

“Finding Descriptions or Definitions of Words on The WWW”

Department Of Information Studies
University Of Sheffield

By Ying Ki Liu
MSc Information Systems

September

2000

Abstract

The information explosion problem is recognised in Information Retrieval. The advent of the Internet and the World Wide Web has exaggerated the problem, documents written over this medium, are free from the constraints of ordinary publishing. The need for a search tool to navigate and make sense of this information is no more apparent than now. Question Answering systems which incorporate many disciplines outside of Information Retrieval is one such tool that can be developed to counter this problem.

The aim of this project is to extract definitions or descriptions of noun phrases from the World Wide Web. This was done by implementing and evaluating a system that retrieves ranked descriptive phrases. The system used a simple pattern matching algorithm to detect candidate answers for queries such as, “What is a description of NATO?” or “What is a description for Steve Jobs?”. This was because other techniques were considered too complicated or required domain knowledge. The free nature of the Internet required a domain independent algorithm. Evaluation metrics used were borrowed from Information retrieval, Information Extraction and the Question Answering Track of TREC.

The results obtained indicate that a pattern matching technique is a viable method of retrieving descriptive phrases. In all, 227 queries were tested against the system, and 96 were used in binary relevance test. Results were achieved for the twenty answer ranks returned and at least 6 rank positions for a query imparted some information. In a separate task, where the answers obtained for 50 queries were compared against a key answer, 86% of key answers can be found in the ranked answers returned to the user.

Contents

Chapter 1. Introduction	5
1.1. Aim And Scope	6
1.2. Brief Methodology	6
1.3. Guide To This Paper	7
Chapter 2. Literature Review	8
2.1. Information Retrieval	8
2.2. Information Extraction	9
2.3. Automatic Text Summarisation	11
2.4. Differences Between The Project And Summarisation or IE or IR.	12
2.5. Question Answering & Systems	13
2.6. Evaluation	18
2.6.1. IR Evaluation	18
2.6.2. IE, Summarisation and QA Evaluation	20
2.6.3. IR Techniques	21
2.6.4. IE, Summarisation and QA Techniques	28
2.7. Project Techniques	31
2.7.1. Pattern Matching	31
2.7.2. Ranking	31
Chapter 3. Methodology & System Implemented	32
3.1. System Architecture	32
3.1.1. User Interface Module	33
3.1.2. Information Source	33
3.1.3. Document Opener & Tag Remover Module	34
3.1.4. DPD System	34
3.2. Experiment Description	35
3.2.1. Query Selection	36
3.2.2. Broad Evaluation Selection	36
Chapter 4. Results & Evaluation	39

4.1. Introduction	39
4.2. Effectiveness	39
4.3. Efficiency	40
4.3. Relevance	43
4.3.1. MRAR Results (considering the first five rank positions only)	43
4.3.2. Extended MRAR Evaluation (considering the first ten ranks and twenty ranks)	44
4.3.3. Binary Test	45
Chapter 5. Discussion	46
5.1. Architecture	46
5.1.1. User Interface	46
5.1.2. Information Source	46
5.2. Evaluation	47
5.2.1. Query Sets	47
5.2.2. Judge	47
5.2.3. Assessors	47
5.2.4. Effectiveness	48
5.2.5. Efficiency	48
5.2.6. MRAR	49
5.2.7. Extended MRAR	50
5.2.8. Binary Test	50
5.3. The System As A Whole	51
Chapter 6. Conclusion	54
6.1. Overview	54
6.2. Further Work	55
References	57
Appendix A: Queries For Binary Test	63
Appendix B: Queries For MRAR Tasks	66
Appendix C: Key Answers	68
Appendix D: Sample Answers	70

Chapter 1. Introduction

With the advent of the World Wide Web(WWW) and the subsequent information explosion. The need for research into methods that can handle sort and make sense of the information present is no more apparent than now. In order to address this query, research has been prominent in the areas of Automatic Text Summarisation(Mani & Maybury(1999)), Information Extraction(Gaizauskas et al.(1998)), Question Answering(Kupiec(1993)) and Information Retrieval.

The different approaches taken by researchers has enabled alternative methods to be used to extract, condense, store and summarise information or to provide answers using queries. It is recognised that the areas shown are not exclusive but rely on the interaction between one another in order to achieve their own objectives. In the areas shown many systems and tools have been developed to tackle the problem of excessive information. In Automatic Text Summarisation, the aim is to produce concise summaries of the documents queried upon. Information Extraction techniques are typically used to extract information from documents and store them as entities for later regeneration, or the creation of a knowledge base. Question Answering is a relatively new area, where the central aim is to retrieve or generate answer passages for a user queries. Fundamentally, Information Retrieval in the classic sense is research in the area of document or reference retrieval. This means that when an user query is given, Information Retrieval systems typically retrieves the document or the reference to that query and not an answer in itself.

The approaches described above seem to point to the fact that tools are needed to deal with the information problem. The most direct source of information described is the idea of Question Answering systems. For example, when a user types a query into a web search engine, (s)he expects a ranked list of document or resource references to be displayed. However what the user really wants is an answer or answers to the query and not references. Quite simply, conventional Information Retrieval systems require more effort to be expended by users to answer a query. Whereas, little effort is required by users in using a Question Answering system.

1.1. Aim And Scope

In order to answer some of the problems associated with the above, the aim of this project is to build and especially evaluate a system that can extract descriptions of noun phrases from the WWW. To this end an user interface to an existing program that retrieves descriptions from free text documents needs to be implemented. The aim is not to produce full Question Answering system, but one which restricts the question type to the description of noun phrases or acronyms. The restriction is placed on the system because Information Extraction techniques although are useful require domain knowledge and parsing. The system is not required to detect precise and exact descriptions as it uses a simple pattern matching approach to extraction. This approach enables multiple descriptions to be extracted from resources of the WWW.

In a way, this system may at first sight be similar to on-line encyclopædias. However, because of the dynamic nature of the WWW, the descriptions extracted to the user will reflect the documents generally available. This is in contrast to on-line encyclopædias where the content is controlled by an author. Some work in developing systems that extract acronyms already exists(Yeates(1999)). An important WWW resource where one can find definitions for acronyms is already available(WWW0).

1.2. Brief Methodology

The main aim stated is to produce a system and evaluate the effectiveness. To build the system an user interface needs to be implemented and be somehow integrated with the existing system that detects noun phrases. It is known that the existing Descriptive Phrase Detection system uses pattern matching as it's detection algorithm. This is because of the simplicity of the approach compared to Information Extraction or Automatic Text Summarisation. The evaluation that takes place seeks to understand if such a simple approach can be used to fulfil a partial Question Answering task, and it's potential as a full Question Answering system.

In order to evaluate, three different points require attention, they are, effectiveness, efficiency and relevance. Effectiveness addresses the performance of

the pattern matching algorithm as a extraction tool. Efficiency attempts to understand how the system is affected under different loads. Finally, relevance is needed to actually assess the output of the system.

1.3. Guide To This Paper

The rest of this paper consists of :

- A literature review.
- The methodology issues of the project.
- The actual evaluation performed.
- The results achieved.
- A discussion and conclusion to the outcome of the project.

The primary focus of the literature review is on the evaluation aspects of systems that are related to the one produced. Parallels between Automatic Summarisation, Information Retrieval, Question Answering systems and Information Extraction evaluation techniques are drawn and analysed. A brief introduction to the above research areas is also covered.

The methodology issues surrounding the project is presented in a chapter of it's own right. This covers the methods of evaluation, data collection, the quality of the data, the quantity of data and a full appraisal of the architecture of the system.

The evaluation and results chapter aims to collate the results that were achieved with the system in place. Various graphs and tables show the performance of the system in different ways. Amongst those measurements includes efficiency as a function of time, effectiveness of the pattern matching algorithms and relevance to user needs.

The discussion part of this paper is a dissection of the system and analyses the need and possible development of system in light of the results.

A conclusion is needed to ascertain the full impact of the research project. The findings will be summarised and a summary produced for possible further development.

Chapter 2. Literature Review

The system devised tested and evaluated falls into many different categories of Information Retrieval(IR). The areas that are of major significance to the project are Question Answering Systems and Evaluation. As well as IR some relevance can be associated with the area of Computer Science known as Natural Language Processing(NLP) and particularly Information Extraction(IE). The other area where the system overlaps with is Automatic Text Summarisation. An overview of the relevant areas will be given.

2.1. Information Retrieval

Since the project falls into the domain of IR, a brief description of the topic is required. According to (Sparck Jones & Willett (1997)), text retrieval has become synonymous with document retrieval which in turn many researchers call information retrieval. The primary processes involved in IR consists of the two main acts, that of Indexing and Searching.

Indexing is described as:

“the way documents, i.e. the items in the file, and requests, i.e. expressions of the user’s information need, are represented for retrieval purposes.”

The Searching act is described as:

“the way the file is examined and the items in it are taken as related to a search query.” (Sparck Jones & Willet(1997:1))

Whilst Indexing and Searching has been the dominant area of study for IR since the inception of the term, there is a need for IR to adapt to the changes brought about by the mass acceptance of the WWW. The areas that requires further study focuses on compliments to the main two acts, which are centred around information seeking behaviour and user-interface interaction.

Examples of interesting IR systems include CODER produced by (Fox & France(1987)). This system applied Artificial Intelligence(AI) techniques to the problem of IR. In particular the system used a series of expert modules which specialised in a predefined task, communicating via a blackboard approach. The

two most promising features of the system, aside from the use of AI, is the use of collections of composite documents used as an information source and, the retrieval of passages of text, as well as the references of documents or documents themselves. This was done at a time when the WWW was not visible and comparisons can be drawn with this project. This is because the information source of this project is comprised not only of full text documents written in a recognised style but also documents containing pictures, graphics and other non relevant items to the task.

This brings up a potential problem with the information source in that as the internet is not only made up of many professional publishers and writers but also it encompasses a broad range of amateur writers.

2.2. Information Extraction

This part of Computer Science majors on the development of systems that analyse and extract specific information from unrestricted amounts of text. (Lehnert(2000)) goes on to suggest that Information Extraction(IE) systems do not attempt to understand the source document completely but merely analyse those specific parts of the text that is relevant for a given query.

IE systems tend to use templates that describe the topic of the information in a frame format, where individual slots within a frame pertain to certain types within the topic. For example (Soderland(1997)) describes a system called CRYSTAL which attempts to extract information from WWW sites that host weather reports. The slots or fields of information in this instance are related to three different constraints that have to be met for a concept definition to be extracted. The concept definition refers to the rules that CRYSTAL has to apply to different domains, in this case, it is weather reports.

There are problems to this approach, (Chowdhury(1999)) sees these as the use of lexicons to understand words, the constraints of natural language and the problem of knowledge base updates and creation.

The issue of lexicons to control vocabulary in NLP systems is perceived to be problematic because of the nature of natural language. Natural language that is, the language in everyday use changes over time and the meaning of words also may be changed. So the compilation of dictionaries to control the vocabulary of systems becomes difficult, as not only creation needs handling but also updating. This problem can be partially solved by the construction of a limited vocabulary located to a specific domain. For example, the 5th MUC conference (Sundheim(1993)) focused the task on two domains, joint ventures and microelectronic chip fabrication.

The problems inherent in natural language such as the use of partial sentences, compound words, abbreviations(e.g. anaphoric references), coded references and incorrect grammar(e.g. unstructured phrases) are difficult to solve. A way to conquer these latter problems is to introduce a comprehensive parser into the system. This allows movement of the parser to catch all the problems that arise from natural language. In order to answer the former problem of compound words, systems need to stream to produce a conceptual understanding of a word given it's locality and inter-relationships with other words. In NLP this is typified by the semantic network or frame based approach of knowledge(conceptual) representation.

With knowledge bases(KBs) dominant in IE, the problem of creation and updates of KBs is very real. The questions that are asked are:

1. What do we include in KBs?
2. How to manage KBs?
3. What is acceptable performance for the system?

The work being done in IE and NLP in general requires a balance between the amount of knowledge present in a KB, the level of knowledge needed for a particular domain and the level of knowledge needed for acceptable performance of the system. The choice of techniques for KBs is reflected in the differing requirements of systems. For example the LaSIE system(Gaisauskas et. al.(1997)) developed by the Computer Science Department at the University Of Sheffield expanded on it's KB in order to extract information from texts in languages other than English.

The problem that still lies within the area of IE is that domain independent knowledge extraction is difficult. Hence the narrow focus for the MUC tasks enables research to be directed to specific domains, and therefore keep systems manageable.

2.3. Automatic Text Summarisation

The central paradigm of Automatic Text Summarisation systems is to produce a summary of a document(s) that is useful to the user in extracting the relevant points in the original document without losing too much detail.

(Sparck Jones(1999)) provides a description of what a summary should be in relation to the original document. It is stated that a definition is:

“A reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source.”

(Sparck Jones(1999:1)).

Further to this, the paper recognises that summarisation systems to date concentrate on two main points that is of text extraction and fact extraction. Text extraction systems attempt to extract information from source documents for use in summaries as it is presented in the original without prior knowledge of it's utility. This contrasts with the fact extraction based approach whereby a decision into what needs to be extracted is made before extraction begins. Sparck Jones calls these open and closed approaches respectively.

Many automatic summarisation systems exist, an important mention to Columbia University's Computer Science Department must be made. Throughout the 1990's Radev in collaboration, has produced many papers outlining systems that implement summarisation. In particular PROFILE in 1997 and SUMMONS a precursor to PROFILE.

(McKeown & Radev(1995)) explains the SUMMONS system. The primary objective was to summarise a series news articles which describe the same event. Development of such systems coincided with the maturing of the Message Understanding Conferences(MUCs) organised by ARPA, now known as DARPA. As such SUMMONS is typical in it's architecture relying on templates similar to that of IE for input. Output is generated by a two stage process which includes the

content planning stage and a linguistic component. The first stage of the process is the selection procedure, from the original document texts, knowledge is stored for later regeneration. The linguistic component selects words that correspond to concepts to include in a cohesive sentence. This secondary component usually involves rearranging the knowledge in the system to form English sentences.

