	
Query 21	-	0	14	
Query 27	42.8	100	19	
Query 32	0	0	40	