FINAL DEGREE PROJECT

Urdu-English Cross-Language Question Answering System

MEDITERRANEAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

JUNE 2006

SUPERVISED BY: PAOLO ROSSO

BY:
GEET FARUKH
2204003
B.Sc. (Hons.)Computer System Engineering
Acknowledgement

I, gratefully acknowledge the cheerful and encouraging aid of Professor Paolo Rosso in the successful completion of this project. In addition, assistance from the UPV student Jose Luis, provided constant help during the implementation and setup of the Question Answering system. Nevertheless, support and provision of the necessary equipment from the university, MUST, eased the course of project development.

Finally, I would like to thank my parents for encouraging me and giving me moral support during the development and implementation stage of our Cross-Language Question Answering system.
The goal of Question Answering system is to provide inexperienced users with a flexible access to the information allowing them to query in natural language, in their native language and obtaining a concise answer. Cross-Language Web-based Question Answering systems allow the user for querying in a language different than the language in which the original web documents are written. However, these systems are mainly suited to English as the target language. In this project we will investigate with a case study, how much the translation of the queries, from the Urdu language to English, could reduce the accuracy of the Question Answering system. The main aim of this final degree project is to compare and analyze the accuracy of the translated query’s answer with those of the direct English query.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PART I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Information Retrieval systems</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Information versus Data Retrieval</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>The User Task</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>The Retrieval Process</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Web as a “Lexical Resource”</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Looking at the web</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Working of a Search Engine</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Meta Tags</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Building a Search</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>The Future of Search</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Natural Language Processing in IR</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Natural Language Processing</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Vector Search</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Question Answering System</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Generic System Architecture</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Question Answering methods</td>
<td></td>
</tr>
<tr>
<td>PART II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Cross-Language QA systems</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Web-Based QA system</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Technical aspects of QA systems</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>Query Reformulation</td>
<td></td>
</tr>
<tr>
<td>7.3</td>
<td>Answer extraction</td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>Experimental Results</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Urdu-English JAVA Translator</td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>Translation System architecture</td>
<td></td>
</tr>
<tr>
<td>8.2</td>
<td>How the Translator Works</td>
<td></td>
</tr>
<tr>
<td>8.3</td>
<td>Searching for keywords</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Conclusion and Future work</td>
<td></td>
</tr>
</tbody>
</table>

Screenshots
Appendix I
Appendix II
1. Introduction

The Web is becoming a universal repository of human knowledge and culture which has allowed tremendous sharing of ideas and information in a scale never seen before. This universe without frontiers has attracted vast attention from millions of people everywhere since the very beginning. Furthermore, it is causing a revolution in the way people use computers and perform their daily tasks. Nearly all kind of information (digital libraries, newspapers collections, cooking recipes etc.) is available in electronic format. These documents may satisfy almost every information need. For instance, home shopping and home banking are becoming very popular and have generated several hundred million dollars in revenues.

Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy one’s information need, the user might navigate the space of Web links (i.e., the hyperspace) searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient. For naive users, the problem becomes harder, which might entirely frustrate all their efforts. The main obstacle is the absence of a well defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality. These difficulties have attracted renewed interest in Information Retrieval (IR) and its techniques as promising solutions. As a result, almost overnight, IR has gained a place with other technologies at the center of the stage.

Automated information retrieval systems were originally used to manage information explosion in scientific literature in the last few decades [1]. Many universities and public libraries use IR systems to provide access to books, journals, and other documents. IR systems are often related to object and query. Queries are formal statements of information needs that are put to an IR system by the user. An object is an entity which keeps or stores information in a database. User queries are matched to documents stored in a database. A document is, therefore, a data object. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates [2].

Among the most visible and extensively used applications of information retrieval research are the Web search engines such as Google and Yahoo.
A search engine, it is a program designed to help find information stored on a computer system such as the World Wide Web, or a personal computer. The search engine allows one to ask for content meeting specific criteria (typically those containing a given word or phrase) and retrieving a list of references that match those criteria. Search engines use regularly updated indexes to operate quickly and efficiently [3].

In recent years, the combination of the Web growth and the explosive demand for better information access has motivated the interest in developing special search systems such as the Question Answering systems. Documents on the web are written in more than 1500 languages [4]. Therefore, it is useful to provide an inexperienced user with flexible access to the information allowing for writing a question in her mother tongue, and obtaining a concise answer.

Question Answering (QA) is a type of information retrieval technique. Given a collection of documents (such as the World Wide Web or a local collection) the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex Natural Language Processing (NLP) techniques than other types of information retrieval such as document retrieval, and it is sometimes regarded as the next step beyond search engines [5].

QA research attempts to deal with a wide range of question types including: fact, list, definition, how, why, hypothetical, semantically-constrained and cross-lingual questions. Search collections vary from small local document collections, to internal organization documents, to compiled newswire reports up to the World Wide Web [5].

QA systems are often suited to English as the target language. Cross-language Question Answering allows for querying the system in a language (e.g. Urdu) which is not the language of the documents (e.g. English). In this project, we approach the challenging Urdu-English Cross-Language Question Answering task in which the questions are placed in the Urdu language for the search of the documents are written in English. The goal is to investigate how much the translation of the queries from Urdu into English language, could reduce the accuracy of the system.

This report is organized in two parts to deal with the outlining generic idea behind the need of development of such a NPL search system and the later explains the technical details considered during the development procedures.

The first part of the report comprises of five sections. Section 2 briefly presents the reader with an introduction and the techniques used in Information Retrieval Systems. Section 3 outlines the terminologies, the architecture and current situation of Search Engines. Section 4 gives a glance at the Natural Language Query (NLQ) technique used in QA systems. Section 5 describes to the reader the idea behind Question Answering systems and the current developments in the field.
The second part of the report talks about the more technical aspects of the project. Section 6 discusses the architecture of our Web-based QA system and then gives a detailed description of the methods used for question reformulation and answer extraction. Section 7 list the technical details about the Urdu-English Translator developed to be adaptable with our QA system. Section 8 presents some comparison between the experimental results obtained with a simple QA system and a Cross-Language QA system. Finally in section 9 we summarize our experiments and the future works.

[1]: http://en.wikipedia.org/iki/information-retrieval
[3]: www.searchenginesshoedown.com
[5]: http://en.wikipedia.org/iki/Question_answering
2. Information Retrieval

Information Retrieval deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he/she is interested. Unfortunately, characterization of the “user information need” is not a simple problem. It can be defined as a natural language declaration of the informational need of a user [1]. For instance, “find documents which discuss the political implications of the Monica Lewinsky scandal in the results of the 1998 elections for the U.S. Congress”.

Consider, for instance, the following hypothetical user information need in the context of the World Wide Web:

Find all the pages (documents) containing information on college tennis teams which: (1) are maintained by an university in the USA and (2) participate in the NCAA tennis tournament. To be relevant, the page must include information on the national ranking of the team in the last three years and the email or phone number of the team coach.

Clearly, this full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a query (or a list of queries) which can be processed by the search engine (or IR system).

In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user [1]. The emphasis is on the retrieval of information as opposed to the retrieval of data.

2.1 Information versus Data Retrieval

Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. In fact, the user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query. A data retrieval language aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression. Thus, for a data retrieval system, a single erroneous object among a thousand retrieved objects means total failure. For an information retrieval system, however, the retrieved objects might be inaccurate and small errors are likely to go unnoticed. The main reason for this difference is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. On the other hand, a data retrieval system (such as a relational database) deals with data that has a well defined structure and semantics.
Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic. To be effective in its attempt to satisfy the user information need, the IR system must somehow `interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This `interpretation' of a document content involves extracting syntactic and semantic information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. Thus, the notion of relevance is at the center of information retrieval. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible [2].

The effective retrieval of relevant information is directly affected both by the user task and by the logical view of the documents adopted by the retrieval system.

2.2 The User Task

The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need [3].

Consider now a user who has an interest which is either poorly defined or which is inherently broad. For instance, the user might be interested in documents about car racing in general. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to car racing. For instance, he might find interesting documents about Formula 1 racing, about car manufacturers, or about the '24 Hours of Le Mans.' Furthermore, while reading about the '24 Hours of Le Mans', he might turn his attention to a document which provides directions to Le Mans and, from there, to documents which cover tourism in France. In this situation, we say that the user is browsing the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.

The figure 2.1 describes the way a user can browse documents for relevant material that he/she needs from the World Wide Web.
Logical View of the Documents

In the figure 2.2, a Logical view of the documents can be analyzed. Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords [3]. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a logical view of the document.

2.3 The Retrieval Process

To describe the retrieval process, we use simple and generic software architecture as shown in Figure 2.3 [4]. First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following:
(a) The documents to be used,
(b) The operations to be performed on the text, and
(c) The text model (i.e., the text structure and what elements can be retrieved).
The text operations transform the original documents and generate a logical view of them.

Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text [3]. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted file as indicated in Figure 2.3. The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times.

![Figure 2.3: The process of retrieving information.](image_url)

With the enormous increase in recent years in the number of text databases available on-line, and the consequent need for better techniques to access this information, there has been a strong resurgence of interest in the research done in the area of IR. For many years, IR research was done by a small community that had little impact on industry.