The PROFILE system developed by (Radev & McKeown(1997)) extends the summary function present from the SUMMONS system. This is done by tracking prior references to given entities, which is especially important to the domain of news articles. For example, if a news article mentions a fact referring to a given entity and no prior knowledge of that entity is present, then important information in the summary would not be present. By using PROFILE, profiles of entities are stored in a database for later regeneration of articles.

2.4. Differences Between The Project And Summarisation or IE or IR.

Differences are present between the project and Summarisation/IE/IR systems. This is due to the fact that the methods employed in extracting the relevant sentences have certain properties that may not be deemed 'conventional'.

The dominant area of study for IR research relates to Indexing and Searching of documents. However the project aims to produce a system that extracts from the documents present on the WWW and not return a document itself or even a reference to it.

Summarisation techniques as the name implies returns a summary of a document that encapsulates the most important aspects of the original document. This contrasts with the system produced here because a summary is not generated in any way and no storage relating to later regeneration is applied.

Where does the system differs from IE? is a more difficult question to answer. IE systems as explained tend to use a slot or database approach to store entities, evidently the system described here does not require or use a storage mechanism after the extraction stage. More importantly, IE systems attempt to

understand in some way the original document extracted from, but this system does not have any relation to understanding. Also IE systems typically extract more than a sentence from a document in order to gain a thorough understanding of an entity, whereas this system will extract one sentence only.

Whilst some differences are mentioned and analysed to an extent, a topic of study closer to the project is the area of Question Answering systems. However, despite these differences the systems implementing any of the topic areas mentioned must all be covered by a common metric. That is the role of evaluation described later.

2.5. Question Answering & Systems

It is noted that IR has roots in document retrieval and reference retrieval. IE is centred around passage retrieval and storage research, whereas summarisation is a blend of technologies that attempt to summarise a document or multiple documents.

Question Answering is a discipline where systems are designed for a specific purpose that may encompass any or all the above areas. As the name implies, QA and QA systems attempt to answer some questions that are often posed in natural language queries. For example, a question posed to a system might be:

- *What is the price of oranges in the UK on average?*

Often given a collection of documents, a system should be able to retrieve relevant documents and process those documents in a way that an answer can be produced to a user. Systems that use approach have to deal with two types of questions, these are termed open and closed classed questions. The difference between the two classes lie in the answer expected from them. Open classed questions are different to closed classed because they typically expect a non definite, often unspecified length answer. This contrasts with systems that expect closed classed questions, which as opposites expect a finite and definite determinable answer. For a more concrete example:

- *What is the capital of the UK?*

- *What are the merits of Shakespeare?*

It can be seen that in the above example, the first question is of a closed class, since the answer expected is finite and definite, in this case the answer would be the word: London. Whereas in the second question, it can be deduced that the answer required is not definite and unspecified in length, since an answer would encompass many points concerning Shakespeare and his work.

Recently, in IR QA systems have been of interest, this has been somewhat stimulated by the QA track of TREC(WWW1). By adopting a new track, the principle aim of TREC is provide a focal point for QA research and development. This is done by providing common metrics for evaluation, sharing ideas and knowledge.

Fundamentally, QA is a part of NLP as the question posed by the user of the system will need analysis by a system using NLP, as well as processing the question, output often in NL needs to be sought. However as already stated, IR is of interest to QA systems because of the document retrieval stage.

For example, the MURAX system developed by Kupiec (Kupiec(1993), Kupiec(1999)) displays the characteristic described. MURAX attempts to answer closed classed questions using an on-line encyclopaedia. Parallels between the project and this approach can be drawn. The use of an encyclopaedia can be a metaphor for the set of documents in a collection. Kupiec broke the QA task into stages, at the first instance the question is dissected to reveal component noun phrase(s). Using these noun phrases, Boolean queries are formed to interrogate the encyclopaedia. Documents are then retrieved and analysed against the noun phrases. Using the documents retrieved, sentences containing one or more noun phrases are extracted with a weighting scheme applied. These form the basis of the answer(s) to the user.

In a later system by (Breck et al.(1999)), NLP approaches are mixed with IR and knowledge representation(KR) to tackle the TREC QA task of the same year. In the Qanda system described, the IR approach is used to rank the documents into a hierarchy using some relevance criteria or index. As the document collection is quite large, only a relatively small proportion of documents are actually used for

extraction and construction of answer purposes. NLP techniques are added into the blend in order to analyse the original question and the documents themselves. In response to the analysis of the original question/query, two internal queries are formed. These are the IR query and KR query. It is the IR query which handles the ranking of the documents in a format suitable to be processed. After selecting relevant documents from the top of the ranked list, a knowledge base(KB) is formed in response to the question. In turn, this KB is then queried to produce a set of candidate answers, which itself uses IR techniques to rank them.

Another interesting result from the QA task specifies the involvement of IE and IR in order to satisfy the QA purpose. (Srihari & Li(1999)) discuss their Textract QA system. Textract consists of three major components, that of the Question Processing section, the Text Processing section and a Text Matcher.

The Question Processing component is responsible for breaking down the input question into templates referring to IE's "Named Entities"(NE). A NE refers to templates in IE that share a type. In a NE template, types are restricted to person, organization, location, time, date, money and percent. In addition to the standard NE types, Textract can identify entities using it's own NE tagger. If a entity is not a standard NE then an alternative template is used. For example,

"Why did the Cultural Revolution Occur in China?" (Srihari & Li(1999:6))

In this question, quite clearly what the question requires is an answer that explains the reasons behind the Cultural Revolution in China. None of the standard NEs are useful in determining a question template. So the following template is constructed instead:

"asking_point: REASON

key_word: {occur, Cultural, Revolution, China}" (Srihari & Li(1999:6))

At the Text Processing stage, a search is made using the natural form of the question into a predefined search engine. This is done to trim the number of documents down in the collection for NE entity processing to take place. Once this has been achieved, the Text Matcher can be employed to matched entities created by the Question and Text processing stages. On completion relevant sentences are extracted from the documents and filtered to adhere to the requirements of the

TREC QA track. In this system the sentences are truncated to 50 bytes to comply with the task.

Thus, it is already observed that QA research is in a period of transition, a earlier approach by O'Connor is considered. At a time when IR research was focusing on the retrieval of document references, (O'Connor(1980)) produced an evaluation of passage retrieval for scientists. O'Connor recognised the utility of passage retrieval as much as conventional IR by benchmarking with the systems already used by lawyers in the form of FLITE and NEXUS. The experiment to evaluate passage retrieval to scientists was done to a medical statistical system called CANCERLIT.

An indirect QA system with focus on IR has been produced by researchers at MIT. In this case, (Chakravarthy & Haase (1995)) introduce the NetSerf system. In this system, the Internet is utilised in order to resolve a series of questions produced monthly by a game called "Internet Hunt". The questions published by this on-line game are resolved by NetSerf's Query Processor. Some assumptions are made regarding the query/question in order for it to work, at that time not all queries could be resolved automatically, so some questions needed to be rephrased. Following the completion of query processing, relevant information archives are produced as output following the matching stage. At the matching stage, the query is matched across information archives and a weighting factor added.

The interesting features of this research that may prove congruent to this project is the use of the Internet as a information source. Another feature that is also of interest is the fact that Miller's Wordnet(Miller(1990)) program is used to detect word collocations and disambiguate them. This latter feature is interesting because of the efficiency of the idea, it allows time to be spent tuning the rest of the system without sacrificing the integrity of the system. This may not prove to match the project produced by the author, but it does give an insight into how senses of words can be built into QA systems.

From the literature, it can be seen that the system introduced in this project cannot be termed a fully functioning Question Answering system. Since Question Answering systems need to fulfil answers for a range of different questions and the type of queries presented is somewhat limited. Also the communication with the system is limited to the query terms, and a narrow focus was made in the style of questions. It is noted that QA systems typically accept closed classed questions as input, this is in contrast with this project where the question is rather more open. In order to evaluate metrics must be determined as to the size of answers given and how these answers are assessed.

Finally, it is noted that QA systems tend to be restrictive in dialogue terms, when a query/question is posed to any given system, an answer is expected and dialogue consequently is terminated once the answer is given. Of course this need not be the case, if you needed to know the age of someone giving the name in the query, and the system returns a sentence related to the age, this would indicate success. But when you try to refer to this person in a preceding question without mentioning their name, for instance, a query relating to their hobbies, then possibly the system will not recognise this and produce an error. An example dialogue:

“USER1: I need a car.

SYSTEM1: Do you want buy or rent one?

USER2: Rent

SYSTEM Where?” (Jokinen(1993:169))

It is shown above that the PLUS system demonstrated by Jokinen is capable of reasoning in questions, and indeed the system can ask questions itself when the need arises. In the IR systems described so far the systems will terminate at the first statement as it is not a question. The system described is restricted in its domain, in that it provides an interface to Yellow Pages only. It could be termed that the goal to achieve in a QA system is contextual reasoning in dialogues in order to satisfy all user questions, and not just those styled in a way to suit a system..

2.6. Evaluation

Commonality exists between all the topic areas mentioned above. The underlying philosophy in all the disciplines is to produce a system that implements the subject area and evaluate it's worthiness. Overviews of IR, IE, summarisation and QA evaluation are discussed with a view to determine the relevance to the project.

2.6.1. IR Evaluation

(Saracevic(1995)) perceives evaluation as a major part of IR. This is because the whole process of evaluating helps the progress of the by critical analysis of the systems produced, and so strengths can be identified and weaknesses also ascertained. A definition of evaluation given by (Saracevic(1995:138)) is:

“Evaluation means assessing performance or value of a system”

To put this into the context of IR, the “information explosion” problem as realised by commentators such as Vannevar Bush, is the task of improving the accessibility of the knowledge of science and technology after the Second World War. In order to fulfil this void in management of information, IR was and is still predominant to reconcile the “explosion” problem.

Some fundamental questions are asked by Saracevic in the IR arena.

“How successful was and is information retrieval in resolving the problem of information explosion in the areas applied?”

“How well does IR support people in situations when they are confronted with problems of seeking, finding, using, and interacting with information from the mass of existing information and myriad of choices available?”

(Saracevic(1995:138))

To address these two problems is the ultimate goal of any system that attempts to handle any amount of information. In this context, the project can use these two questions as broad categories of relevance to IR and hence undertake some IR evaluation.

In addition to this, Saracevic comments on the differing approaches to IR evaluation. For example, questions that need to be answered in evaluation of IR are:

- Where should IR system evaluation begin and end?
- What and how to include in IR evaluation?
- What should not be included in IR evaluation?

As IR systems are typically composed of a series of modules that interact, the question of where do we start and end evaluation can be solved by splitting the system into it's components and evaluating it from a module to module basis. Although this problem of location is solvable by the modular approach, the objectives of the evaluation needs to be taken into consideration.

Saracevic points to six classes of evaluations. These are based on the levels of: 1) Engineering, 2) Input, 3) Processing, 4) Output, 5) Use and User and 6) the Social level.

At the Engineering level, it is necessary to look at the performance related to the architecture of the IR system produced. Objective focus should be on the computational ability, analysed with metrics that deal with hardware performance, software performance and maintenance.

At the Input level, the evaluation should be directed at the coverage of the IR area under investigation. A way that this could be achieved would be to observe the inputs and the content of the system.

At the Processing level, evaluation techniques centre around the desire to analyse the way in which inputs to the system are processed and converted to a output process format. The assessment of algorithms, techniques and approaches are valid to this level.

From the Output level, the principles of assessment could be focused on the interactions with the system and feedback. This is done to analyse the output of the system.

At the Use and User level, the IR system is evaluated as almost a consumer product, as criteria with relation to markets and fitness for purpose are focal points.

Finally at the Social level, evaluation observes the effect of the system at large. Environmental effects of the system are scrutinised, for example, questions directed maybe:

- 1) What is the effect of the system on current research?
- 2) How does the system benefit the world at large?
- 3) Is there scope for collaborative projects arising from the research?
- 4) How does the system project IR to wider audience?

The idea of modularity in IR evaluation is a worthy attempt at solving a difficult question, however there are some pitfalls. The most prominent being the fact that IR systems can be black box systems, meaning that although the system may comprise of a series of inter-connecting modules, the actual process being dealt in each one may not be visible. The other major factor must be that IR systems are produced with the focus being on evaluation at one level only.

As examples, (Brown(1995)) focuses the objective for the INQUERY system on the performance as a function of speed of retrieval. (Knaus and Schäuble(1994)) focus their evaluation not only on the efficiency of their system but also it's effectiveness by recall and precision methods, described later on.

General evaluation of IR has now been addressed, a look at the systems of IE, Summarisation and QA evaluation is in order.

2.6.2. IE, Summarisation and QA Evaluation

Much of the interest in the IE area has been stimulated by the tasks set out by DARPA's MUC programme. Evaluation plays a key role in the comparison of systems processes. Much of what Saracevic describes in the evaluation of IR is relevant to IE, this is in respect to the six classes of evaluations. The same could be applied to Summarisation and QA systems. Some definite evaluation techniques requires discussion.

2.6.3. IR Techniques

The techniques that are commonly applied to IR systems are the two metrics of Recall and Precision. These two metrics have at their heart one item of commonality. This is the notion of Relevance.

Relevance

(Saracevic(1975)) reports on the notion of relevance in use in Information Science in general. In this step of Information Science research, it is noted that the two measurements of Recall and Precision, have been in mass usage in IR systems since the beginning of the science. Central to Recall and Precision is the judgement of Relevance. Saracevic acknowledges that a definite paraphrased description of Relevance is not possible. The very attempt to condense a description for Relevance is perceived to be naive.

Thus it can be seen that in fact Relevance is a complex interplay of a number of factors which correspond to users and systems. Two early theories for Relevance are shown. These are the System view of Relevance and the Destination view of Relevance.

At the System view level of Relevance, Relevance is considered to be the actual processes involved in a IR system. This System and it's effect/relation to a question/subject/user is considered most important.

This is in contrast to the Destination view. Relevance at this view centres on the human judgement on the relation between documents of information and a question asked usually as a topic.

In addition to these views, were two main others, that of Pertinence and Pragmatic views. The former studies the relationship the knowledge of a subject and current literature. The latter Pragmatic view of Relevance is concerned with the relationship between the problem, as demonstrated to a user and the information provided. This view places focus on the use of information and preference as a base for selection of Relevance.