Today, however, the situation is considerably different. Retrieval techniques based on IR research have found their way into major information services (for example, West Publishing's WIN system, Individual's clipping service) and the World Wide Web (for example, Google and Yahoo) [4]. Many of the features once considered too esoteric for the typical user, such as "natural language" queries, ranked retrieval results, term weighting, "query-by-example", and query formulation assistance, have become common and, indeed, necessary in most IR products.

Given the speed with which industry has adopted the results of IR research, the IR community is faced with identifying major new directions. The emergence of new applications such as "digital libraries" is both an opportunity and a challenge [4]. These applications provide unique opportunities as test beds for evaluating and stimulating research, but the challenge for IR researchers is to define and pursue research programs that maintain their relevance in a rapidly changing environment. One problem is
that the priorities that IR researchers place on research issues are not necessarily the same as those of companies and government agencies that use and sell IR systems. Understanding those priorities and the operational experience behind them will be part of the process of deciding which issues are of fundamental importance and which are more transient.

[1]: www.searchtools.com/info/info-retrieval.html
[2]: www.dcs.gola.ac.uk/keith/Chatpter.1/Ch.1.html
[4]: Ricardo Baeza Yates and Berthier Ribeiro Neto –“Modern Information Retrieval”
3. **The Web as a “Lexical Resource”**

When most people talk about Internet search engines, they really mean World Wide Web search engines. Before the Web became the most visible part of the Internet, there were already search engines in place to help people find information on the Net. Programs with names like "gopher" and "Archie" kept indexes of files stored on servers connected to the Internet, and dramatically reduced the amount of time required to find programs and documents [1]. In the late 1980s, getting serious value from the Internet meant knowing how to use gopher, Archie, Veronica and the rest.

**Search Engines**

The term "search engine" is often used generically to describe both crawler-based search engines and human-powered directories. These two types of search engines gather their listings in radically different ways [1].

**Crawler-Based Search Engines**

Crawler-based search engines, such as Google, create their listings automatically [1][2]. They "crawl" or "spider" the web, then people search through what they have found.

If you change your web pages, crawler-based search engines eventually find these changes, and that can affect how you are listed. Page titles, body copy and other elements all play a role.

**Human-Powered Directories**

A human-powered directory, such as the Open Directory, depends on humans for its listings [1]. You submit a short description to the directory for your entire site, or editors write one for sites they review. A search looks for matches only in the descriptions submitted.

Changing your web pages has no effect on your listing. Things that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The only exception is that a good site, with good content, might be more likely to get reviewed for free than a poor site.

"Hybrid Search Engines" Or Mixed Results

In the web's early days, it used to be that a search engine either presented crawler-based results or human-powered listings. Usually, a hybrid search engine will favor one type of listings over another. However, it does also present crawler-based results, especially for more obscure queries [1][3].

Today, most Internet users limit their searches to the Web, and so is our discussion limited to search engines that focus on the contents of Web pages.
3.1 Looking at the Web

Before a search engine can tell you where a file or document is, it must be found. To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called spiders, to build lists of the words found on Web sites. When a spider is building its lists, the process is called Web crawling. In order to build and maintain a useful list of words, a search engine's spiders have to look at a lot of pages [2].

How does any spider start its travels over the Web? The usual starting points are lists of heavily used servers and very popular pages. The spider will begin with a popular site, indexing the words on its pages and following every link found within the site [2]. In this way, the spidering system quickly begins to travel, spreading out across the most widely used portions of the Web.

![Diagram of a Web spider crawling the web](image)

*Figure 3.1: "Spiders" take a Web page's content and create key search words that enable online users to find pages they're looking for.*

3.2 Working of a Search Engine

In this section we intend to describe how Google.com began as an academic search engine. They built their initial system to use multiple spiders, usually three at one time. Each spider could keep about 300 connections to Web pages open at a time. At its peak performance, using four spiders, their system could crawl over 100 pages per second, generating around 600 kilobytes of data each second [3].

Keeping everything running quickly meant building a system to feed necessary information to the spiders. The early Google system had a server dedicated to providing URLs to the spiders. Rather than depending on an
Internet service provider for the Domain Name Server (DNS) that translates a server's name into an address, Google had its own DNS, in order to keep delays to a minimum.

When the Google spider looked at an HTML page, it took note of two things:

- The words within the page
- Where the words were found

Words occurring in the title, subtitles, meta tags and other positions of relative importance were noted for special consideration during a subsequent user search. The Google spider was built to index every significant word on a page, leaving out the articles "a," "an" and "the" (i.e. the stop words). Other spiders take different approaches.

These different approaches usually attempt to make the spider operate faster, allowing users to search more efficiently. For example, some spiders will keep track of the words in the title, sub-headings and links, along with the 100 most frequently used words on the page and each word in the first 20 lines of text. Lycos is said to use this approach to spidering the Web.

Other systems, such as AltaVista, go in the other direction, indexing every single word on a page, including "a," "an," "the" and other "insignificant" words. The push to completeness in this approach is matched by other systems in the attention given to the unseen portion of the Web page, the meta tags.

3.3 Meta Tags

Meta tags allow the owner of a page to specify key words and concepts under which the page will be indexed. This can be helpful, especially in cases in which the words on the page might have double or triple meanings -- the meta tags can guide the search engine in choosing which of the several possible meanings for these words is correct. There is, however, a danger in over-reliance on meta tags, because a careless or unscrupulous page owner might add meta tags that fit very popular topics but have nothing to do with the actual contents of the page. To protect against this, spiders will correlate meta tags with page content, rejecting the meta tags that don't match the words on the page.

All of this assumes that the owner of a page actually wants it to be included in the results of a search engine's activities. Many times, the page's owner doesn't want it showing up on a major search engine, or doesn't want the activity of a spider accessing the page. Consider, for example, a game that builds new, active pages each time sections of the page are displayed or new links are followed. If a Web spider accesses one of these pages, and begins following all of the links for new pages, the game could mistake the activity for a high-speed human player and spin out of control. To avoid situations like this, the robot exclusion protocol was developed. This
The Web as a Lexical Resource

protocol, implemented in the meta-tag section at the beginning of a Web page, tells a spider to leave the page alone -- to neither index the words on the page nor try to follow its links.

3.4 Building a Search

Searching through an index involves a user building a query and submitting it through the search engine. The query can be quite simple, a single word at minimum. Building a more complex query requires the use of Boolean operators that allow you to refine and extend the terms of the search [5]. The Boolean operators most often seen are:

**AND** - All the terms joined by "AND" must appear in the pages or documents. Some search engines substitute the operator "\+" for the word AND.

**OR** - At least one of the terms joined by "OR" must appear in the pages or documents.

**NOT** - The term or terms following "NOT" must not appear in the pages or documents. Some search engines substitute the operator "\-" for the word NOT.

**FOLLOWED BY** - One of the terms must be directly followed by the other.

**NEAR** - One of the terms must be within a specified number of words of the other.

**Quotation Marks** - The words between the quotation marks are treated as a phrase, and that phrase must be found within the document or file.

3.5 The Future Search

The searches defined by Boolean operators are literal searches -- the engine looks for the words or phrases exactly as they are entered. This can be a problem when the entered words have multiple meanings. "Bed," for example, can be a place to sleep, a place where flowers are planted, the storage space of a truck or a place where fish lay their eggs. If you're interested in only one of these meanings, you might not want to see pages featuring all of the others. You can build a literal search that tries to eliminate unwanted meanings, but it's nice if the search engine itself can help out.

One of the areas of search engine research is concept-based searching. Some of this research involves using statistical analysis on pages containing the words or phrases you search for, in order to find other pages you might be interested in. Obviously, the information stored about each page is greater for a concept-based search engine, and far more processing is required for each search. Still, many groups are working to improve both results and performance of this type of search engine. Others have moved on to another area of research, called natural-language queries.

The idea behind natural-language queries is that you can type a question in the same way you would ask it to a human sitting beside you -- no need to keep track of Boolean operators or complex query structures. The most popular natural language query site today is AskJeeves.com, which parses the query for keywords that it then applies to the index of sites it has built. It
The Web as a Lexical Resource

only works with simple queries; but competition is heavy to develop a natural-language query engine that can accept a query of great complexity.

3.6 What Document Features Make a Good Match to a Query?

We have discussed how search engines work, but what features of a query makes for good matches? Let's look at the key features and consider some pros and cons of their utility in helping to retrieve a good representation of documents/pages [6].

- **Term frequency**: A measure of how often a term is found in a collection of documents. It is discussed in more detail in the next section.

- **Location of terms**: Many search engines give preference to words found in the title or lead paragraph or in the metadata of a document. Some studies show that the location — in which a term occurs in a document or on a page — indicates its significance to the document [7].

- **Link analysis** [8]: Web-based search engines have introduced one dramatically different feature for weighting and ranking pages [6]. Link analysis works somewhat like bibliographic citation practices, such as those used by Science Citation Index. Link analysis is based on how well-connected each page is, as defined by Hubs and Authorities, where Hub documents link to large numbers of other pages (out-links), and Authority documents are those referred to by many other pages, or have a high number of "in-links" (J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. 1998, pp. 668-77).

- **Popularity**: Google and several other search engines add popularity to link analysis to help determine the relevance or value of pages. Popularity utilizes data on the frequency with which a page is chosen by all users as a means of predicting relevance. While popularity is a good indicator at times, it assumes that the underlying information need remains the same.

- **Date of Publication**: Some search engines assume that the more recent the information is, the more likely that it will be useful or relevant to the user. The engines therefore present results beginning with the most recent to the less current.

- **Length**: While length per se does not necessarily predict relevance, it is a factor when used to compute the relative merit of similar pages. So, in a choice between two documents both containing the same query terms, the document that contains a proportionately higher occurrence of the term relative to the length of the document is assumed more likely to be relevant.

- **Proximity of query terms**: When the terms in a query occur near to each other within a document, it is more likely that the document is relevant to the query than if the terms occur at greater distance.
• **Proper nouns** sometimes have higher weights, since so many searches are performed on people, places, or things. While this may be useful, if the search engine assumes that you are searching for a name instead of the same word as a normal everyday term, then the search results may be peculiarly skewed.

[3]: http://computer.howstuffworks.com/search-engine1.htm
[5]: www.webreference.com/content/search/how.html
[6]: www.infotoday.com/searcher/may01/liddy.htm
[7]: www.searchenginedictionary.com/ishtml#Inversedocumentfrequency
[8]: http://searchenginewatch.com/searchtoday/article.php/2158431
4. Natural Language Processing in Information Retrieval Research

4.1 Natural Language Processing

To avoid forcing searchers to memorize Boolean or other query languages, some systems allow them to type in a question, and use that as the query: this is known as "Natural Language Processing" (NLP). The simplest processing just removes stopwords and uses a vector search or other statistical approach. Some sophisticated systems try to extract concepts using linguistic analysis, and match those against concepts extracted by the indexer [1]. Others try to categorize the form of the question and use it to define the query, so "who is" questions are not treated the same as "how many" or "why": a good example of this approach is the AskJeeves system.

4.2 N-grams

N-grams are a popular technique in statistical natural language processing. An n-gram is a sub-sequence of n items from a given sequence.

More concisely, an n-gram is a model of predicting $X_i$ based on $X_{i-1}$, $X_{i-2}$, ..., $X_{i-n}$. In application to language modeling, because of computational limitations and the open nature of language (there are infinitely many possible words), independence assumptions are made so that each word depends only on the last n words [2].

An n-gram of size 1 is a "unigram"; size 2 is a "bi-gram" (or, more appropriately but less commonly, a di-gram); size 3 is a "trigram"; and size 4 or more is simply called an "n-gram".

For parsing, words are modeled such that each n-gram is composed of n words. For a sequence of words, (for example "the dog smelled like a skunk"), the trigrams would be: "the dog smelled", "dog smelled like", "smelled like a", and "like a skunk". For sequences of characters, the 3-grams (sometimes spelled "trigrams") that can be generated from "good morning" are "goo", "ood", "od ", "d m", " mo", "mor" and so forth. Some practitioners preprocess strings to remove spaces, others do not. In almost all cases, punctuation is removed by preprocessing. N-grams can also be used for sequences of words or, in fact, for almost any type of data. They have been used for example for extracting features for clustering large sets of satellite earth images and for determining what part of the Earth a particular image came from.

By converting an original sequence of items to $n$-grams, it can be embedded in a vector space (in other words, represented as a histogram), thus allowing the sequence to be compared to other sequences in an efficient manner [2]. For example, if we convert strings with only letters in the English alphabet into 3-grams, we get a $26^3$-dimensional space (the first dimension measures the
number of occurrences of "aaa", the second "aab", and so forth for all possible combinations of three letters). Note that using this representation we lose information about the string. For example, both the strings "abcbab" and "bcbab" give rise to exactly the same 2-grams. However, we know empirically that if two strings of real text have a similar vectorial representation (such as a small cosine distance) then they are likely to be similar. Other metrics have also been applied to vectors of n-grams with varying, sometimes better, results. For example z-scores have been used to compare documents by examining how many standard deviations each n-gram differs from its mean occurrence in a large collection, or corpus, of documents (which form the "background" vector).

4.3 Vector search

Vector searching takes the query and compares it to the text indexed to find the best match on the most words using complex mathematical formulae [3]. This approach works best for multi-word natural-language question such as "What is the normal weather in Berkeley?" and in finding documents similar to an existing document. Even on very large indexes, searching is very quick. However, the algorithms can use relevance ranking that many people find disconcerting: some pages can be ranked high even if they do not contain all the search terms. See also Boolean searching.

In the vector-space model, each document in a corpus is represented as a list or weighted ‘vector’ of the words (or phrases) it contains [4]. A portion of a document and its associated weighted vector are shown in the table 4.1a and table 4.1b, respectively. Each word that occurs in the table 4.1b has an associated weight, which is intended to represent the relative importance of that word in determining the theme of the document.

*Term frequency [5]:* How frequently a query term appears in a document is one of the most obvious ways of determining a document's relevance to a query. While most often true, several situations can undermine this premise. First, many words have multiple meanings — they are polysemous. Think of words like "pool" or "fire." Many of the non-relevant documents presented to users result from matching the right word, but with the wrong meaning. TF is combined with inverse document frequency (IDF) as a means of determining which documents are most relevant to a query. TF is sometimes also used to measure how often a word appears in a specific document.

*Inverse Document Frequency [6]:* A measure of how rare a term is in a collection, calculated by total collection size divided by the number of documents containing the term. Very common terms ("the", "and" etc.) will have a very low IDF and are therefore often excluded from search results. These low IDF words are commonly referred to as "stop words".

Also, in a collection of documents in a particular domain, such as education, common query terms such as "education" or "teaching" are so common and
occur so frequently that an engine's ability to distinguish the relevant from the non-relevant in a collection declines sharply [4]. Search engines that don't use a tf/idf weighting algorithm do not appropriately down-weight the overly frequent terms, nor are higher weights assigned to appropriate distinguishing (and less frequently-occurring) terms, e.g., "early-childhood."

**Table 4.1a: Portion of a document**

Grave's disease is an autoimmune disease of the thyroid gland that affects the ability of the patient to retain and metabolize iodine.

**Table 4.1b: Weight of each vector**

<table>
<thead>
<tr>
<th>Term</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grave's</td>
<td>5.6</td>
</tr>
<tr>
<td>Thyroid</td>
<td>2.7</td>
</tr>
<tr>
<td>Gland</td>
<td>2.5</td>
</tr>
<tr>
<td>Iodine</td>
<td>2.3</td>
</tr>
<tr>
<td>Autoimmune</td>
<td>1.3</td>
</tr>
<tr>
<td>Disease</td>
<td>1.2</td>
</tr>
<tr>
<td>Patient</td>
<td>1.0</td>
</tr>
<tr>
<td>Metabolize</td>
<td>0.9</td>
</tr>
<tr>
<td>Retain</td>
<td>0.9</td>
</tr>
<tr>
<td>Affects</td>
<td>0.5</td>
</tr>
<tr>
<td>Ability</td>
<td>0.2</td>
</tr>
<tr>
<td>Is</td>
<td>0.0</td>
</tr>
<tr>
<td>An</td>
<td>0.0</td>
</tr>
<tr>
<td>Of</td>
<td>0.0</td>
</tr>
<tr>
<td>The</td>
<td>0.0</td>
</tr>
</tbody>
</table>

This weight is usually some function of the frequency of the term in the document (term frequency, TF), so that terms that occur more often in the document will be given higher weights. To prevent common words with less semantic value from dominating the vector, term weights are typically normalized by IDF (the Inverse Document Frequency of the term), which varies inversely with the frequency of the term in the corpus as a whole. Many formulations of the TF × IDF weighting scheme exist. One of these is:

\[(1 + \log_{10}TF) \log(\frac{N}{DF})\]

where DF is document frequency and N is the size of a corpus.

The individual weights in a vector collectively determine the direction that the vector will point in ‘word space’ [6], wherein each dimension of the space corresponds to a single word or phrase in the document in figure 4.1. In this way, the direction of the vector encodes the content of the underlying document. In c, two documents are shown in a two dimensional space (for
simplicity). The angle (q) between the vectors can be used to measure the similarity between the contents of the underlying documents.

![Diagram](image)

*Figure 4.1: direction of each vector in ‘word space’.*

---

[1]: http://research.microsoft.com/npl/
[3]: Christopher D. Manning and Hinrich Schutze – “Foundations of Statistical Natural Language Processing”
[4]: UAI2003-QA-DI.pdf
[5]: www.searchenginedictionary.com/ishtml#Inversedocumentfrequency
[6]: www.cs.bham.ac.uk/~7Epxc/nlpa/2002/AI-HO-introNPL.html
5. Question Answering

Question answering (QA) is a type of information retrieval. Given a collection of documents (such as the World Wide Web or a local collection) the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex natural language processing (NLP) techniques than other types of information retrieval such as document retrieval, and it is sometimes regarded as the next step beyond search engines [1].

QA research attempts to deal with a wide range of question types including: fact, list, definition [2], How, Why, hypothetical, semantically-constrained, and cross-lingual questions. Search collections vary from small local document collections, to internal organization documents, to compiled newswire reports, to the World Wide Web.

- Closed-domain question answering deals with questions under a specific domain (for example, medicine or automotive maintenance), and can be seen as an easier task because NLP systems can exploit domain-specific knowledge such as ontologies [1].
- Open-domain question answering deals with questions about nearly everything, and can only rely on general ontologies and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer [1].