Whilst these explanations of Relevance are admirable in their own right, other research has postulated alternate theories. In a study of Relevance and its applicability in IR, (Mizzaro(1998)) concluded that Relevance exists as a point in four-dimensional space. At the highest level, Mizzaro states that Relevance exists as a relation between two main entities.

In the first group, documents, surrogates and information are members. This first group is cornered around the idea that information is an entity. That is the representation of information given to users is important in Relevance. The second group is centred on identifying the problem that a user needs to resolve.

It is these two groups that are accepted to be central in a relation to describe Relevance. However, Mizzaro insists that two other dimensions are of equal or more importance. In the third group, topic, tasks and context are members. Here the concept is to decompose the members of the second group that is queries, requests, information needs and problems into specific details. The idea is that the topic that the user is referring to, will affect the Relevance. Other factors that affect Relevance are the task to which retrieved documents are used for, the context with which interaction takes place. This context factor does not seek to examine the task or the topic involved but instead tests external factors. For example, the prior knowledge of users.

The final dimension that is of concern to first two groups is the effect of time. This is important to users of IR and users of IR systems as there is need to resolve a problem(the second group) using resources(the first group) in a context that involves time. There is a need to quantify or measure relevance at a given point in time. This is because the documents/surrogates/references returned to the user may not be considered to be relevant at a given time, but later perception may make it relevant. In effect the following diagram shows the situation regarding relevance as a point in four-dimensional space.

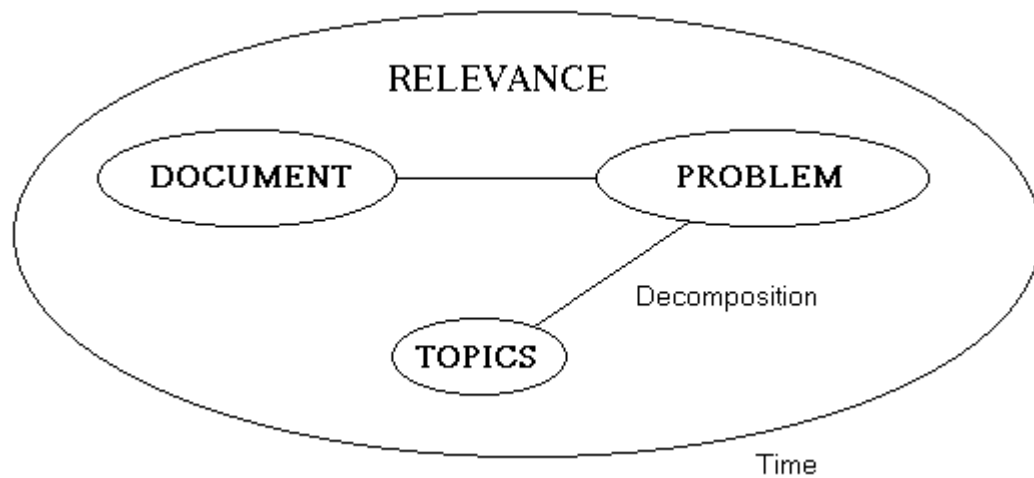


Figure 1- Pictorial Representation of Mizzaro's Four Dimensions of Relevance

In the above figure it is shown that time is of major importance in relevance serving as a background to the relationship between documents and problems. Problems itself is a relationship between topics that identify the context issues of the needs of users.

In a review of relevance (Cosijn & Ingwersen(2000)) declare that, despite the enormous amount of research in IR and Relevance, the concept is still not fully or well understood. In this study later models of Relevance are appraised, in particular a later model of Relevance system of Saracevic is dissected. The pertinent features of this model is the thinking that multiple levels of relevance exist along with different related attributes. Earlier work cited include the theory of Relevance developed at Syracuse University. Whereas Saracevic's earlier model (Saracevic(1975)) leaned more to a systems approach to the question of Relevance, the emphasis of Relevance developed by Syracuse University was on users. Relevance was defined by the information filtered to the user and information needs of the user. It can be observed that this concept places effort into the understanding of the quality of the relationship between information and information requirements of subjects.

Thus, two broad categories of Relevance has been established. On the one side, the systems approach centres relevance as a part of the system itself. This

provides contrast to the user approach to Relevance. Instead of placing focus on the actual systems in place, the actual interaction between users and systems are deemed more important. To clarify this latter approach, Relevance is more related to the actions and thoughts of users and the overall context of the interaction.

(Cosijn & Ingwersen(2000)) attempt to augment existing models of Relevance into a cohesive one blending together most of the theories postulated so far. To this effect, a system of relevance is described whereby a modified version of Saracevic's later system is used. Where Saracevic includes five different manifestations of Relevance labelled, System/algorithmic, Topical/subject, Cognitive/pertinence, Situational/utility and Motivational/affective, the revised model replaces the latter manifestation with one labelled Socio-Cognitive. In addition to this change, a time dependency factor is added as envisaged by Mizzaro. This time factor increases the dependency across attributes, with the Socio-Cognitive attribute being the most time dependent and Algorithmic least. Where the Socio-Cognitive attribute differs from Saracevic's Motivational./affective attribute is that the focus of the relation is on the problem at hand, perceived in a socio-cultural context.

To review so far, it has been seen that Relevance is considered in many different ways. Different levels of Relevance are modelled and different dimensions of relevance has been seen. However some of the shortcomings of Relevance requires some coverage.

In (MacCall(1998)) the problem of conventional metrics in IR to compare to Internet IR is examined. The basic underlying problem identified is the subjective nature of Relevance assessments. Relevance assessments are normally determined by an expert judge or users. Since humans are inherently different, different assessments may be made given the same information. Some studies cited attempt to counter this argument by explaining that the critical documents retrieved by a system is largely stable.

However, the possibility of variance is still present and as such can be considered as an error in measurement. To tackle this MacCall calls upon Kyburg's studies of error measurement. It is stated that:

“relevance judgments can serve as the criterion for IR performance evaluation can hold only as long as they are “ideal” (i.e., based on a valid theoretical derivation) and there is a method of determining error rates so as to quantify the error variation of that measurement procedure.” (MacCall(1998:20).

In order to achieve this ideal, the level of ability of all judges of Relevance to a given domain must be measured some way. That is, the cognitive ability of the judge prior to the task, the knowledge of the judge and their perception of the task.

An overview of the problems of assessment and describing Relevance has been addressed, but because of the typically large database environments that IR operates in and the subsequent high retrieval in numbers of documents(using any of the methods described previously), ranking becomes an important issue.

(Järvelin & Kekäläinen(2000)) recognises that most laboratory tests linked with Relevance were of a binary nature. Due to this restrictive scale of Relevance, different degrees of Relevance cannot be measured. With modern IR systems, results returned to the user consistently number more documents than users have time to examine. To demonstrate this, one could search on a web search engine using any query present in the English language and see the amount of results returned. Järvelin & Kekäläinen found that typically, users examine the top ranked documents in a results list only, with the documents near the end of the list receiving little or no attention. With this observation in mind, it can be shown that documents with a low rank number are more useful than ones with a high number. This is caused by the lower likelihood of users examining high numbered ranked documents. A simple way described to administer this bias is to introduce a discounting function for returned documents. This allows for more quantifiable value to be held by those at a lower rank position compared to those at a high value. However this discounting function needs to be aware of user's persistence in examining high numbered ranked documents. Järvelin & Kekäläinen suggest that the discounting

function be made but not projected too steeply, to allow weighting to be progressive across rank positions.

So far, it has been introduced that Relevance is central to IR, with different models of Relevance, a method to aid the problem of judgements and finally a weighting function for ranking relevant documents. The original dual measures in IR require attention.

Recall & Precision

Recall and Precision in an IR perspective are easy to calculate if the number of items in a collection is known. Succinct definitions of these two measures are easy to come by (Harman(1995:251&252), Chowdhury(1999:205)):

“Recall = $\frac{\text{number of relevant items retrieved}}{\text{total number of relevant items in collection}}$

Precision = $\frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}}$ ”

These two basic measures have been in use in IR since their introduction in the 1950's(Saracevic(1975)), their use now in the WWW environment is questioned by (MacCall(1998)). Since these measures were adopted from a time where the documents concerned are static and the fact that much of WWW content is dynamic leads to apparent formulae implementation problems.

In traditional IR, the number(quantity) of documents can be counted definitely as the physical mediums in which they were presented did not allow for changes. However in electronic environments where documents change dynamically, the question asked is whether a WWW source could be considered the same document after alteration. The argument is furthered by MacCall concerns the makeup of Recall and Precision ratios. It is stated that because both Recall and Precision are in fact derived metrics and not fundamental metrics, then there is need to keep the units of measurements in these derived metrics constant. The dynamic nature of the WWW intervenes with this in that document become different so an

actual number cannot be calculated. This leads to a possible situation where classic Precision and Recall measurements become unusable.

(Baeza-Yates & Riberio-Neto(1999)) seems to agree with this theory, stating that to calculate maximum recall requires detailed knowledge of all the documents in the collection. However in modern collections where size becomes a issue, the numbers cannot be captured easily. Further problems are identified by Baeza-Yates & Riberio-Neto. Recall and Precision measure only effectiveness over a series of queries based on processing in a batch mode. However most IR activities today centre around the interactivity of the system. To aid this latter interactive view, measures of IR systems must somehow measure or quantify the quality of the information in the retrieval process. In a user-oriented process, distinct measures are proposed but only coverage is relevant to this project.

$$\text{Coverage} = \frac{\text{Relevant documents retrieved that are known to a user}}{\text{Documents in collection that are known by users}}$$

Ranking

Ranking of documents has been covered briefly with a view of Relevance, a fuller understanding needs to take place. (Baeza-Yates & Riberio-Neto(1999)) describes the major difference between the three different models of IR, that of the Boolean model, the vector model and probabilistic models. There lies advantages and disadvantages with each model, however since ranking takes place in the project system, it is suitable to aim the discussion in the direction of the vector model.

The most useful aspect of the vector model is the move away from the assumption present in the Boolean model concerning use of binary relevance criteria to determine document usefulness. In the vector approach, it can be seen that index terms in queries are assigned non-binary weights, therefore a comparison in terms of the relativity of similarity can be obtained. This enables documents to be ranked in orders of similarity. The most widely used weighting scheme to achieve this vector model is a tf-idf scheme. A full derivation given by (Baeza-Yates & Riberio-Neto(1999:29)) is:

“Let N be the total number of documents in the system and n_i be the number of documents in which the index term k_i appears. Let $freq_{i,j}$ be the raw frequency of

term k_i in the document d_j (i.e., the number of times the term k_i is mentioned in the text of the documents d_j). Then, the normalised frequency $f_{i,j}$ of term k_i in document d_j is given by

$$f_{i,j} = \text{freq}_{i,j} / \max_l \text{freq}_{l,j}$$

where the maximum is computed over all terms which are mentioned in the text of the document d_j . If the term k_i does not appear in the document d_j then $f_{i,j} = 0$. Further, let idf_i , inverse document frequency for k_i , be given by

$$idf_i = \log N / n_i$$

This can be summarised by:

$$\text{Term weight} = \frac{\text{Frequency of term in a document}}$$

$$\text{Number of Documents in collection which term occurs}$$

Adapted from Chowdhury((1999:89))

2.6.4. IE, Summarisation and QA Techniques

Much of the IE discipline stems from the desire to accomplish tasks set out by the MUCs. Many of the tasks set borrow standard Recall and Precision measures for evaluation. (Voorhees & Tice(2000)) observes that the IE community is traditionally operated on the assumption that a definite finite answer key exists for comparison. This answer key encompasses all possible and acceptable responses. In this evaluation metric, any keys given by systems must match the answer key exactly. This was seen as a problem as different people have different perceptions on what is right and what is acceptable. Using this as a basis, it is said that a single comprehensive answer key cannot be constructed.

The main technique in IE for evaluation has been addressed, some methods of the use of evaluation in Summarisation is considered. In an early article by (Rath et al.(1961)), human relevance judgements were tested against machines. The experiment concerned the generation of abstracts by the selection of sentences in a document. This is similar to the task of Automatic Text Summarisation today. The

test that took place ascertained whether human judges will pick and rank twenty sentences from the document in the same way as an automatic abstracting program. It was concluded that little agreement between subjects and machines existed. This experiment seemed to show that in order to assess summarisation techniques, it is important to use one criteria for relevance only, hence ensuring all results obtained operated on a common level.

(Salton et al.(1997)) furthers this experiment in the evaluation of automatic text summaries. It is suggested that automatic generated summaries of articles are compared to human generated summaries. This evaluation was done by allowing a user to compare both summaries and then decide which one is best. In the paper, the term “satisfaction” was given as the criteria for measurement for assessing each summary. It was found that of the three algorithms of automatic text summarisation tested, 44-46% agreement was achieved in comparison with human generated summaries. These results are deemed acceptable to evaluate the performance of automatic text summarisers because agreement between different human generated summaries were of 45.81%, so human generated summaries performed on a level with automatic summarisation algorithms.

The main problem as perceived by (Firmin & Chrzanowski(1999)) is that all the evaluation methods that are used to compare summaries still rely on the fundamental notion that there is a single best or correct summary. Clearly, this is not the case since different people have different requirements and therefore seek different summaries. (Sparck Jones(1999)) advocates the comparison model, but alternative models are also proposed.

Sparck Jones notes that an important method of evaluating summaries is to compare it to the original source/document(s). Despite it's importance it is not rigorous enough in common implementation. Another idea reviewed is to examine summary texts to see if answers to questions can be obtained, that are obtainable from the original document(s). This latter method is more interesting as it allows a “quality of information” assessment. The methods described are all said to be limited in their utility if context factors are not taken into account. This suggests

that a framework for a summary evaluation strategy be developed that captures the context factor and therefore can obtain a level of purpose.

In QA systems, evaluation is relatively new area however a number of factors affect evaluation(adapted from Breck et al.(1999:4)):

- *“The types of questions, e.g. factual versus explanatory, context-dependent or not.*
- *The document genres, e.g., news stories, expository texts from encyclopedias, technical reports etc.*
- *The requirements for a successful answer, e.g. whether the answer is a document extract or synthesised, whether it is provided with document context”.*

Along with these factors exist two different overall methods. These are on-line and off-line evaluations. These two extend the idea of human and computer judges as used in automatic text summarisation. In an on-line evaluation, humans are used to score system answers, in an off-line evaluations programs score answers against a gold reference standard(Breck et al.(1999)).

Alternatively, earlier evaluation done on the original MURAX system by Kupiec(Kupiec(1993)) compared answers in the different rank positions produced by the system. In this evaluation, the questions used were of a very closed nature, in that all that a answer requires to the question is a simple noun phrase. This enabled Kupiec to determine whether an answer lies in the top rank, top three ranks, the top five ranks and a category excluding all these. Before any assessment actually took place, the on-line encyclopedia was tested by humans using common sense, to see that an answer exists for all the questions. Kupiec concluded that 74% of the answers produced by MURAX were in the top five ranks.

This study by Kupiec seems to have influenced the choice of metrics adopted by the QA track of TREC. In this evaluation(Voorhees & Tice(1999)), participants to the QA task were asked to submit five ranked answers to two hundred fact-based short answer questions. Restrictions were made on the answer strings submitted, in one run, answer strings were to be at most fifty bytes long, in another run the answer length could as long as two hundred and fifty bytes. Human relevance judgements

were made to each answer string, in order to score for an answer, that answer must answer the question unambiguously. In the answer ranks returned, the best answer judged would be used to calculate the reciprocal answer rank. For example, if the best answer returned was at rank position three then the reciprocal answer rank would be one third. Hence,

$$\text{Reciprocal Answer Rank(RAR)} = 1 / \text{Best Answer Rank}$$

The measure adopted by TREC was an average of all RARs of the questions resulting in a mean reciprocal answer rank(MRAR) metric.

2.7. Project Techniques

2.7.1. Pattern Matching

Pattern matching is a technique widely used in NLP to detect key phrases for QA or IE (e.g MURAX (Kupiec(1999), CIRCUS in (Riloff & Lorenzen(1999))). Kupiec hypothesised that answers to the question or query terms in the question will be found near the query terms. Using this hypothesis and some knowledge obtained from the question, MURAX was able to extract answers with high precision.

The system described uses a pattern matching technique because of two points, that of detecting the URLs in a web page returned to the user interface and to detect sentences containing the query phrase. The reason behind this is because pattern matching is perceived to be of use as it :

- Is domain-independent
- Requires no knowledge base
- Is relatively simple to implement

Due to the fact that the restriction on the question is of the type:

“*What is a description for X?*” Where X is the query noun phrase.

It can be deduced that certain patterns will more of use than others, hence the patterns used will be indicative of those that are used to describe noun phrases.

2.7.2. Ranking

The system will return ranked answers using a formula based around the tf-idf weighting scheme.

Chapter 3. Methodology & System Implemented

3.1. System Architecture

The architecture of the system described by this paper is best summarised by the following diagram:

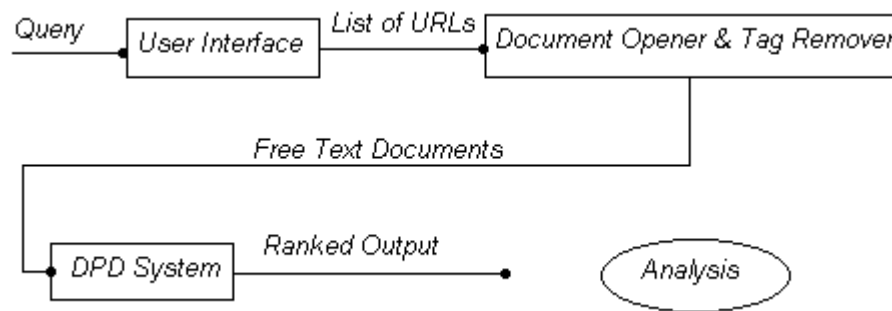


Figure 2: Architecture of the System

The system accepts queries of the form: “*What is a description for X?*”, Where X is the query noun phrase. However because of the lack of a NLP parser, the full question cannot be inputted, so an abbreviation is used instead. The question to be processed by the system is reduced to the noun phrase to be described. This noun phrase(query in the above diagram) is supplied to the user interface along with the number of URLs(Uniform Resource Locators) to be retrieved, and also an output file name. The user interface module processes this information to produce a file containing the URLs associated with the query from the web search engine. This file containing the URLs is then passed onto a module that removes the mark up language that is associated with web resources, in order to produce free text documents. This is then passed onto the DPD(Descriptive Phrase Detection*)

* This system was developed by a Hideo Joho, former student at the University of Sheffield, he also implemented the Document Opener And Tag remover module, now integrated with the original DPD system.

system along with the original query, which analyses the documents and produces a ranked list of 20 answer sentences.

3.1.1. User Interface Module

This module was designed and implemented by the author, this enabled users to select the number of URLs to be extracted from a web search engine, specify a relative file name to write the URLs to and input the query. Implementation of this module was done in the JAVA environment because of the desire inter-operating system compatibility/portability. However, because there was a need to retrieve the relevant documents associated with a query, this module was dedicated to one search engine(Google(WWW2)).

3.1.2. Information Source

As explained above a search engine was used in the retrieval process. This was decided upon because it is the easiest way to navigate the WWW. If this was not the case other tools would be needed to find and store documents from the WWW. The search engine decided upon was Google, this was chosen because of two reasons, the main one being that the source for the web pages in Google was easy to decipher and the relevant URLs were easily detected and stored. The second reason is that Google can handle Boolean queries. This was needed as the user can type a query in to the system without the '+' symbol in between query terms, this needs to be added into the query at the point where the web pages are returned. This is because an exact match between the query terms and the URLs pointed to is needed, so that pattern matching can be begun by the DPD system. For example, the query: Douglas Adams. In this example, Google may return URLs that point to documents containing one term only or both terms, since we are only interested in the latter the query that should be submitted to the search engine is: Douglas+Adams.

The use of a web search engine in the way described has benefits as programming additional tools is a resource intensive task as commented above. However there may be drawbacks in that the documents/resources that are interrogated by the DPD system is determined by that engine, hence there is no control on the part of the system.

3.1.3. Document Opener & Tag Remover Module

This module was implemented using a package that is available for the operating system implemented on, and hence is not described. However it is sufficient to know that it is capable of removing mark up from most WWW resources.

3.1.4. DPD System

This system is responsible for detecting all the sentences that contain the query terms from the collection. The free text documents returned by the Tag Remover module were analysed using pattern matching. It is known that the DPD system integrated uses the ranking method described in the previous chapter. The DPD system uses a pattern matching technique in order to detect candidate answers. In addition it identifies the number of sentences that match a particular pattern. Nine different patterns are detected by the system. These patterns are described in the table below with the conditions required for it to be detected.

Pattern	Condition(s)[where DP is a descriptive phrase and X is the query]
Is a	X is a DP
Acronym	... DP(X) ... or ... X(DP) ...
Such as	... DP such as X ...
Appositive	... X, DP, is/was ... or ... X, DP, were ... or ... X, which DP, ... or ... X, the DP, ... or ... X, a DP, ...
Such X as	... such DP as X ...
Especially	... DP, especially X ...
Including	... DP, including X ...
And other	... X and other DP ...
Or other	... X or other DP ...

Sentences that contained any the above patterns were given scores according to the system, which enabled answer ranking to take place. This pattern(boost)

score was combined with a sentence score to produce a combined score, it was this combined score that was used to rank the sentences returned to the user.

This pattern score was based on the following weighting applied to the sentence identified. It is also noted that sentences detected with the query terms but not with a pattern match scored a boost score of 1.

Pattern	Boost Score
Is a	9
Acronym	9
Such as	7
Appositive	7
Such X as	5
Especially	5
Including	5
And other	3
Or other	3

The sentence score was calculated by the DPD system(using a formula based upon Inverse Document Frequency(IDF)) by giving each sentence a score based on the significance of that sentence in the sentence collection. This sentence collection was merely all the sentences extracted with the query terms by the system.

3.2. Experiment Description

The experiment that took place involved a number of factors. These are broadly Query Selection and Evaluation Selection.

3.2.1. Query Selection

This was needed to select appropriate noun phrases to test the system with. The method used was to ask colleagues for suggestions and from these, a subset was taken. These were used as input to the system and an evaluation took place. There were two points to be made about the query, the first being that it must be a noun phrase and valid. The reason noun phrases are required is because of the desire to find descriptions of words that are not found in conventional dictionaries. For a query to be valid it must return some URLs.

3.2.2. Broad Evaluation Selection

This is required to ascertain what we need to measure from the system and hence what we need to satisfy these measurements.

- It was decided to show the effectiveness of the pattern matching approach, this would be done by calculating the distribution and coverage of the patterns.
- The efficiency of the system was also evaluated, in that the time taken to resolve queries was noted for some queries.
- An approach to ascertain the relevance of the system answers was needed. This was done by the appointment of a judge to read through the answers given by the system to queries. For this part of the evaluation, a thorough understanding of the system or even the query was not necessary. For this relevance assessment, a ranked answer scored 1 if some information regarding the query term was present, for example, “Tony Blair is a political leader”, for the query ‘Tony Blair’. A ranked answer scored 0 if no information whatsoever could be inferred from the sentence, for example, “Tony Blair, Tony Blair, Margaret Thatcher painted”, for the query ‘Tony Blair’.
- A secondary approach to assess the relevance of the answers is to compare the answers from the system to a key answer metric analogous to the technique used in IE. This approach required the construction of key answers for the system to be compared with. This was done by distributing questionnaires to four assessors to fill in. Scoring was done to observe whether the output answers from the system contained a ‘golden’ sentence in common with assessors’ opinions. To this end, investigation into whether the best answer was achieved in the top five ranks (as in QA in TREC), in the top ten ranks or in the top twenty ranks was made. To clarify this, the judge was to assess a key answer compared

to the first five ranks only and decide if there was sufficient correlation between one of the five positions and the key answer, to the warrant that position scoring as the best answer. This was then done with the first ten positions and then all twenty positions. For example, considering the first five ranks only, if the key answer was contained in the information given at rank position three in its entirety then that answer would be considered the best answer in the top five. Then considering the first ten ranks only, if a better answer than that at rank position three was found between ranks five and ten say for instance, at rank eight, then that would be considered the best answer in the top ten. Similarly for the twenty ranks. The answers given in the ranks did not have to match the key answer exactly but should impart a similar level of information.

3.2.3. Judge Selection

The judge was needed in the first relevance task that is to read through the answers given by the system in order to make binary relevance judgements. The judge was also needed to firstly combine the answers produced by the assessors, this was done in order to produce a key answer. The way this was done was to combine the judgements of the assessors into one key answer, where only one assessor answer was available for combination, then the judge supplied an extra one for combination. Combination is to include any description made by either assessor or made by both assessor into a clear concise sentence ending in a full stop(period) marker. Due to amount of effort involved in this task, the author was appointed to this role.

3.2.4. Assessor Selection

The selection of assessors was important in the second relevance task, this is because they were required to produce a English sentence for each query. This required that the assessors had some sort of knowledge of the query term otherwise random answers may be given. To alleviate this problem, the queries were divided up into two groups. Each group of queries were posed to two assessors to determine a key answer. The judge is appointed to resolve the key answer if only one answer was given by one assessor. No prior training was given to the assessors, just the instruction that for every query term given, to write down a full English sentence describing the query term.

Chapter 4. Results & Evaluation

4.1. Introduction

In total 247 queries were tested on the system, of these only 227 produced results that were of use, these are said to be valid queries. The remaining 20 queries produced either no output that could be assessed or the program failed for no obvious reason, hence these queries were discarded.

4.2. Effectiveness

This was measured by the coverage and distribution of the patterns. In all 227 valid queries were tested to evaluate this metric. Nine patterns exist, to ascertain a coverage number for each pattern the following steps are taken. For each valid query, if there is at least one sentence that matches a pattern, then that query is said to be covered for that pattern. The results are tabulated below:

Pattern	Number Of Queries Covered	Coverage Percentage
And Other	139	61.2%
Such As	141	62.1%
Appositive	202	89.0%
Is A	206	90.7%
Including	127	55.9%
Especially	27	11.9%
Such X As	46	20.3%
Or Other	32	14.1%
Acronym	178	78.4%