5.1 Generic System Architecture

The first QA systems were developed in the 1960s and they were basically natural-language interfaces to expert systems that were tailored to specific domains. In contrast, current QA systems use text documents as their underlying knowledge source and combine various natural language processing techniques to search for the answers [2].

Current QA systems typically include a question classifier module that determines the type of question and the type of answer. After the question is analyzed, the system typically uses several modules that apply increasingly complex NLP techniques on a gradually reduced amount of text. Thus, a document retrieval module uses search engines to identify the documents or paragraphs in the document set that are likely to contain the answer. Subsequently a filter pre-selects small text fragments that contain strings of the same type as the expected answer [2]. For example, if the question is "Who invented Penicillin" the filter returns text that contain names of people. Finally, an answer extraction module looks for further clues in the text to determine if the answer candidate can indeed answer the question.
5.2 Question Answering methods

QA is very dependent on a good search corpus - for without documents containing the answer, there is little any QA system can do. It thus makes sense that larger collection sizes generally lend well to better QA performance, unless the question domain is orthogonal to the collection. The notion of data redundancy in massive collections, such as the web, means that nuggets of information are likely to be phrased in many different ways in differing contexts and documents, leading to two benefits [3]:

a) By having the right information appear in many forms, the burden on the QA system to perform complex NLP techniques to understand the text is lessened.
b) Correct answers can be filtered from false positives by relying on the correct answer to appear more times in the documents than instances of incorrect ones.

Shallow [4]
Some methods of QA use keyword-based techniques to locate interesting passages and sentences from the retrieved documents and then filter based on the presence of the desired answer type within that candidate text. Ranking is then done based on syntactic features such as word order or location and similarity to query.

When using massive collections with good data redundancy, some systems use templates to find the final answer in the hope that the answer is just a reformulation of the question. If you posed the question "What is a dog?", the system would detect the substring "What is a X" and look for documents which start with "X is a Y". This often works well on simple "factoid" questions seeking factual tidbits of information such as names, dates, locations, and quantities.

Deep [4]
However, in the cases where simple question reformulation or keyword techniques will not suffice, more sophisticated syntactic, semantic and contextual processing must be performed to extract or construct the answer. These techniques might include named-entity recognition, relation detection, coreference resolution, syntactic alternations, word sense disambiguation, logic form transformation, logical inferences (abduction) and commonsense reasoning, temporal or spatial reasoning and so on. These systems will also very often utilize world knowledge that can be found in ontologies such as WordNet, or the Suggested Upper Merged Ontology (SUMO) to augment the available reasoning resources through semantic connections and definitions.

More difficult queries such as Why or How questions, hypothetical postulations, spatially or temporally constrained questions, dialog queries, badly-worded or ambiguous questions will all need these types of deeper understanding of the question. Complex or ambiguous document passages likewise need more NLP techniques applied to understand the text.
Statistical QA, which introduces statistical question processing and answer extraction modules, is also growing in popularity in the research community [3]. Many of the lower-level NLP tools used, such as part-of-speech tagging, parsing, named-entity detection, sentence boundary detection, and document retrieval, are already available as probabilistic applications.

5.3 Cross-Language Question Answering system

Cross Language Evaluation Forum (CLEF) [5] is a reference workshop to evaluate Information Retrieval and QA systems operating on European Languages. The CLEF workshop is organized each year to compare the participants systems in a wide range of categories, such as Mono-, Bi- and Multilingual Information Retrieval, Interactive Cross Language Information Retrieval, Multiple Language QA, etc. An important extension of these tasks is the Multilingual QA which accepts questions in any language on multilingual document collections. In this task, it is very interesting the use of methodologies of document (or passage) retrieval as independent as possible of the language.

The QA system, adapted with modifications, in our project is the QA JIRS-based systems for Spanish CLEF-2005 [5]. However, the system is modified to be able to extract snippets in English with a list of English stop-words.

The figure 5.1 list the accuracy achieved by different systems that were presented at the Spanish CLEF 2005.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>inao051eses</td>
<td>42.00%</td>
</tr>
<tr>
<td>tova051eses</td>
<td>41.00%</td>
</tr>
<tr>
<td>upv051eses</td>
<td>33.50%</td>
</tr>
<tr>
<td>alia051eses</td>
<td>33.00%</td>
</tr>
</tbody>
</table>

*Figure 5.1 Statistical result of the Spanish CLEF-2005*

Additionally, a Translator is added to the system to make it Cross-Language. Thus with experimental observations, a JAVA QA translator has been implemented to interpret the Urdu language query into English with a maximum accuracy. This interpreter is intended to translate queries and not sentences.

In the later sections we will analyze the results obtained with the English QA system and the Urdu-English Cross-Language QA system.
Question Answering


[2]: Boris Galitsky – “Natural Language Q.A system”

[3]: http://acl.ldc.upenn.edu/A/A00/A00-1023-pdf

by Rohini Srihari, Weili- “A Question Answering System Support by Information Extraction”

[4]: www.cadeproject.com/csharp/tf.df.asp

Nowadays there is a large amount of digital documents accessible from the web. These documents may satisfy almost every information need. However, all these documents are practically useless without the appropriate mechanisms that help users to find the required information when they need it.

For several decades, researchers have pursued the goal of developing computational machinery with the ability to generate answers to freely-posed questions. General question answering systems depend on techniques for analyzing questions and for composing answers from some corpus of knowledge. This is a challenging problem because the corpus may not contain an explicit matching answer or may contain multiple variants of relevant answers or answer components.

We have been interested in procedures that enlist the poorly-structured but copious resources of the Web for answering questions. Web-based question answering systems typically employ rewriting procedures for converting components of questions into sets of queries posed to search engines, along with techniques for converting query results into one or more answers. These systems are capable to answer questions formulated by the users in natural language. For instance, given the question like “What is the capital of Pakistan”, a QA system responds “Islamabad” instead of returning a list of related documents to the country of Pakistan.

6.1 Technical aspects of Question-Answering Systems

Most text retrieval systems operate at the level of entire documents. In searching the web, complete web pages or documents are returned. There has been a recent surge of interest in finer-grained analyses focused on methods for obtaining answers to questions rather than retrieving potentially relevant documents or best-matching passages from queries—tasks information retrieval (IR) systems typically perform. The problem of question answering hinges on applying several key concepts from information.

Recent developments in QA are mainly focused on answering factual questions (those having a simple named entity as the answer), and are mainly suited to English as the target language. These kind of systems performed relatively well when retrieving 250-byte passages, but less well when they attempted to locate the concrete answers (restricted to 50-byte long).

In contrast to many rich natural language approaches, there have been developments of question answering system that attempt to solve the difficult matching and extraction problems by reformulations.

In figure 6.1 a general architecture of the proposed system can be seen. It consists of three major modules:
6.2 Query Reformulations

Given a question, this module generates a set of query reformulations. These reformulations are expressions that have the probability of being used in writing down the expected answer. Several experiments were performed in order to determine the most general and useful reformulations. Many questions and answers were analyzed to observe the relationships between the answer phrase and its corresponding question phrase. To better illustrate the reformulations used, we analyze the question:

What is the capital of Pakistan?

In the algorithms described below, a question Q is said to be a set of words, i.e.:

\[ Q = \{w_0, w_1, \ldots, w_{n-1}\} \]

Here \( w_0 \) corresponds to the wh-word, that is to say the questioning word, and \( n \) indicates the number of words of the question. On the other hand, a query reformulation \( R \) is represented as a symbol string. This string consists of
Web-Based QA Systems

words, spaces, and quotation marks, and it satisfies the format of a conventional search engine query.

For instance the reformulation \( R = w_i w_j \) corresponds to the query \( w_i \text{ AND } w_j \), while the reformulation \( R = "w_i w_j" \) to the query “\( w_i \text{ AND } w_j " \).

Now let us take a look at the five reformations that the QA system use inorder to extract possible answers to the factual questions.

First reformulation: “bag of words”

This reformulation is the set of non stop-words, that are a list of words without the prepositions, conjunctions and articles, of the question. It is built as follows:

1. For each \( w_i \) \( Q | i > = 1 \)
2. If \( w_i \) is not a stop word
3. \( R \neg w_i \)
4. Save \( R \)

The reformulation generated for the example query is:

Who received the Nobel Prize of Peace in 1992?

Second reformulation: “components”

In this reformulation, the question is divided in components. A component is an expression delimited by a preposition. Therefore, a question \( Q \) with \( m \) prepositions is represented by a set of components:

\[ C = \{c_1, c_2, \ldots, c_{m+1}\} \]

Each component \( c_i \) is a substring (subset of words) of the original query. New reformulations are defined combining these components as follows:

1. Determine the set of components \( C \) from \( Q \)
2. \( R = "c_1" "c_2" \ldots "c_{m+1}" \)
3. Save \( R \)
4. For each permutation \( C' \) of \( C \)
5. \( R = "c'_1 c'_2 \ldots c'_{m+1}" \)
6. Save \( R \)

Some examples of this kind of query reformulations are:

“received the Nobel Prize” “of Peace” “in 1992”
“of Peace received the Nobel Prize in 1992”
“in 1992 received the Nobel Peace Prize”
Third reformulation: “components without the first word”

In order to construct this set of reformulations, the main verb of the question (commonly expressed by the word w1) is eliminated, and then we apply the method of reformulations by components.