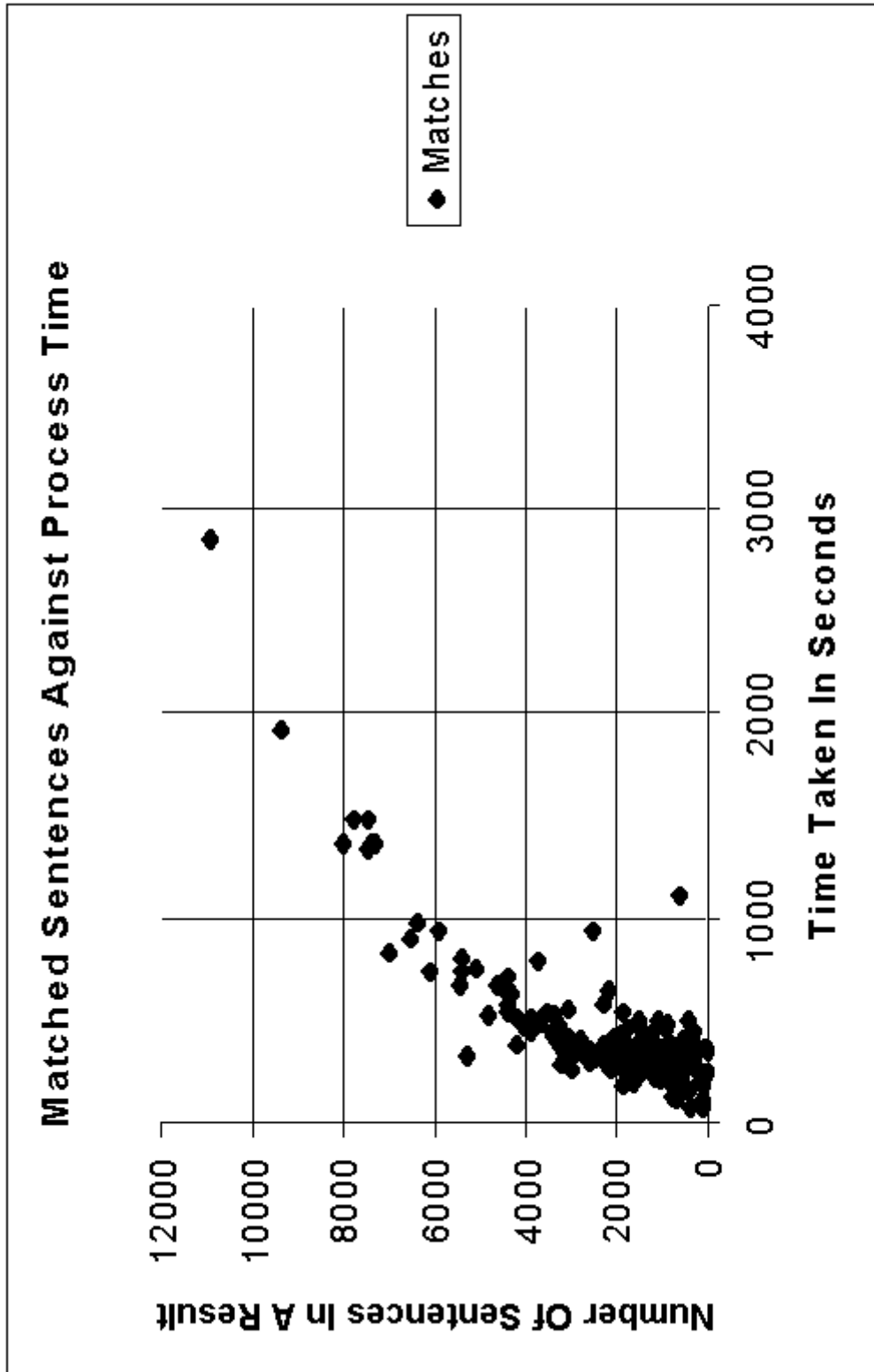
To calculate the distribution of the patterns, the total number of sentences that matched a pattern was calculated against the total number of sentences matching all

patterns. In the 227 valid queries, the total number of sentences that matched all patterns was 14180. The following table gives a breakdown of each pattern and its distribution percentage.

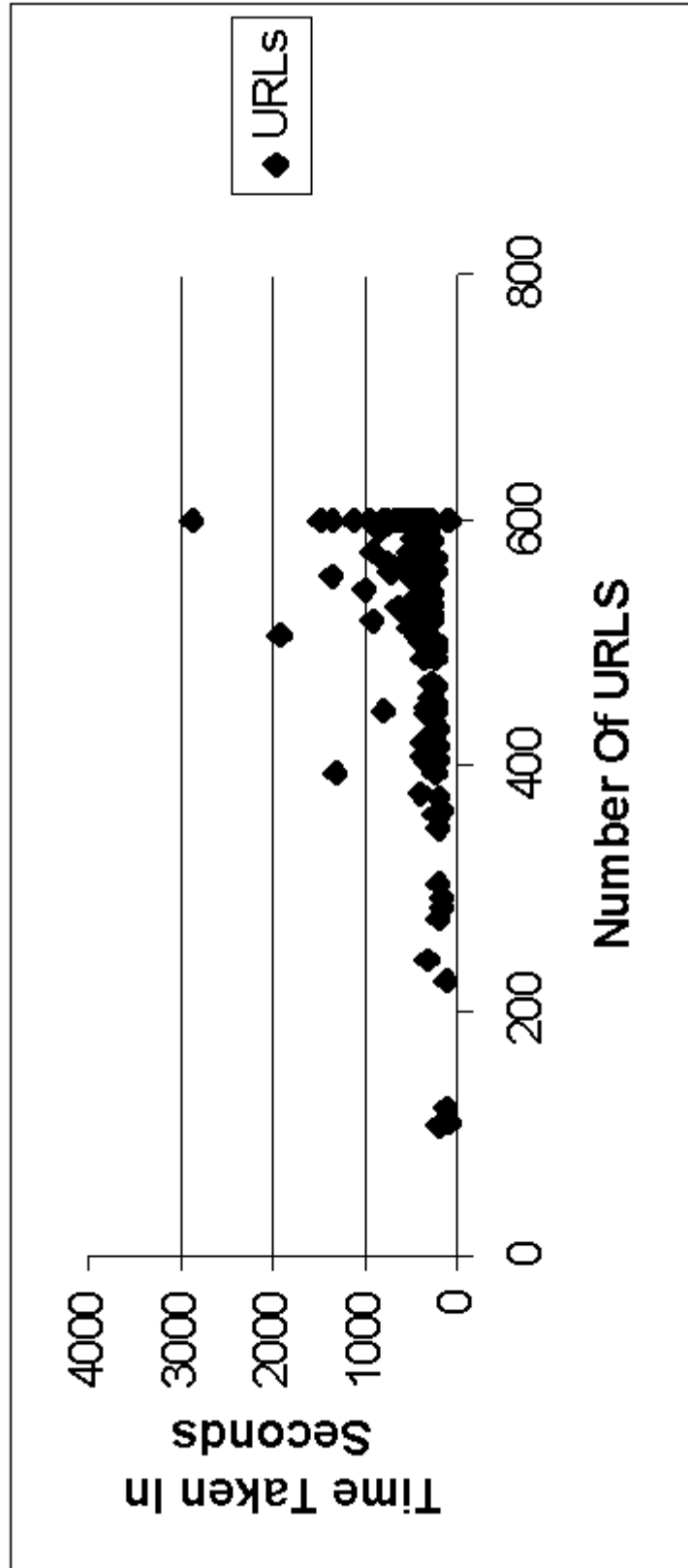
Pattern	Number Of Sentences Matching The Pattern	Distribution Percentage
And Other	719	5.071%
Such As	560	3.949%
Appositive	3037	21.417%
Is A	4335	30.571%
Including	375	2.645%
Especially	43	0.303%
Such X As	63	0.444%
Or Other	84	0.592%
Acronym	4964	35.007%

4.3. Efficiency

This was measured by timing how long it took for the system to process a query. This time measurement was compared against the two points. Firstly, against the number of documents retrieved for that query, i.e. in this case the number of URLs. Secondly, the time for a process was measured against number of sentences matched for a query. The following graphs show the results.



Process Time Against Number Of URLs Retrieved

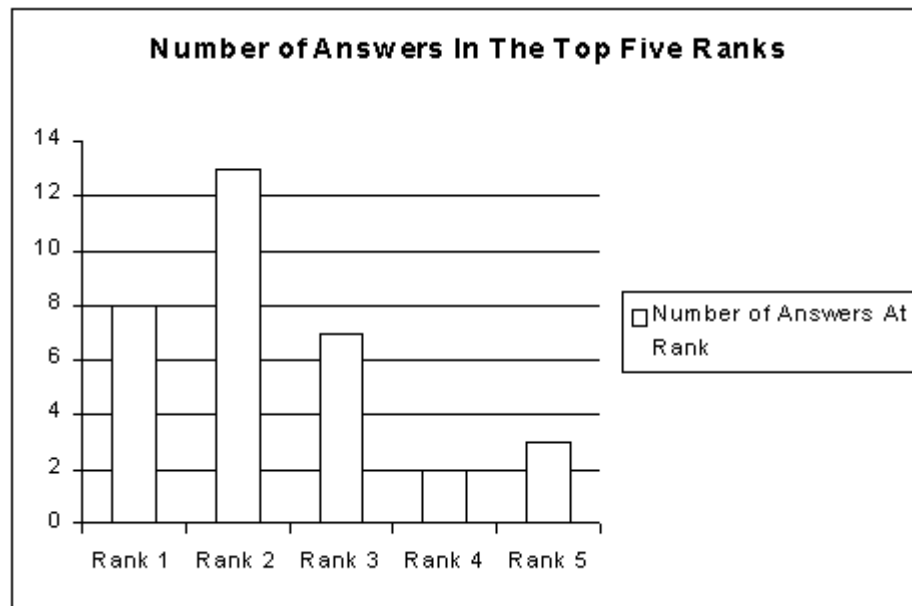


4.3. Relevance

This was achieved in two ways, first, by adopting the MRAR metric of the TREC QA track in comparing the answer given in the top five ranks to a key answer. This was extended somewhat in that after this was done, the answer was compared to the top ten ranks to see if there was a better answer there than in the top five, this was also done in the top twenty(all) ranks returned. The second way that this relevance was assessed was to see if the answer contained any information in relation to the query phrase, binary relevance judgements were made to each query's answer ranks. If there was information in an answer rank then that rank position scores one, if no information was present then that rank position scores zero.

4.3.1. MRAR Results (considering the first five rank positions only)

To evaluate this result, 50 valid queries were considered. The key answer was compared to each answer rank, the best answer scored one, other answer ranks scored zero. The following graph shows the number of answers in the top five ranks.



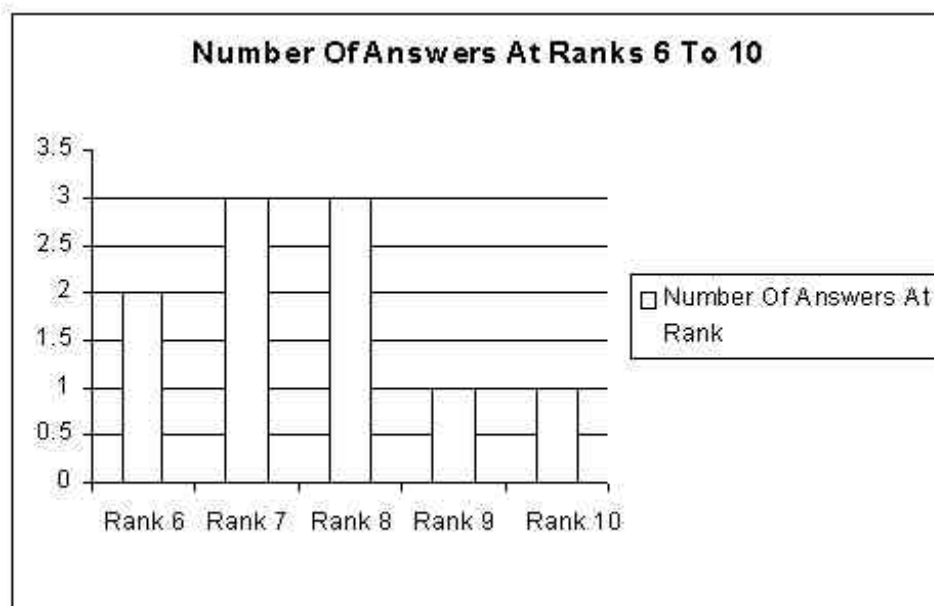
In this evaluation, it can be seen that 33 answers for 50 queries appear in the top five ranks, of these 28 answers actually appear in the top three ranks. To calculate the

MRAR metric used by the QA task in TREC, the reciprocal of the answer rank is taken, when this is done here the sum total for 50 queries is 17.933333, resulting in a MRAR score of 0.358666.

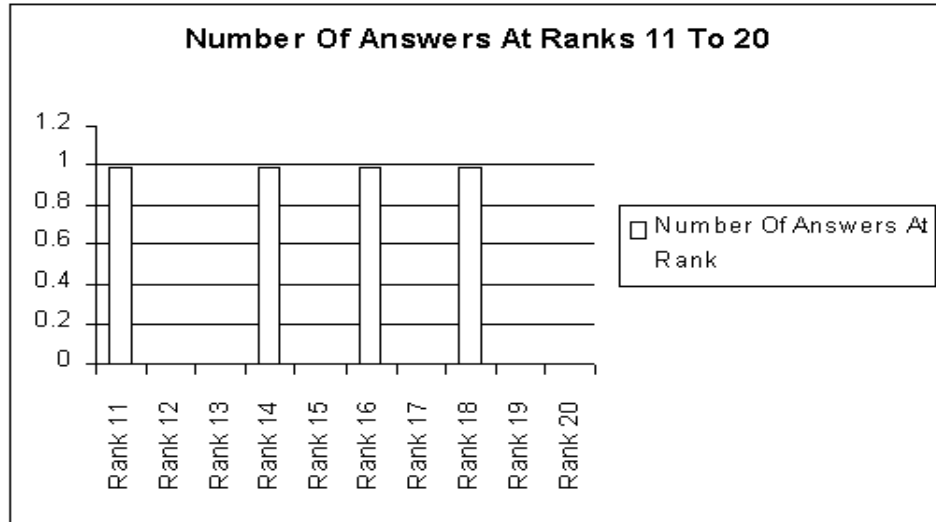
4.3.2. Extended MRAR Evaluation (considering the first ten ranks and twenty ranks)

At the beginning of the project, focus was placed on the evaluation aspect to the system. To achieve this goal, it was decided to investigate all the ranked answers for the fifty queries in the MRAR evaluation. Focus then was placed on the first ten ranked answers, these were analysed against the key answer, if an answer appears between ranks 6 and 10 that is better than in ranks 1 to 5, then that rank position scores 1. The second extension is to compare the first twenty ranks against the key answer, if an answer in ranks 11 to 20 is better than in the first 10 ranks, then that rank position scores 1. The graphs following shows the results achieved.

It can be seen that the rank position answers of 6 to 10 that supersede those in ranks 1 to 5 is only ten answers. It can also be seen that four answers in ranks 11 to 20 supersede those in the ranks 1 to 10. This means that of the fifty queries tested here, ten answers match closely the key answer in ranks 6 to 10 and overall four best answers are achieved in ranks 11 to 20.



The actual figures show that the system performance in detecting answers in the top five ranks achieved a score of 66% for that task, for detecting answers in the top ten ranks the score of 82% was achieved, and finally for detection in the top twenty ranks the score was 86%.



4.3.3. Binary Test

The second part of the relevance test was to see if any of the ranked sentences for each query contained any information relating to the query. To test this part 110 queries were used, of these 96 were valid queries, the remaining 14 were discarded for this purpose. Of these 96 valid queries, the average score for the queries was 16.15625, that is on average each query had 16.15625 rank position answers that provided some information. Every single valid query had at least one ranked answer that produced some information for the query, the co-lowest score for a query was 6 answers that provided information and the co-highest all 20 ranked answers provided some information. The table below shows how each ranked position fared.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
75	81	85	83	81	83	82	77	76	79	74	81	81	74	72	75	72	70	69	69

The top row represents the rank position of an answer, the number below that rank position indicates the number of queries that provided some information at that rank position (where maximum = 96, i.e. the number of valid queries for this task). The average number of queries that had some information for a rank position is therefore 76.95.

Chapter 5. Discussion

5.1. Architecture

5.1.1. User Interface

The user interface programmed by the author proved quite easy to use in its native form. Development was done on a PC running Windows 98(OSR 2), using the J Builder environment for Java. The advantages of this approach to user interface design is that components for common desktop metaphors are already available for use and are familiar to users. However, because of the initial desire for portability across different platforms the interface was tested against another platform(other than the development one). All of the functions of the user interface remained intact but the presentation of the user interface suffered aesthetically.

5.1.2. Information Source

The decision to use a search engine for the retrieval task made implementation of the system easier, but the use of this search engine meant that the author had no control over the documents that were presented to the DPD system. It would have been ideal to still use a search engine but have some sort of filtering option available to control the flow to the DPD system. The user interface extracts up to 600 URLs for the DPD system, this was restricted because the web search engine used(Google(www2)) presents this figure of URLs normally or at most. Other web search engines such as (Altavista(www4)) were considered but because of the limited amount of documents referenced, in this case 200, this option was discarded.

Another point of interest as regards to the information source is that as no control was exerted on the documents presented to the DPD system, it couldn't be established in the answers for queries where the sentences extracted came from. If this could be incorporated further evaluations could have taken place. For example in chapter 2, it was noted that in Automatic Text Summarisation, it is common to compare summaries with the original source document(s). If it could be seen where

the source of sentences extracted came from, then a comparison between the sentence and the source document could be made.

5.2. Evaluation

5.2.1. Query Sets

The query sets used can be found in the appendices. Interesting points to note about query sets, is the number used in each evaluation. In the effectiveness test, 227 valid queries were used to evaluate the pattern coverage and distribution metrics. The same queries were used to evaluate the efficiency of the system. In all, 490127 sentences were extracted by the DPD system, this meant that in order to fully evaluate the system in Relevance terms would have required enormous effort. So it was decided to use a subset of the 227 to evaluate the MRAR results, the extended task and the binary test. The numbers used in each case, 50, 50, 96 were used because of time restrictions. The ideal case would have been to evaluate all the valid queries in order to produce evaluations. However, it is fair to say from the results achieved that the query sets for each task seemed large enough.

5.2.2. Judge

The appointment of the author as a judge meant that the resolution of key answers for the MRAR tests was always achieved. It also meant that the judge's decision on key answers was final in the MRAR tasks. This means that if the judge not had any idea what a certain query's description may be, then a key answer may not be formed, however this situation did not occur for the 50 queries in the MRAR tasks. The judge reported little difficulty in reading the answers, this suggests that the majority of answers given by the DPD system were comprehensible.

5.2.3. Assessors

The decision to appoint assessors in order to resolve the key answers for the MRAR tasks was taken quite late in the evaluation cycle. Initially, the judge felt that one answer would be enough to deem returned answer accuracy. Due to the fact that perceptions of answers differ amongst people, it was realised that multiple answers were needed. To this end, two answers for each query were combined to form one answer as described previously. The assessors appointed were given clear instructions to form a sentence description for the query terms, but no prior training

was given and certainly no examples. The key answers combined by the judge reflect the assessors judgements, it is noted that although a sentence was required for each query item, assessors typically formed a short sentence consisting of less than 10 words. The attitudes and backgrounds of the assessors were not taken into account, but it is noted that this may have an effect on the key answers determined.

5.2.4. Effectiveness

The sole concern of the effectiveness task was the performance of the pattern matching algorithm. Correspondence with (Saracevic(1995)) can be seen here because this could called evaluation at the processing level. The results achieved indicate that the matching of 'Is A' and 'Acronym' and 'Appositive' scored most, and a low number of matches for 'Especially' and 'Such X As' exist. The high scores seem to point to the fact that the high number of web documents retrieved showed the patterns can be of good use to detect candidate answers. Whereas conversely, the low scores may indicate that on the WWW, documents which ascribe to these patterns are relatively uncommon.

Various reasons could account for the low scores, the one postulated here is that WWW documents are not always structured as normal print versions, and indeed much of the material on the WWW is written by individuals with little regard for basic text constructs. As the documents were not filtered in any way prior to processing by the DPD system, these documents could not be ruled out. The other theory is that writers/authors of documents just do not use the low score patterns as much as the higher score ones.

5.2.5. Efficiency

The efficiency metric was included in order to ascertain the engineering level of the project. Any system built today must be measured by it's performance and, measuring how long the system takes to execute a query is a valid way of assessing performance. Two comparisons were made: 1) the number of sentences matched and the time it takes to execute a query and 2) the time it takes to execute a query and the number of URLs presented to the DPD system.

From the results achieved for the first task, the graph shows almost a linear relationship the number of sentences matched and the time taken to process a query. This was largely expected because as sentence collection size increases then the ranking of sentences will take longer.

For the second task the same query process times were plotted against the number of URLs retrieved by the user interface. This graph is meant to show that there is no correspondence between the number of URLs and the process time. As the graph testifies, this is certainly the case.

Another point to note from this exercise was the long times associated with some queries. For example, the two longest process times for “NATO” and “ISO” took 1930 and 2863 seconds respectively. This amounted to almost 80 minutes of processing time. Clearly a user friendly QA system should return faster times than those.

5.2.6. MRAR

Whilst the system described within is not a full QA system, it was felt that the MRAR metric of the QA task of TREC seemed relevant. This is because the DPD system returns a ranked list of answers that were compared to a key answer metric. The difference lies in that those systems submitted to the QA track of TREC are required to answer closed classed questions and return five ranked answer strings of lengths of restricted length. The DPD system designed here is really dealing with the idea of an open classed question, with the answer not restricted by length but by sentence. Also the system returns a ranked list of twenty answer sentences rather than five strings.

The MRAR metric was still used because it seems to be the sole common measurement used for QA systems. In order to evaluate the score for our system, the ranked list was used to compare the first five ranks only. The score of 0.358666 seems to be quite good if compared to the results of the TREC track (Voorhees & Tice(1999)). Even though it is not the same task, the procedure of evaluation is similar. So the score seems respectable and if inserted in the track to compare with

other systems, it would lie in the top 15, the Textract system described in chapter 2 was a superior system heading the table with a score of 0.617.

5.2.7. Extended MRAR

As the DPD system returns 20 ranked answer sentences to the user, it was felt that there was a need to establish whether the answer ranks beyond the first five may of relevance. The results obtained seem to suggest that the DPD system is capable of detecting the majority of key answers within the top ten. Of the fifty queries tested it was already known that 33 queries could be satisfied by considering the first five ranks only, but the extended task revealed that a better answer, that is one that satisfies the query whilst simultaneously not giving any irrelevant information could be found in ranks 6 to 10, for ten of the queries. Furthermore, it is shown that in ranks 11 to 20 four queries could be satisfied better than in ranks 1 to 10. This does not mean that only 33 queries were satisfied in total because the some of the answers that has been said to supersede ones before might not have had an answer in the ranks before.

The figures show that the system performed adequately in detecting answers in the top five ranks achieving a score of 66% for that task, for detecting answers in the top ten ranks the score of 82% was achieved and finally for detection in the top twenty ranks the score was 86%. They go on to suggest that as the number of ranked answers increase then the possibility of finding a key answer becomes correspondingly bigger.

5.2.8. Binary Test

This test revealed some good performance figures for this system. It was exceptional to observe that from the answer ranks given to the user, all 96 queries considered had at least 6 rank positions that gave some information to the user. This leads to the conclusion that the system is functioning on a level that was beyond expectation. Of course, the criteria as to what was information, in relation to the query is important because, if the assessment was to say that binary relevance was achieved if there was text in an answer, then every rank will score 1. It was not such a simple criteria.

5.3. The System As A Whole

In common with the thinking of the QA track of TREC, a thorough of discussion of the assessment is detailed as follows.

The focus on evaluation of the system seems to have been achieved, the notable exception that is recognised, is the lack of correlation between answers given by the DPD system and the source of the answers. It would have been a good idea to evaluate the source and against the extracted sentence(s). The actual results and evaluation achieved was done in a consistent fashion and a good sized sample query base was used.

As with any system there were problems in the output. These were discovered when applying the MRAR task, the extended task or the binary test.

A general problem encountered was that some queries returned ranked answers that were not in the English language. It was realised after evaluation had finished, that the problem lies in the implementation of URL retrieval by the user interface. The problem was that by using a search engine to retrieve URLs, the user interface had not encoded the information that filters the URLs to contain English documents only. This problem can be resolved easily.

For an example query where the MRAR tasks produced an interesting result, consider the query 'Mike Tyson'. The key answer established by the judge was 'The popular American Heavyweight boxer often associated with his notorious behaviour inside and outside the ring'. One of the answers given by the system at rank position 4 was: *'Mike Tyson Photo Gallery Mike Tyson is the greatest boxer of all time!'*. Quite clearly, this does not correspond with the key answer, but will score in the binary test because it gives some information about 'Mike Tyson', namely that he is a boxer and there is a photo gallery of Mike Tyson's.

Another example, where the system falls down in the extended MRAR task is with the query 'NATO'. The key answer that was used as a comparison was 'North Atlantic Treaty Organisation', however the system extracted answers that were

describing Nato's various actions, the best answer though achieved was at rank position 12: *'NATO basic goals: NATO is a defensive alliance based on political and military cooperation among independent member countries, established in accordance with Article 51 of the UN Charter.'* Although this seems to be a good answer under normal circumstances, the metric used in the extended MRAR evaluation required that this answer be passed over. However this answer would still score for the binary test, as it imparts some information for the query.

A query whereby the system failed in the same way as the above query was with 'FIFA'. The key answer derived was 'The world football governing body'. However, all the answer sentences in the ranks given related to the computer game series of the same name. This seems to support the argument that different assessors will form different relevance judgements and therefore create different key answers. It is entirely possible that had the assessors had contact with the video game of the same name then it may have been given as a key answer. This argument gives credence to the binary test, in that as key answers vary with different assessors it may be prudent to measure the quality of answers by employing a simple binary judgement.

Another query whereby the output from the system needs addressing is for 'Venus Williams'. The key answer expected was 'A tennis sensation' or 'Ladies Wimbledon Champion 2000'. None of the ranked answers produced the second key answer. This query demonstrated the problem of old links on the WWW. In some answer ranks, information on Venus's age was given that was wrong and out of date. This supports Mizzaro's (Mizzaro(1998)) notion that relevance is also attributed to time.

A final example given here is the query 'TREC'. The key answer obtained was 'Text Retrieval Conference'. None of the ranked sentences produced these four words, however in a few of the ranked answers the topic of IR was discussed. At rank position 5: *'In the six years since the beginning of TREC, the state of the art in retrieval effectiveness has approximately doubled, and technology transfer among research labs and between research systems and commercial products has accelerated.'*

A small problem that exists in the present architecture of the system is the duplication of answers across different ranks. The root cause of this is not known, however a probable cause identified is the duplication of URLs passed to the DPD system. This is in turn caused by the duplication of resources pointed to by the search engine or by multiple URLs for single sites. A possible counter to the former problem is simply to remove duplicate URLs, prior to processing by the DPD system, this will not affect single sites that have multiple URLs.

It has been established that relevance is a broad issue, and two different, contrasting attempts have been made of assessing it in this system. It has been postulated that the QA track of TREC serves as a good basis in evaluating QA systems, or partial QA systems in this case. The further development of the MRAR metric to encompass the top ten and top twenty answers seems to be creditable. The results achieved for this system for the top twenty answers(86% of queries answered by key answer match) certainly bears this out.

In all, the augmentation of the DPD system to incorporate the WWW seems to be a success with minor problems present. Solutions to a majority of these problems have been proposed, and it is foreseen that further tuning of the system will make it even more successful.

Chapter 6. Conclusion

6.1. Overview

It has been investigated whether a system that produces descriptions of noun phrase from documents on the WWW is feasible. A background to the areas concerned, IR, IE, QA and Automatic Text Summarisation has been discussed. Some of the techniques of evaluation has been used to test the prototype system described within. A methodology has been presented that captures the important metrics that are applicable to assessing the performance of the system. The results of the experiment that took place are shown along with a thorough assessment.

Chapter 2 gives the introduction to the areas applicable to the development of this system. IR is relevant to the project as retrieval takes place at the user interface stage. Also the relevance metric of IR is discussed comprehensively and the results show that the relevance is important and is different to different people, at different times. The issues relating common ranking techniques is shown to have relevance to the DPD system. The coverage of IE, QA and Summarisation show that the different disciplines are beginning to merge. Indeed, the project described within uses the techniques of IE, Summarisation and QA systems in some way. The pattern matching approach could be considered an IE technique and Summarisation technique. The TREC QA approach to evaluation has been used and augmented.

Chapter 3 shows the methodology of the experiment and the architecture of the system. It has been shown that by combining the existing DPD system with the user interface, a partial QA system can be implemented. The ranking method of the DPD system has been discussed as well as the patterns used to match within documents.

Chapters 4 and 5 revealed the results of the experiment for discussion. The results shown suggests that pattern matching is a good way of detecting descriptions for queries. This is concluded because of the high scores present in the relevance tests considered. However, it was also shown that the system performed badly at the

efficiency evaluation. This suggests that further work in developing the algorithms may be necessary.

To conclude the following points are of relevance to the system described:

The system has the potential to provide descriptive information of a noun phrase. Techniques from IE, IR, Automatic Text Summarisation can be merged to form a QA system.

Simple pattern matching can detect descriptive phrases to a high rate.

The number of sentences matched affect the efficiency of the system

Simple IDF-based term weighting can be useful for ranking of answers.