Some examples of these reformulations are:

“in 1992 the Nobel Peace Prize”
“the Nobel Prize“of Peace” “in 1992”

Fourth reformulation: “components without the first and second words”

Here, an assumption about the presence of an auxiliary verb is made. Thus it is necessary to omit the words w1 and w2 from the question, and then to apply the method of reformulations by components.

Two examples of this kind of reformulations are:

“Nobel Peace Prize in 1992”
“in 1992 Nobel Peace Prize”

It is evident from the previous examples that some reformulations do not make much sense. For instance, “in 1992 of Peace received the Nobel Prize”. However, the probability of finding snippets from them is very low. On the contrary, most reformulations are syntactically correct, and they produce several snippets containing the desired answer.

Here it is worth mentioning to note that some reformulations allow collecting snippets using a synonym of the main verb of the question. For instance, the query reformulation “Nobel Peace Prize in 1992” allows extracting snippet(s) with the phrase: “won the Nobel Peace Prize in 1992”.

Fifth reformulation: “verb movement”

One of the main observations after examine a list of factual questions was that the verb is frequently used right after the wh –word, i.e. right after the questioning word. It is also know that in order to transform an interrogative sentence into a declarative one is necessary to eliminate the verb, or to move it to the final position of the sentence. The resulting sentence is expected to be more abundant in the web that the original one.

Therefore to take advantage of this phenomenon, but without using any kind of linguistic resource, we propose to build a set of query reformulations eliminating, or moving to the end of the sentence, the first and second words from the question. The reason to include the second word was to consider the cases when an auxiliary verb exists.
The algorithm to build these reformulations is as follows:

1. Set $w_0 = ""$
2. $R = "w_1 w_2 \ldots w_{n-1}"
3. Save $R$
4. For $i = 1$ to $2$
5. $R = "w_{i+1} w_{i+2} \ldots w_{n-1}"
6. Save $R$
7. $R = "w_{i+1} w_{i+2} \ldots w_{n-1} w_{i-1} w_i"
8. Save $R$

Two examples of these kind of reformulations are:
“the Nobel Peace Prize in 1992 received”
“Nobel Peace Prize in 1992”

Snippets recollection

Once the set of reformulations has been generated and sent to a search engine (currently we are using Google), this module collects the returned snippets. Here is an example of a snippet collected from the reformulation “the capital of Pakistan is”:

…..1 Projects *Islamabad is the Capital of Pakistan* 1 Introduction Many capital of Pakistan, the first team of experts of the consultant arrived at Rawalpindi early in November 1959 and a methodical collection of data was ..

6.3 Answer extraction

This module extracts all the n-grams (from unigrams to pentagrams) from the collected snippets. Each n-gram is defined as a possible answer to the given question. Then, using some statistical criteria, it ranks the n-grams by decreasing likelihood of being the correct answer. The top five n-grams are selected as possible answers.

Following is a description of the three different methods for the n-gram extraction and ranking.

Method of relative frequency:
1. Extract the twenty most frequent unigrams.
2. Compute the relative frequency of each unigram.

If $G_1$ is the set of frequent unigrams, and $f_w$ indicates the frequency of occurrence of the unigram $w \in G_1$, then the relative frequency of $w$ is computed as follows:

$$P_w = \frac{f_w}{\sum_{i \in G_i} f_i}$$
3. Determine all the $n$-grams, from bi-grams to pentagrams, built from the set frequent unigrams.
4. Rank the $n$-grams based on their relative frequency. The relative frequency of a $n$-gram $g = (w_1… wn)$ is computed as follows:

$$P_g = \frac{1}{n} \sum_{i=1}^{n} P_w$$

Here, $P_w$ indicates the relative frequency of the unigram $w \in g$.

5. Select the top five $n$-grams as possible answers. Applying this method to an example question, “Who received the Noble Prize in 1992?”, we obtained these answers:

Menchu       0.05541
Rigoberta Menchu      0.05074
Rigoberto      0.04607
Rigoberta Menchu recibio     0.04005
guatemalteca Rigoberta Menchu    0.03860

*Method of regular expressions:*

1. Extract the twenty most frequent unigrams that satisfy a given typographic criteria (i.e., words starting with an uppercase letter, numbers and names of months).

2. Determine all the $n$-grams, from bi-grams to pentagrams, built from the set frequent unigrams.

3. Rank the $n$-grams, in decreasing order, based on the number of words.

4. Select the top five $n$-grams as possible answers.

For the example question we obtained the following $n$-grams:

Rigoberta Menchu Tum
Rigoberta Menchu Recibio
Rigoberta Menchu
Menchu Tum
Menchu Recibio

*Method of regular expressions plus a compensated frequency:*

1. Extract the twenty most frequent unigrams that satisfy a given typographic criteria (i.e., words starting with an uppercase letter, numbers and names of months).
2. Determine all the n-grams, from bi-grams to pentagrams, built from the set frequent unigrams.

3. Rank the n-grams based on their compensated relative frequency. The compensated relative frequency of a n-gram \( g(n) = (w_1 \ldots w_n) \) is computed as follows:

\[
\text{compensated relative frequency} = \frac{j(i)}{|G_i|} f(j(i))
\]

where \( Gi \) is the set of n-grams of size i, \(|Gi|\) indicates the cardinality of this set, \( j(i) \) is an n-gram \( j \) of size i contained in \( g(n) \), and \( f(j(i)) \) is the frequency of occurrence of this n-gram.

4. Select the top five n-grams as possible answers. Applying this method we obtained the following answers:

- Rigoberta Menchu 0.07418
- Rigoberta Menchu Tum 0.05753
- Menchu 0.05541
- Rigoberta Menchu Recibio 0.05143
- Rigoberta 0.04607

It is important to notice that the method of relative frequency favors the short n-grams, while the method of regular expressions the large ones. The last method, the method of regular expressions plus a compensated frequency, combines the advantages of both previous methods.

However it applies a compensation factor in order to avoid favoring the short answers. This method produced the best results in the experiments.

### 6.4 Experimental results

Table 6.1: English QA system experimental results

<table>
<thead>
<tr>
<th>English QA system Results</th>
<th>Bags of words</th>
<th>Components</th>
<th>Component missing word-0</th>
<th>Component missing word-01</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67</td>
<td>14</td>
<td>26</td>
<td>23</td>
<td>13</td>
</tr>
</tbody>
</table>

From the results shown in Table 6.1, it is clear that “Bags of Words” reformulation produces the best of the results. However, in some cases where the user question is considerably complex and long, the reformulation is not efficient.
From the results we conclude that the system work best with “Bags of Words” reformulation when the questions are well structured, simple and short.
7. **Urdu-English Translation**

7.1 **An Introduction to Urdu**

The language of the road, and most widely spoken and understood language in Pakistan and India is Urdu (also called Hindustani, Hindi, Bihari, Dakhini). More than 220 million people in subcontinent regard it as their mother tongue and is actively used by 400 million people in India and Pakistan in their daily life at work and home. Outside subcontinent large Urdu speaking communities are found in USA, UK, Mauritius, South Africa, Yemen, Uganda, Singapore, Nepal, New Zealand and Germany [1].

Urdu is the national language of Pakistan and an official language of State of UP in India. It unites all people and all communities, whatever their mother tongue.

It has been influenced and enriched by Dravidian, Turkish, Farsi, Arabic, Portuguese and English.

Urdu is written in modified Arabic script, from right to left. It has developed some very sophisticated and beautiful calligraphical scripts for printing. The most popular is "Nastaliq". Most of the books, newspapers and general printed matters are printed in this script.

Significant work has been done on the linguistic aspects of the language like orthography and lexicography. However with the advent of computers it is a natural desire of its adherents to harness the power of the computer to increase their productivity and efficiency in the usage of the language.

When you send your message, sell you product having label and instructions only in English to the markets in India and Pakistan, you are perhaps convinced that everybody would understand it, after all English is the official language in these two counties. But the fact is that by doing so you are reaching only a tiny fraction of these huge markets, 5% of the potential customers at the most. When you say it in Urdu you are understood by 90% of Pakistani and some 30% of Indians.

Urdu is approximately the twentieth most populous natively spoken language in the world.

[1]: [www.urdusolutions.com](http://www.urdusolutions.com)
7.2 Urdu-English Translation System Architecture

Any translator may translate the words in a phrase (question, answer or sentence) into another language. But translating into a meaningful phrase from one language to another is another task.

Since translation plays a key role in Cross-Language QA systems, it is required to get as accurate translation as possible from source language (i.e. Urdu language) to target language (i.e. English language).

7.2.1 Translation system architecture

The following figure 8.1 illustrates the layout of the proposed translation system:

1. User input native language query
2. Words are counted in the input phrase
3. Translation of each word into English
4. Formulating the positioning of each word

![Figure 7.1 Translator architecture](image)
7.2.2 How the translator works

After performing several experiments in order to determine the most general and precise structure of a question in Urdu language. It has been observed that in Urdu language, a subject is placed before stating the key questioning word.

For example, a question in Urdu:
*Pakistan ka capital kya hai?*  
*(What is the capital of Pakistan?)*

Let us analyze this question word by word.

<table>
<thead>
<tr>
<th>Urdu</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pakistan</td>
<td>Pakistan</td>
</tr>
<tr>
<td>Ka</td>
<td>of</td>
</tr>
<tr>
<td>Capital</td>
<td>capital</td>
</tr>
<tr>
<td>Kya</td>
<td>what</td>
</tr>
<tr>
<td>Hai</td>
<td>is</td>
</tr>
</tbody>
</table>

*A question in Urdu*

There are many ways to frame a question in Urdu language, since Urdu is considered as a more expressive language. Therefore, we have adopted the most commonly used Urdu query structure.

The factual question placed by a user, must be in a format as follows:

```
Question subject -- Questioning keyword
Pakistan ka capital -- Kya hai
```

This is the format the user should input his/her Urdu factual question.

Word of Urdu broken down into “tokens”

The Urdu question is broken down into words, “tokens”, to get a translation of every Urdu token into English. Thus, it is required to have a knowledge base of Urdu to English words at a backup. However, here we were faced with some difficulties. Urdu is a language that uses a lot of expressive words for which there is not a single meaningful English translation.

To tackle this problem some techniques were adapted for generating a more precise and accurate question.

1. The adjective before a main proper noun in the Urdu question must be connected to it (noun) with a “-“.

For example:

*Urdu question:* “Pehla EU ka Finnish-Commissionary koun tha? “
*English question:* “Who was the first Finnish Commisionary of EU?”
2- Some words in Urdu are same as English.

3- Urdu has many extra words for which there is no English translation. When the Urdu question is broken down to tokens some tokens will not have any translation. Thus, we have to get rid of such words at the end.

To tackle this problem, we replaced a “-” to get rid of these redundant words. However, a blank “ “ could not be placed as in some cases we want to keep words with no translation (e.g. Name of person, country, car etc.).

For example:
Urdu question: “Davis Cup kis mulk nay jeeta?”
English translation: “What country won the Davis Cup?”

7.3 Searching for keywords

Once the tokens are all translated, the next step is to formulate the tokens to generate a meaningful English question. As specified before, in an Urdu question the questioning keyword is placed after the subject of the question. Thus, the system tries to locate this keyword. Once the keyword is located, the system proceeds to the next step of formulating the English question.

Formulating with stop words:

Here, stop words are the prepositions that are generated from the Urdu question. In an Urdu question, the proper noun is always placed before a proposition and the common noun or the descriptive noun of the proper noun is place after the preposition. However, this format is a reverse in an English question.

For example:

Urdu question: “Italy ka waziryAzum koun hai?”
English translated question after formulation: “Who is president of Italy?”

In this above question the Urdu term “ka” is translated as “of” in English. It is the article in the question. Thus the proper noun “Italy” must be placed after and the common noun “president” is placed before the preposition.

Translated question

Finally the translated question is stored in a text file. However, due to word by word translation some articles are often missed in the generated question. But, considering the fact that these articles don’t play a vital role in the QA system, not much concern is given to this fact.
Moreover, in some cases, the generated questions may not be in correct format, but reformulation in the QA system helps to retrieve a better and in many cases accurate answers. Nevertheless, sometimes the meaning of the questions maybe lost.

Experiments have been carried out using this translator to translate some 100 experimental questions into English.

### 7.4 Experimental results

The test questions used to test the English QA system were first manually translated to Urdu language. Then these questions were passed on to the Urdu-English translator developed to work with the QA system.

However, once translated, these questions were saved to a file which is later passed to the QA system for snippets’ collection. Table 7.1 list the results obtained, with the five reformulations (described in chapter 6), passing to the system the test questions.

<table>
<thead>
<tr>
<th></th>
<th>Urdu-English QA system Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bags of words</td>
<td>Components</td>
</tr>
<tr>
<td>54%</td>
<td>12%</td>
</tr>
<tr>
<td>Component missing</td>
<td>Component missing word-01</td>
</tr>
<tr>
<td>word-0</td>
<td>10%</td>
</tr>
<tr>
<td>16%</td>
<td>28%</td>
</tr>
</tbody>
</table>

From the table 7.1 it can be observed that the “Bags of Words” reformulation generated the best of the results. This shows that once omitting the questioning word, the remaining phrase of the question can be found in the answer of that specific question. However, “Component missing word-0”, produce the least of the snippets, because in some cases the word next to the question word is the main word of the question. Since the question is being translated from Urdu to English, the Translator, itself, omits the articles and other helping words, in “Component missing word-0” and “Component missing word-01”, so this reformulation, in some cases, tends to drop of the “subject” of the question.
8. **Urdu-English QA system versus English QA system**

Let us take a look at the comparison of the two systems. To estimate the performance of our Urdu-English QA system, we performed some experimental runs with the test questions on the English QA system, the results of which had been listed in chapter 6 in table 6.1.

*Table 8.1: Comparison of results for English questions vs. Urdu questions*

<table>
<thead>
<tr>
<th>COMPONENTS</th>
<th>COMPONENT NO</th>
<th>WORD-0</th>
<th>COMPONENT NO</th>
<th>WORD-01</th>
<th>VERBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>67%</td>
<td>14%</td>
<td>26%</td>
<td>23%</td>
<td>32%</td>
</tr>
<tr>
<td>Urdu</td>
<td>54%</td>
<td>12%</td>
<td>10%</td>
<td>16%</td>
<td>28%</td>
</tr>
</tbody>
</table>

From the comparison table, it is clear that the loss of accuracy is not much significant. However, efficiency of the system is reduced by putting the Translator in front. This is because when the system had translated the question from Urdu to English, the original phrase losses the standard question structure. Thus the system, unaware of these uncertain changes, sometimes reformulates the question phrase, losing the meaning. Moreover the connecting words in a question, like is, was, did, does, etc, are usually omitted during the translation procedures, thereby producing a improper sentence phrase in verb reformulations.

Below, in figure 8.1, the graphical representation sums up the results of the table 8.1 with target (English) language and proposed (Urdu) language.

*Figure 8.1: Graphical representation of table 8.1*
9. Conclusion and Future Work

After carrying out several experiments on the designed system, we concluded that in the QA system the keywords that are likely to appear in an answer must always be present in the question. However, in some cases the structure of the question is not of much concern since reformulations help get meaning out of the phrase.

During the development cycle, we came across some significant features of Urdu and English language. In Urdu language, usually the subject of the question is placed before the questioning word occurs in the phrase. Thus, our Translator has been designed considering this constrain. But, in some cases, when the question is considerably long, the Translator may not be able to produce 100% results and therefore resulting in improper English question phrase.

The pie-chart below, in figure 9.1, represents the answered snippets collected from the English test questions and Urdu-English test questions.

![Pie chart showing Answered Questions: English 56; 44% and Urdu 72; 56%](image)

*Figure 9.1: Comparison of the correct snippets collected*

Finally we conclude that Urdu Language can be used to implement a Cross-Language QA system. It is necessary to maintain, as close as possible, a standard English Language question phrase structure, where the first word is usually the questioning word itself. The later part of the question is usually part of the answer phrase.

Overall, in our snippets’ collection we observed that the best of the results were obtained in the “ANDS” (i.e. the Bags of Words) reformulation. This is mainly because, majority of English answer phrases are structured to contain the whole question phrase except the questioning word itself. Thus, keeping this point in mind, our QA is designed so that the first questioning word is always dropped before sending the user query to “Google” search. But in some cases where the Questioning word is placed in the middle or later half
of the question, “Bags of Words” reformulation to form an answering phrase, may not produce accurate results

On the other hand, “Verbs” reformulation is done five times. In this reformulation, the snippet recollection was, to some extend, accurate. However, in our experiments, it was observed that this type of reformulation word best with long questions because in shorter question phrases, there is a risk of losing the important words during reformulations. It is said that in a question phrase the article next to the question word is placed to just before the answer word in majority of the sentences. For example, in the following question, let us see the placing of the article “is”:

Q: What is the capital of Pakistan?

A: The capital of Pakistan is Islamabad.”

Furthermore, in future, more translators can be added to get better results and to obtain phrases as close as possible to English language.
Appendix I Cross-Language QA System

List of Urdu test questions:

5?

18?

21?
Appendix I

Cross-Language QA System

95?? ?? ?? ??e ?? ?????