The quality of a descriptive phrase may be influenced by the information source, and by some attributes of queries and users.

Evaluation is a key criteria to consider when developing QA or partial QA systems.

Extending the QA evaluation used by TREC seems to be beneficial in extracting good answers.

Relevance of answers can be determined by users and may differ through time and personal perceptions.

6.2. Further Work

The ideal system would accept a broader range of questions and incorporate some NLP features. For example at the interface stage, queries submitted via natural language could be processed by a parser to validate the question. From IE, features such as domain knowledge could be added to augment the pattern matching strategy. This will enable spurious answers to be discarded. A more basic suggestion is to increase the amount of patterns that the system is able to match. For example, the pattern: “DP ... also known as X”, where DP is the descriptive phrase and X the query, could be added as a possible candidate.

Despite these ideas, a more fundamental problem needs to be addressed first, it has been seen that performance of the system in efficiency terms can be described only as adequate at present. The lengthy times associated with processing may be

prove to be this system's main stumbling block. The high number of sentences detected for each query means that, trimming of the number of sentences needs to take place before ranking in order to alleviate the problem of lengthy process times. It also means that may be, adding more patterns to be matched is not such a worthwhile extension.

Finally, an important addition to the project as discussed earlier is to filter the URLs from the search engine somehow to enable multiple duplicates of answers to be eliminated. This filtering function could also have the effect of controlling the flow of documents to the DPD system.

A grounding into the complexities of developing a good QA system to deal with open classed questions has been addressed. Improvements to the existing system need to be made to achieve a fully functioning system.

References

Baeza-Yates, R. & Riberio-Neto, B.(1999). *Modern Information Retrieval*. New York: ACM, Addison Wesley Longman.

Breck, E., Burger, J., House, D., Light, M. & Mani, I.(1999). “Question Answering from Large Document Collections”. [Online][Last Visited 04 August 2000]

Available: http://www.mitre.org/support/papers/tech-papers99_00/breck_question/index.shtml

Brown, E.W. (1995). “Fast Evaluation of Structured Queries for Information Retrieval”. In Fox, E.A., Ingwersen, P. & Fidel, R.(eds.)(1995). *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp30-38. Seattle, Washington: ACM Press.

Chakravarthy, A.S. & Haase, K.B. (1995). “NetSerf: Using Semantic Knowledge to Find Internet Information Archives”. In Fox, E.A., Ingwersen, P. & Fidel, R.(eds.)(1995). *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp4-11. Seattle, Washington: ACM Press.

Chowdhury, G.G. (1999). *Introduction to Modern Information Retrieval*. London: Library Association.

Cosijn, E. & Ingwersen, P.(2000). “Dimensions of relevance”. In *Information Processing and Management*, 36. pp533-550. Pergamon.

Firmin, T. & Chrzanowski, M.J.(1999). “An Evaluation of Automatic Text Summarization Systems”. In Mani, I. & Maybury, M.T.(1999)(eds.). *Advances In Automatic Text Summarization*. pp 325-336. Camb, Mass: MIT.

Fox, E.A. & France, R.K. (1987). “Architecture of an expert system for composite document analysis, representation, and retrieval”. In *Journal of Approximate Reasoning*, 1, 151-175. Also In Sparck Jones, K. & Willet, P. (eds.) (1997).

- Readings In Information Retrieval*. pp 400-413. San Francisco: Morgan Kaufmann.
- Gaizauskas, R., Humphreys, K., Azzam, S. & Wilks, Y.(1997). “Concepticons vs. Lexicons: An Architecture for Multilingual Extraction”. In Pazienza, M.T.(ed.)(1997). *Lecture Notes in Artificial Intelligence, Subseries of Lecture Notes in Computer Science. Information Extraction, A Multidisciplinary Approach to an Emerging Information Technology*. pp 28-43. Berlin: Springer-Verlag.
- Harman, D.(1995). “The TREC Conferences”. In Sparck Jones, K. & Willet, P. (eds.) (1997). *Readings In Information Retrieval*. pp 247-256. San Francisco: Morgan Kaufmann.
- Järvelin, K. & Kekäläinen, J.(2000). “IR evaluation methods for retrieving highly relevant documents”. In Belkin, N.J., Ingwersen, P. & Leong, M.(eds.)(2000). *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp 41-48. Athens, Greece: ACM Press.
- Jokinen, K.(1993). “Reasoning about Coherent and Cooperative System Responses”. In Adorni, G. & Zock, M. (eds.)(1993). *Lecture Notes in Artificial Intelligence, Subseries of Lecture Notes in Computer Science. Trends in Natural Language Generation*. pp 168-187. Berlin: Springer-Verlag.
- Kupiec, J.M. (1993). “MURAX: A robust linguistic approach for question-answering using an on-line encyclopedia”. In *Proceedings of the Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp 181-190. Pittsburgh: ACM Press.
- Kupiec, J.M.(1999). “MURAX: Finding and Organizing Answers from Text Search”. In Strzalkowski, T.(ed.)(1999). *Natural Language Information Retrieval*. pp 311-332. Dordrecht, Netherlands: Kluwer Academic.

Knaus, D. & Schäuble, P. (1994). "Effective and Efficient Retrieval from Large and Dynamic Document Collections". In Harman, D.K.(ed.)(1994). *The Second Text REtrieval Conference(TREC-2)*. Washington: US Government Printing Office.

Lehnert, W.(2000). *Information Extraction*. [Online][Last Visited 04 August 2000]
Available: <http://www-nlp.cs.umass.edu/nlpie.html#IEpubs>

MacCall, S.L. (1998). "Relevance Reliability in Cyberspace: Toward Measurement Theory for Internet Information Retrieval". In *ASIS '98 Information Access in the global Information Economy, Proceedings of the 61st Annual Meeting of the American Society for Information Science*. pp 13-22. Pittsburgh, PA: Information Today Inc.

McKeown, K. & Radev, D.R. (1995). *Generating Summaries of Multiple News Articles*. [Online][Last Visited 06 June 2000]
Available: <http://www.cs.columbia.edu/~radev/publication/publications.html>

Miller, G.A.(1990). "WordNet: An On-line Lexical Database".
Available: <http://www.cogsci.princeton.edu/~wn/>
Also In *International Journal of Lexicography*. 3(4).

Mizzaro, S.(1998). "How Many Relevances in IR?". In *Interacting with Computers*. 10. pp 395-322.
[Online][Last Visited 08 August 2000]
Available:
http://www.dcs.gla.ac.uk/~johnson/papers/hci_and_ir/first_workshop.html

O'Connor, J. (1980). "Answer-Passage Retrieval by Text Searching". In *Journal of the American Society for Information Science*. pp227-239.

Radev, D.R. & Mckeown, K.(1997). "Building a Generation Knowledge Source using Internet-Accessible Newswire". [Online][Last Visited 06 June 2000]
Available: <http://www.cs.columbia.edu/~radev/publication/publications.html>

Also In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. pp221-228.

Rath, G.J., Resnick, A. & Savage, T.R.(1961). “The Formation of Abstracts By the Selection of Sentences”. In Mani, I. & Maybury, M.T.(1999)(eds.). *Advances In Automatic Text Summarization*. pp 287-291. Camb, Mass: MIT.

Riloff, E. & Lorenzen, J.(1999). “Extraction-Based Text Categorization Generating Domain-Specific Role Relationships Automatically”. In Strzalkowski, T.(ed.)(1999). *Natural Language Information Retrieval*. pp 167-198. Dordrecht, Netherlands: Kluwer Academic.

Salton. G., Singhal, A., Mitra, M. & Buckley, C.(1997). “Automatic Text Structuring And Summarization”. In Mani, I. & Maybury, M.T.(1999)(eds.). *Advances In Automatic Text Summarization*. pp 341-355. Camb, Mass: MIT.
Also In *Information Processing and Management*, 33(2), pp193-207.

Saracevic, T.(1975). “Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science”. In Sparck Jones, K. & Willet, P. (eds.) (1997). *Readings In Information Retrieval*. pp 143-165. San Francisco: Morgan Kaufmann.

Saracevic, T. (1995). "Evaluation of Evaluation in Information Retrieval". In Fox, E.A., Ingwersen, P. & Fidel, R.(eds.)(1995). *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp138-146. Seattle, Washington: ACM Press.

Soderland, S.(1997). “Learning to Extract Text-based Information from the World Wide Web”. [Online][Last Visited 04 August 2000]
Available: <http://www-nlp.cs.umass.edu/pubs/Soderland-KDD97.pdf>
Also In *Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*.

Sparck Jones, K. (1999). "Automatic summarizing: factors and directions". In Mani, I. & Maybury, M.T. (eds.) (1999). *Advances In Automatic Text Summarization*. pp1-12. Camb, Mass: MIT.

Sparck Jones, K. & Willet, P (eds.) (1997). *Readings In Information Retrieval*. . San Francisco: Morgan Kaufmann.

Srihari, R. & Li, W.(1999). "Information Extraction Supported Question Answering". [Online][Last Visited 08 August 2000]
Available: <http://trec.nist.gov/pubs.html>

Strzalkowski, T.(ed.)(1999). *Natural Language Information Retrieval*. Dordrecht, Netherlands: Kluwer Academic.

Sundheim, B.M.(1993). "5th Message Understanding Conference—Call for Participation". [Online][Last Visited 05 July 2000]
Available: http://www.dcs.gla.ac.uk/idom/irlist/new/1993/93-x-1-145/5th_Message_Understanding_Conference_Call_for_Participation.html

Voorhees, E.M. & Tice, D.M.(1999). "The TREC-8 Question Answering Track Evaluation". [Online][Last Visited 10 August 2000]
Available: <http://trec.nist.gov/pubs.html>

Voorhees, E.M. & Tice, D.M.(2000). "Building a Question Answering Test Collection". In Belkin, N.J., Ingwersen, P. & Leong, M.(eds.)(2000). *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp 200-207. Athens, Greece: ACM Press.

WWW0. *Acronym Finder*. [Online][Last Visited 31 July 2000]
Available: <http://www.mtnds.com/af/>

WWW1. *TREC Home Page*. [Online][Last Visited 08 August 2000]
Available: <http://trec.nist.gov/>

WWW2. *Google Home Page*. [Online][Last Visited 25 August 2000]

Available: <http://www.google.com/>

WWW3. *Altavista Home Page*. [Online][Last Visited 08 August 2000]

Available: <http://www.altavista.com/>

Yeates, S.(1999). “Automatic Extraction of Acronyms from Text”.

[Online][Last Visited 31 July 2000]

Available: <http://www.cs.waikato.ac.nz/~nzdl/publications/>

Appendix A: Queries For Binary Test

Query Number	Query Phrase
1	Patrick Kluivert
2	Sun Microsystems
3	Area 51
4	Bill Gates
5	Tony Blair
6	Imperial College
7	Tiger Woods
8	Millenium Dome
9	Star Trek
11	FIFA
12	Albert Einstein
13	Iron Curtain
14	Bamboo Curtain
15	JCB
16	Hewlett Packard
17	Sushi
19	Microsoft
20	Mike Tyson
21	Apple
22	Steve Jobs
23	Lennox Lewis
24	Damon Hill
25	Chianti
26	TREC
27	Cisco
28	Xerox
29	Flymo
30	Venus Williams
32	Charles Darwin
33	Edward Elgar

34	Robert Louis Stevenson
35	Mozart
37	Fox Mulder
38	Janet Jackson
39	Tesco
40	Zinedine Zindane
41	Vladimir Putin
42	Vindaloo
44	Margaret Thatcher
45	Neil Kinnock
46	Yorkshire Pudding
47	Claret Jug
48	Jules Rimet
49	NIST
50	BSI
51	United Nations
52	Enid Blyton
53	NATO
54	Chris Tarrant
55	Beatles
56	Rupert Murdoch
57	Aaron Spelling
58	Ronan Keating
59	DARPA
60	CERN
61	Calculus
62	Julia Roberts
63	Winnie the Pooh
64	Saki
65	Teletubbies
66	Orson Welles
68	McDonalds
69	Ravioli

70	Netscape
71	Andre Agassi
72	Hippocrates
73	Zeus
74	SDI
75	Star Wars
76	Nescafe
77	EDS
78	GMC
79	Chow Mein
80	Baskin Robbins
81	CSC
82	KPMG
83	Jeffery Archer
84	Bill Clinton
85	George Bush
86	UB40
89	ICL
90	Clarkes
91	Harry Potter
92	Pokemon
95	Wasabi
96	Richard Branson
97	Virgin
98	Borland
100	Morrisons
102	Vauxhall
104	Viglen
105	Trevor McDonald
106	Jerry Springer
108	Chambers
109	Nissan
110	Mercedes

Appendix B: Queries For MRAR Tasks

Query Number	Query Phrase
1	Patrick Kluivert
2	Sun Microsystems
3	Area 51
4	Bill Gates
5	Tony Blair
6	Imperial College
7	Tiger Woods
8	Millenium Dome
9	Star Trek
11	FIFA
12	Albert Einstein
13	Iron Curtain
14	Bamboo Curtain
15	JCB
16	Hewlett Packard
17	Sushi
19	Microsoft
20	Mike Tyson
21	Apple
22	Steve Jobs
23	Lennox Lewis
24	Damon Hill
25	Chianti
26	TREC
27	Cisco
28	Xerox
29	Flymo
30	Venus Williams

31	Jasmin Tea
32	Charles Darwin
33	Edward Elgar
34	Robert Louis Stevenson
35	Mozart
37	Fox Mulder
38	Janet Jackson
39	Tesco
40	Zinedine Zindane
41	Vladimir Putin
42	Vindaloo
44	Margaret Thatcher
45	Neil Kinnock
46	Yorkshire Pudding
47	Claret Jug
48	Jules Rimet
49	NIST
50	BSI
51	United Nations
52	Enid Blyton
53	NATO
54	Chris Tarrant

Appendix C: Key Answers

1. Patrick Kluivert – Dutch Football International, a star of Euro 2000.
2. Sun Microsystems – A multinational computer equipment, operating systems supplier.
3. Area 51 – The top secret US Air Force Base in Nevada.
4. Bill Gates – The multibillionaire boss of Microsoft Inc.
5. Tony Blair – The Prime Minister of Great Britain.
6. Imperial College – A college of the University of London.
7. Tiger Woods – A millionaire, Nike endorsed golf star.
8. Millenium Dome – The multimillion pound attraction situated in Greenwich, London.
9. Star Trek – A popular science fiction television series, also made into a series of films.
10. FIFA – The world football governing body.
11. Albert Einstein – The renowned scientist often associated with the theory of relativity.
12. Iron Curtain – The communist regime in the Eastern Block nations especially the USSR.
13. Bamboo Curtain – The regime that existed in the Far East nations especially China.
14. JCB – The supplier of the JCB credit card; A recognised world leader in excavator manufacture.
15. Hewlett Packard – The supplier of IT hardware, particularly printers and computers.
16. Sushi – A Japanese dish made from raw fish and rice.
17. Microsoft – The multi-billion pound US software and operating systems giant.
18. Mike Tyson – The popular American Heavyweight boxer often associated with his notorious behaviour inside and outside the ring.
19. Apple – A fruit; A computer hardware manufacturer.
20. Steve Jobs – Silicon Valley based entrepreneur, founder of Apple Computer Inc.
21. Lennox Lewis – A Heavyweight Champion.
22. Damon Hill – A Formula One star of the 1990s, former F1 champion.
23. Chianti – A region of Italy famous for it's fine wines.