96?? ?? ?? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ?? ?? ??? ??
English Questions

1. Where is Euskirchen?
2. At what festival is the "Leone d'Oro" awarded?
3. What does the abbreviation IRA stand for?
4. What group did John Lennon belong to?
5. What country was the recipient of aid from "Opération Turquoise"?
6. In what country was the "Real Plan" against inflation carried out?
7. What is the technical name for the "mad cow disease"?
8. What country is the estate agent Schneider from?
9. What country invaded Kuwait in 1990?
10. Under what other name is the EFA project also known?
11. What Nobel prize was Solzhenitsin awarded?
12. What is Mika Kaurismäki's brother's name?
13. Who is the Italian Prime Minister?
14. What country hosted the Winter Olympic Games?
15. What year was South Yemen's secession declared?
16. Where was Kazem Radjavi assassinated?
17. What is the capital of Iran?
18. What award did director Quentin Tarantino's film, "Pulp Fiction", win in the Cannes Film Festival?
19. What country was the communist leader Najibullah from?
20. What office did Rabbani hold when the civil war broke out in Afghanistan in 1992?
21. In all of the EU, how many people have the right to vote for a new European Parliament?
22. Who was the first Finnish Commissary of the EU?
23. What country is Edouard Balladur from?
24. What year did Ayrton Senna die?
25. How many Formula 1 championships did the Brazilian racing driver Ayrton Senna win?
26. What was the name of the singer and head of Nirvana?
27. Who is the president of the European Commission?
28. What year was the Berlin Wall torn down?
29. How many currency changes has Brazil undergone throughout its history?
30. What city is the European Parliament in?
31. At the World Football Championship, in what place did Italy finish?
32. How many European countries form part of the G7?
33. What does the abbreviation ETA stand for?
34. Who is the president of Spain?
35. On what date was the MIR space station put into orbit?
36. On what day of the year is the Spanish Constitution honored?
37. Who won the Wimbledon tournament?
38. What country won the Davis Cup?
39. In what brand of baby food were pesticides found?
40. What team won the NBA championship?
41. What country withdrew its troops from Bosnia?
42. Who won the European Final Four basketball tournament?
43. What is the highest peak of the Alpes?
44. What year did the Russian troops withdraw from Latvia?
45. What Russian president attended the G7 meeting in Naples?
46. On what date did the referendum concerning Norway’s incorporation into the EU take place?
47. What two ethnic groups are in conflict in Ruanda?
48. What ship did the oil tanker "New World" crash into?
49. Where did the accident of the oil tanker "Mar Egeo" take place?
50. Under what flag did the oil tanker "Aegean Sea" sail?
51. How much oil was spilled in the accident of the oil tanker "Aegean Sea"?
52. How much was the estimated damage caused by the accident of the oil tanker "Aegean Sea"?
53. Who paid the indemnities for the accident of the oil tanker "Aegean Sea"?
54. Who was the prosecuting attorney at the trial for the accident of the oil tanker "Aegean Sea"?
55. What is the abbreviation for the "World Wildlife Fund"?
56. How many countries form part of the International Whaling Commission?
57. Above what continent was the ozone hole detected?
58. Where did the annual meeting of the International Whaling Commission take place?
59. What year was the World Bank created?
60. What year was the International Monetary Fund created?
61. How many countries form part of the IMF?
62. Who is the president of Peru?
63. At what conference were the WB and the IMF created?
64. What is the currency of Vietnam?
65. Who is the author of the Brady Plan?
66. What office did Brady hold in the U.S. Treasury?
67. What company did Nicholas Brady found?
68. What Peruvian president was accused of bribery in the awarding of a railway construction contract?
69. What French ex-minister was sent to prison for corruption?
70. On what day did Vicente Albero, Spanish Minister of Agriculture, Fisheries and Food, resign?
71. On what day did the Spanish Parliament pass a bill for the creation of a special prosecutor's office to fight against corruption?
72. What Prime Minister was issued an indictment for corruption by the prosecutor's office in Milan?
73. On what day was Silvio Berlusconi issued an indictment for bribing Treasury agents?
74. On what day did the public prosecutor of Milan, Antonio di Pietro, resign?
75. What French nuclear submarine had an accident?
76. How many crew members died on the submarine "Emeraude"?
77. When did the Chernobyl disaster take place?
78. How many countries was the USSR divided into?
79. How many people is it estimated will die as a result of the Chernobyl disaster?
80. What is the world suicide rate?
81. How high is the Everest?
82. What does PLO stand for?
83. How many countries attended the "MED 21" conference?
84. Who is the president of the Palestinian National Authority?
85. What day did the Intifada begin?
86. What British company belongs to the Airbus consortium?
87. How many people died in the Airbus accident at the Nagoya airport?
88. What is the world population?
89. When was the International Socialist Organization founded?
90. What Nobel Prize was Willy Brandt awarded?
91. Where did the 1996 Olympic Games take place?
92. What country is the painter Piet Mondrian from?
93. What city is the Prado Museum in?
94. How much was paid for "La Femme qui Pleure I", an engraving by the Spanish painter Pablo Picasso?
95. What city is "Broadway" in?
96. What was the real name of the Yugoslavian president "Tito"?
97. What is the capital of Croatia?
98. What is the population of Belgium?
99. What country is the largest consumer of alcohol?
100. What country is Tegucigalpa in?
List of Urdu to English translated questions:

1. Where is Euskirchen?
2. Which festival in got award to Leone-d'Oro?
3. What is IRA's meaning?
4. 
5. 
6. Which country in happen inflation's against Real Plan?
7. What is disease technical of Mad-Cow name?
8. Which country from is Estate agent Schneider?
9. 
10. Which other name from known is EFA project?
11. Which Nobel prize get Solzhenitsin to?
12. 
13. Who is italy of president?
14. Which country's host in happen winter's Olympic-Games?
15. Where happen South Yemen's declaration of seccession?
16. Where happened Kazem Radjavi's death?
17. What is capital of Iran?
18. 
19. Which country from was Communist leader Najibullah?
20. Which office in Rabbani was at in 1992 civil-war in afghanistan happened?
21. 
22. Who was first Finnish-Comissionry of EU?
23. Which country from is Edouard Balladur?
24. Which year happened Ayrton Senna's death?
25. 
26. Who is Nirvana's head and singer?
27. Who is European-Commission's president?
28. When broke Berlin Wall?
29. 
30. Which city in is European Parliament?
31. 
32. How-many European countries are part of G7?
33. What is ETA's meaning?
34. Who is president of Spain?
35. Which date on orbit in kept MIR Space Station?
36. 
37. Which did win tournament of Wimbledon?
38. Which country win Davis cup?
39. Which food in pesticides found babies's?
40. Which team did win NBA championship?
41. Which country did army withdraw from Bosinia?
42. Which did win European Final Four tournament of basketball?
43. Which is Alpes's highest peak?
44. Which year left Russian army from Latvia?
45. Which Russian's president did attend's G7's meetings in Naples?
46.
47. Which two ethnic group are fighting in Ruanda?
48. Which ship in crash happen New-World tanker of oil?
49. Which flag's with sail's tanker of oil Aegean Sea?
50. How-much oil fell tanker of oil Aegean Sea's in accident?
51. How-much happen tanker of oil Aegean Sea's accident's approximately damages?
52. Who gave tanker of oil Aegean Sea's accident's money of damages?
53. Who was tanker of oil Aegean Sea's accident's prosecuting-attorney of trial?
54. What is World Wild Life fund's meaning?
55. How-many country together make are International Whaling Commission?
56. Which country's above get Ozone hole?
57. Where happen International Whaling Commission's yearly meetings?
58. Which year was made world bank?
59. Which year was made International Monetary Fund?
60. How-many country together make IMF?
61. Who is president of Peru?
62. Which company made Nicholas-Brady did?
63. Which French ex-minister went jail Curruption's on charges?
64. Which day resignation gave Agriculture, Fisheries and Food's spanish minister Vicente Albero did?
65. How-many crew's death happened in Submarine-Emeraude?
66. When happen accident of Chernobyl?
67. How-many divided in countries happen USSR?
68. How-many people's are Chernobyl's dieing from disaster of risk?
69. What is world's rate of suicide?
70. How high is Everest?
71. What meaning are PLO's?
72. How many countries do attend are MED 21 in conference?
73. Who is Palestinian National Authority's president?
74. Which day happen start of Intifada?
75. How-many people were died Nagoya's airport's accident of Airbus in?
76. What is world's population?
77. When formed International Socialist Organization?
78. Which Nobel Prize get Willy Brandt to?
79. Where happen 1996's Olympic Games?
80. Which is from country Painter Piet Mondrian?
93. Which is in city Prado Museum?
95. Which is in city Broadway?
96. What is Yugoslavia's president Tito's real name?
97. What is capital of Croatia?
98. What is Belgium's population?
99. Who is biggest in world consumer of alcohol?
100. Which is in country Tegucigalpa?
Urdu-English QA system snippet collection:

1. Where is Euskirchen?
Snippets:
*Bags of Words*: Euskirchen (district) - Wikipedia, the free encyclopedia
Map of North Rhine-westphalia highlighting the district Euskirchen Euskirchen is a Kreis (district) in the south-west of North Rhine-westphalia, Germany

*Component*: Apta: Germany’s ‘Stadtbus’ System Expanding; ‘Best Practices’ Book Another example is Euskirchen, a regional center of 51000 south of Bonn, which tallied a record four million riders on its Stadtbus system last year

*Comp. missing w0*: Euskirchen (district) - Wikipedia, the free encyclopedia
Map of North Rhine-westphalia highlighting the district Euskirchen Euskirchen is a Kreis (district) in the south-west of North Rhine-westphalia, Germany

*Verbs*: Apta: Germany’s ‘Stadtbus’ System Expanding; ‘Best Practices’ Book Another example is Euskirchen, a regional center of 51000 south of Bonn, which tallied a record four million riders on its Stadtbus system last year

2. Which festival in got award to Leone-d'Oro?
*Bags of Words*: Ennio Morricone In 1958 he got a job at the Rai as music assistant, but he resigned on the Leone d'Oro Lifetime Achievement Award at the 52nd Venice Film Festival

3. What is IRA's meaning?
*Bags of Words*: The Ira's international (Irish Republican Army's ties to The Ira's International The Irish Republican Army (Ira) has, still, many friends campaign against "imperialism" (meaning principally the United States)

4.

5.