24. TREC – Text REtrieval Conference.
25. Cisco – Cisco Systems; The leader in internetworking solutions.
26. Xerox – A famous photocopier/document/printer company.
27. Flymo – A leader in lawnmower production.
28. Venus Williams – A tennis sensation; Ladies Wimbledon Champion 2000.
29. Jasmin Tea – A Chinese tea famous for it's aroma.
30. Charles Darwin – The writer of “The Origin of Species”; First who postulated the theory of Natural Selection.
31. Edward Elgar – A famous English composer.
32. Robert Louis Stevenson – A writer of many books including “Treasure Island”.
33. Mozart – A classical music genius.
34. Fox Mulder – A TV character played by David Duchovony in the TV show “The X-Files”.
35. Janet Jackson – The youngest of the famous musically gifted Jackson Family.
36. Tesco – Britain's leading supermarket.
37. Zinedine Zidane – French soccer star, a star of France 98 and Euro 2000.
38. Vladimir Putin – The Premier of Russia.
39. Vindaloo – A fiery hot Indian curry dish.
40. Margaret Thatcher – A member of the House of Lords; Former British Prime Minister.
41. Neil Kinnock – Labour Party member; A member of the European Parliament.
42. Yorkshire Pudding – A dish usually produced to accompany roast beef.
43. Claret Jug – The trophy presented to the winner of golf's Open Championship.
44. Jules Rimet – The father of the World Cup football tournament.
45. NIST – National Institute of Science and Technology.
46. BSI – British Standards Institute.
47. United Nations – The international humanitarian agency.
48. Enid Blyton – Famous children's book author; Author of books relating to the “Famous Five”.
49. NATO – North Atlantic Treaty Organisation.
50. Chris Tarrant – Radio host; Television Presenter; Celebrity.

Appendix D: Sample Answers

These answers are for the MRAR task, they represent the best answer that matches the key answer for that query.

1. Patrick Kluivert (Rank Position 5) – “(Patrick Kluivert, insofar as he plays with the Dutch national team and FC Barcelona, is also sponsored by Nike)”.
2. Sun Microsystems (Rank Position 5) – “Plaintiff Sun Microsystems is the developer and licensor of the JAVA Technology, which comprises a standardized application programming environment that affords software developers the opportunity to create and distribute a single version of programming code that is capable of operating on many different, otherwise incompatible systems platforms and browsers.”.
3. Area 51 (Rank Position 14) – “The first detailed satellite images of Area 51, the top-secret Air Force test site in Nevada, apparently reveal nothing out of the ordinary.”.
4. Bill Gates (Rank Position 14) – “March 5, 1998 Following intense grilling from Senate Judiciary Committee Chairman Orrin Hatch, Bill Gates, the Microsoft Corp. chairperson, conceded that the company does restricts its Internet partners' ability to deal with its rivals.”.
5. Tony Blair (Rank Position 10) – “Tony Blair Tony Blair is the current Prime Minister of the United Kingdom.”.
6. Imperial College (Rank Position 3) – “In case you were wondering, Imperial College is a large part of the University of London.”.
7. Tiger Woods (Rank Position 8) – “ELPERS Tiger Woods Bunker Mentality Woods Supports Actors' Strike; Won't Film Nike Ad Tiger Woods, the No. 1 player in golf and winner of 21 tournaments around the world, was no different than more than 100,000 actors.”.

8. Millenium Dome (Rank Position 2) – “While ordinary Britons thronged the streets, Queen Elizabeth, the Prime Minister, Mr. Tony Blair and around 10,000 other celebrities and guests saw in the New Year in the Millenium Dome, a large tent-shaped structure that has become the newest landmark on the London skyline.”.
9. Star Trek – no ranked answer.
10. FIFA – no ranked answer.
11. Albert Einstein (Rank Position 7) – “Albert Einstein was a physicist and Nobel Laureate, know primarily for the Theory of Relativity”.
12. Iron Curtain (Rank Position 3) – “Since the fall of the Iron Curtain, the countries of the Former Soviet Union and the Soviet Bloc have been emerging from decades of an informational black-out of positive images and constructive cultural input from the West.”.
13. Bamboo Curtain (Rank Position 11) – “In recent years the defeat of communist insurgences in Thailand and Burma, coupled with the lowering of the Bamboo Curtain in China and Laos as both those countries slowly switch to free trade, has opened some parts of the Golden Triangle to the outside world for the first time in decades.”.
14. JCB (Rank Position 6) - “Amount: Y5 billion Issue Date: May 29, 2000 Due Date: May 27, 2005 Coupon: 1.45% Covenants: Negative Pledge Commissioned Company: Yes ¥ JCB is a bank-affiliated credit card company, ranking top in the card transactions in value.”.
15. Hewlett Packard (Rank Position 3) – “Hewlett Packard is the recognized leader in Network Management solutions.”.
16. Sushi (Rank Position 9) – “Sushi is a type of Japanese delicacy that consists mainly of fish and rice.”.

17. Microsoft (Rank Position 7) – “ On April 28, the Justice Department and 17 state attorneys general asked Jackson to break the company into two parts - one that would develop and market the Windows operating system and one that would develop Microsoft's other software and Internet products, such as the Microsoft Office suite of programs.”.
18. Mike Tyson (Rank Position 2) – “For those of you who don't know, Mike Tyson, a former heavyweight champion, lost his title two years ago and in the very next fight became so furious that he bit a part of opponent Evander Holyfield's ear off.”.
19. Apple (Rank Position 1) – “Followup-To: comp.emulators.apple2 Date: 10 Mar 1998 08:34:43 GMT Organization: Leng in the Cold Waste Message-ID: Reply-To: Alex Maddison Summary: This FAQ contains a guide to cross-platform emulators of Apple 8-bit and 16-bit computers - primarily the Apple][series - and other emulator resources including disk-images (software).”.
20. Steve Jobs (Rank Position 8) – “But while Steve Jobs, the 42-year old maverick and founder of Apple Computer, is considered an icon in the Silicon Valley as a techno-futurist, Gandhiji shunned technology (the spinning wheel et al).”.
21. Lennox Lewis (Rank Position 3) – “WBC heavyweight champion Lennox Lewis, who is currently training in the Poconos for his rematch with his IBF/WBA heavyweight counterpart Evander Holyfield, is so confident of his chances at becoming the next undisputed heavyweight champion that he is already looking ahead to fighting Mike Tyson.”.
22. Damon Hill (Rank Position 1) – “Damon Hill hits out at Michael Schumacher Damon Hill, the most recent British winner of the Formula 1 World Championship in 1996, talked for the first time in length about life after retiring from Formula 1 last night on BBC TV's On Side sporting talk show.”.
23. Chianti – no ranked answer.

24. TREC – no ranked answer.
25. Cisco (Rank Position 16) – “Cisco is the clear market leader in traditional networking markets and it is extremely well positioned in the emerging service provider equipment and integrated voice and data equipment markets”.
26. Xerox (Rank Position 5) –“ Xerox, based in Stamford, Conn., is a leader in the document business, integrating fax machines, printers, scanners, and copiers, as well as PC and workstation software.”.
27. Flymo (Rank Position 1) –“ successful strategic product development Integration is the key to success Flymo, based in Newton Aycliffe, County Durham, are Europe's largest lawn mower manufacturer.”.
28. Venus Williams (Rank Position 4) –“ Venus Williams Signs On as Member of USA Tennis Spokesteam By Brian Walker Venus Williams, a rising star on the COREL WTA TOUR, has signed on as a member of the USA Tennis Spokesteam.”.
29. Jasmin Tea – no ranked answer.
30. Charles Darwin (Rank Position 2) – “Charles Darwin And the Evolution Revolution REBECCA STEFOFF Charles Darwin, from Oxford's ongoing Oxford Scientists series, is a biography of the naturalist and biologist that will

introduce young adults to Darwin's achievements and theories, including his theory of natural selection and origin of species.”.

31. Edward Elgar (Rank Position 2) – “Sir Edward Elgar Sir Edward Elgar Sir Edward Elgar, who rose from obscurity to become England's greatest composer for 200 years, was born on 2nd June 1857, at Broadheath near Worcester.”.

32. Robert Louis Stevenson (Rank Position 1) –“ Richard Dury what's here site has text site has video Information on author Robert Louis Stevenson. features Includes a biography, several images, etext of his essays and novels ('Treasure Island,' 'Kidnapped,' 'Strange Case of Dr. Jekyll and Mr. Hyde,' etc.), and other miscellaneous information. hot tips Start with the Life and Works Outline for a good introduction to Stevenson.”.

33. Mozart (Rank Position 3) –“ Wolfgang Amadeus Mozart, the uncommonly gifted and prolific composer, died 209 years ago at age 35 of a common killer-- from natural rather than unnatural ailments.”.

34. Fox Mulder (Rank Position 1) –“ DAVID DUCHOVNY Birth Date: August 7, 1960 Birthplace: New York City, USA Hair: Brown Eyes: Green Height: Six feet Education: English literature major at Yale University Background information: After making his debut in a beer commercial, he appeared in various features such as Twin Peaks, Beethoven (1992), Chaplin (1992), Kalifornia (1993) before landing the role of Special Agent Fox Mulder.”.

35. Janet Jackson (Rank Position 2) –“ Janet Jackson Vital Stats Born: May 16, 1966
Hometown: Gary, Indiana Occupation: Musical performer Relatives: Youngest
sister of superstar Michael Jackson and other members of the entertaining
Jackson family Love Interests: Divorced from James DeBarge; reportedly dating
Rene Elizondo Official Sites o Janet Jackson Official Site o Virgin Records:
Janet Jackson o E!”.
36. Tesco (Rank Position 4) –“ Tesco, the largest food retailer in the UK, has
ploughed £21 million into its Internet store so far, and is injecting a further £35
million into its online venture this year alone.”.
37. Zinedine Zindane – no ranked answer.
38. Vladimir Putin (Rank Position 2) –“ Starting with Yuri Andropov, proclaimed to
be a jazz-loving "closet liberal", followed by his protege Gorbachev, and later
Prime Minister Primakov, Stepashin, and now Vladimir Putin, the new President-
designate of the Russian Federation - each of them was greeted in his time as the
final solution to Russia's problems, as a brave reformer and a guarantor of
stability.”.
39. Vindaloo (Rank Position 1) –“ Vindaloo is a South Indian curry, always the
hottest thing on the menu, but simple enough in its ingredients”.
40. Margaret Thatcher (Rank Position 6) –“ News Release Former British Prime
Minister Margaret Thatcher to Speak at William Paterson University Lady

Margaret Thatcher, the former prime minister of Britain, will continue the 20th anniversary season of the Distinguished Lecturer Series at William Paterson University in Wayne with a lecture on December 10.”.

41. Neil Kinnock (Rank Position 1) –“ Near the top of the pyramid he has appointed two vice-presidents: Neil Kinnock, a former leader of Britain's Labour Party, who previously held the commission's transport job and will now be expected to reform the commission to make it more open and simple; and Loyola de Palacio, a Spanish conservative, who will handle the commission's relations with the European Parliament, as well as transport and energy.”.

42. Yorkshire Pudding (Rank Position 2) –“ Yorkshire Pudding Yorkshire Pudding is the traditonal accompaniment to Roast Beef (along with English Mustard or Horseradish Sauce).”.

43. Claret Jug (Rank Position 2) –“ Looking at the Claret Jug, the historic trophy given to the British Open champion, Lawrie declared, "It's just an incredible feeling and seeing this thing is totally amazing.”.”.

44. Jules Rimet (Rank Position 1) –“ The concept of a World Cup Soccer Championship was conceived by two Frenchmen, Jules Rimet, the president of FIFA, and Henri Delaunay, its General Secretary.”.

45. NIST – no ranked answer.

46. BSI (Rank Position 2) –“ Registration entry for : British Standards Institution -
BSI Description: BSI is a non-profit distributing independent organisation whose
core activity is standardization entry last updated on: 01 April 98 Address:.
. ”.
47. United Nations (Rank Position 18) –“ At the humanitarian needs assessment
meeting in Tbilisi in March 1996, the three governments, donors, United Nations
Agencies, non-governmental (NGOs) and other organisations agreed that the
strategy for the Caucasus humanitarian programme would address both
emergency relief and rehabilitation issues.”.
48. Enid Blyton (Rank Position 2) –“ ALISON KERR considers the reality behind an
innocent icon FEW authors have been as prolific or as widely read as Enid
Blyton, the creator of such children's favourites as The Famous Five, The Secret
Seven and, of course, Noddy, whose All New Adventures are brought to life in a
production at Edinburgh's King's Theatre this week.”.
49. NATO – no ranked answer.
50. Chris Tarrant (Rank Position 2) –“ Chris Tarrant, the popular TV and radio show
host, and Rose Hill, British Paralympic marathon racer, together launched the
Stannah Power Chair at The Howard Hotel in London.”.