6. Which country in happen inflation's against Real Plan?

7. What is disease technical of Mad-Cow name?
*Bags of Words*: Enzyme Degrades Mad Cow Disease Prion These transmissible prions – believed to be the cause of bovine spongiform encephalopathy (BSE), the technical name for mad cow disease, as well as the

8. Which country from is Estate agent Schneider?
*Bags of Words*: Victoria County - Indianola Database The agent soon after left the country, and your petitioner did not and has not to this German at E Schneider's was supposed to be T Miller for visits
9. Which other name from known is EFA project?

"Bags of Words": Eurofighter Typhoon: Information From Answers com The project has been named and renamed a number of times since its inception, having been known as Efa (European Fighter Aircraft), Eurofighter.

10. Which Nobel prize get Solzhenitsin to?

"Bags of Words": Famous Russians i think he is the one whose name is Aleksandr Solzhenitsin with nobel prize -1970 literature- i love this quotation from him

11. Which Nobel prize get Solzhenitsin to?

"Bags of Words": Famous Russians i think he is the one whose name is Aleksandr Solzhenitsin with nobel prize -1970 literature- i love this quotation from him

12. Who is italy of president?

"Bags of Words": Bbc News | Europe | Italy president vetoes media bill Carlo Azeglio Ciampi returns to parliament a media law which critics say favours tycoon Pm Silvio Berlusconi

"Component": Sweden - Euro - Worldpress org core of the founding countries—among which we hope to see also the Italy of President [Carlo Azeglio] Ciampi, a great, longtime supporter of Europe

"Verbs": Sweden - Euro - Worldpress org core of the founding countries—among which we hope to see also the Italy of President [Carlo Azeglio] Ciampi, a great, longtime supporter of Europe

13. Who is italy of president?

"Bags of Words": Bbc News | Europe | Italy president vetoes media bill Carlo Azeglio Ciampi returns to parliament a media law which critics say favours tycoon Pm Silvio Berlusconi

"Component": Sweden - Euro - Worldpress org core of the founding countries—among which we hope to see also the Italy of President [Carlo Azeglio] Ciampi, a great, longtime supporter of Europe

"Verbs": Sweden - Euro - Worldpress org core of the founding countries—among which we hope to see also the Italy of President [Carlo Azeglio] Ciampi, a great, longtime supporter of Europe

14. Which country's host in happen winter's Olympic-Games?

"Bags of Words": Lillehammer Olympic Games The first winter Olympics were held at Chamonix, France, in 1924, Thus it is of utmost importance that host countries try to minimize their destructive

15. Where happen South Yemen's declaration of seccession?

16. Where happened Kazem Radjavi's death?

17. What is capital of Iran?

"Bags of Words": Tehran - Welcome to Iran's Capital - Tehran's Official Homepage Welcome To Tehran, Capital of Iran, This tower which represents the city of Tehran is part of Azadi (Freedom) cultural complex which is located in Tehran's

"Component": Tehran - Welcome to Iran's Capital - Tehran's Official Homepage Welcome To Tehran, Capital of Iran, This tower which represents the city of Tehran is part of Azadi (Freedom) cultural complex which is located in Tehran's

"Verbs": Tehran - Welcome to Iran's Capital - Tehran's Official Homepage In the reign of agha Mohammad Khan, the founder of the Qajar dynasty (1776-1925), when Tehran was chosen as the capital of Iran, it enjoyed a boom
19. Which country from was Communist leader Najibullah?

“Bags of Words”: Democratic Republic of Afghanistan - Wikipedia, the free encyclopedia Najibullah then reorganized his government to face the mujahedin alone. A new constitution took effect in November, 1987. The name of the country was

20. Which office in Rabbani was at in 1992 civil-war in Afghanistan happened?

“Bags of Words”: Crisis of Impunity - Afghanistan’s Civil Wars In June 1992 Rabbani became president of Afghanistan, while Hikmatyar continued. The civil war in Afghanistan can be characterized as a non-international

22. Who was first Finnish “Component” omissionry of EU?

23. Which country from is Edouard Balladur?

“Bags of Words”: Edouard Balladur -- Encyclopædia Britannica Edouard Balladur French neo-Gaullist politician, prime minister of France from 1993 to who fought many battles for his country — on the Western front, “Comp. missing w01”: The Parliamentary system of France The current Prime Minister is Edouard Balladur He was appointed at the end of March 1993. To appreciate to the full the links between the President of the “Verbs”: The Parliamentary system of France The current Prime Minister is Edouard Balladur He was appointed at the end of March 1993. To appreciate to the full the links between the President of the

24. Which year happened Ayrton Senna’s death?

“Bags of Words”: The death of Ayrton Senna On May 1st 1994 during the San Marino Grand Prix, the unthinkable happened and 34 year old Ayrton Senna would never race again “Comp. missing w01”: Buy used car: Renault cars for sale, select model to browse Follow the Williams team from the thrill of winning the 1992 and 1993 World Championships, through the tragedy of Ayrton Senna’s death in 1994, “Verbs”: National Geographic Channel In 1994, millions of people witnessed Ayrton Senna’s death, yet the tragedy is still unresolved. Find out what caused this catastrophic event

26. Who is Nirvana’s head and singer?

“Bags of Words”: Albums of the artist singer songer: Nirvana - album albums songs Nirvana is primarily known for their head banging revolutionary grunge music Nirvana (made up of guitarist/singer, songwriter Kurt Cobain, bassist Krist
27. Who is European Commission's president?

“Bags of Words”:  Charlemagne | The Angela and Jose show | Economist.com Mismatched as they may seem, Angela Merkel, the German chancellor, and Jose Manuel Barroso, the European Commission's president, have struck up the most

“Component”:  Charlemagne | The Angela and Jose show | Economist.com Mismatched as they may seem, Angela Merkel, the German chancellor, and Jose Manuel Barroso, the European Commission's president, have struck up the most

“Verbs”:  Charlemagne | The Angela and Jose show | Economist.com Mismatched as they may seem, Angela Merkel, the German chancellor, and Jose Manuel Barroso, the European Commission's president, have struck up the most

28. When broke Berlin Wall?

“Bags of Words”:  Berlin Wall 1989 On 10th November 1989, I arrived in Berlin, the day after The Wall was opened They all loved Berlin Some time after midnight the party broke up

“Component”:  The Fall of the Berlin Wall 1989 The Fall of the Berlin Wall, which will always be used as a symbol for the end of the Cold War, made the "West" available in the middle of the "East",

“Verbs”:  Berlin Wall Online - Chronicle of the Berlin Wall history includes A history of Berlin Wall in text and photographs and what happened to the Wall after its fall in 1989

29.

30. Which city in is European Parliament?

“Bags of Words”:  European Parliament - Committee on Budgetary Control - Committee on Budgets the City of Strasbourg may have charged the European Parliament inflated

“Verbs”:  European Parliament - Committee on Budgetary Control - Committee on Budgets the City of Strasbourg may have charged the European Parliament inflated

31.

32. How many European countries are part of G7?

33. What is ETA's meaning?

“Bags of Words”:  Reference com/Encyclopedia/ETA For other meanings of Eta, see Eta Euskadi Ta Askatasuna, or Eta (IPA: ??ta), In its early years, Eta's activity seems to have consisted mostly of

34. Who is president of Spain?

“Component”:  Jose Luis Rodriguez Zapatero President of Spain or Prime Minister? In theory, the president of Spain is not a president Keywords Jose Luis Rodriguez Zapatero, President of Spain
“Verbs”: President George W Bush welcomes the President of Spain Jose Maria Aznar to the Oval Office Nov 28 "Recently, Spain has arrested al Qaeda members and has"

35. Which date on orbit in kept MIR Space Station?

“Bags of Words”: A history of space stations The US space agency Nasa dates the first proposal for a manned station to 1869, and in 1986 launched the first module of the Mir space station

36.

37. Which did win tournament of Wimbledon?

38. Which country win Davis cup?

“Bags of Words”: Usatoday.com - Ancic's win clinches first Davis Cup for Croatia Ancic's win clinches first Davis Cup for Croatia Bratislava, Slovakia (AP) — Croatia won its first Davis Cup title Sunday, with Mario Ancic beating Michal

“Comp. missing w0”: Cpc Bbs Archive - croatia win davis cup World of Sport - come on my boyslux lux lux lux lux

“Comp. missing w01”: Bbc Sport | Tennis | Croatia win first Davis Cup title Mario Ancic beats Slovakia's Michal Mertinak to lead Croatia to their first Davis Cup win

“Verbs”: Independent, The (London): Tennis: Ancic nerve holds to win Davis Cup Full text of the article, "Tennis: Ancic nerve holds to win Davis Cup for Croatia" from Independent, The (London), a publication in the field of News

39. Which food in pesticides found babies's?

40. Which team did win NBA championship?

“Bags of Words”: Nba.com: Pistons One Win From Claiming Nba Championship Detroit 88, La Lakers 80 Pistons One Win From Claiming Nba Championship &quot;So there's no need even thinking about what no team did&quot;

“Comp. missing w01”: Historylink Essay: Supersonics win Nba Championship on June 1, 1979 Historylink is a historical database and Website devoted to chronicling the history of Seattle and Washington State history

“Verbs”: Scholastic News: Spurs Win Nba Championship Spurs Win Nba Championship Spurs Win Nba Championship By Ezra Billinkoff Monday, June 27—The San Antonio Spurs overcame early difficulties to defeat

41. Which country did army withdraw from Bosinia?

42. Which did win European Final Four tournament of basketball?

43. Which is Alpes's highest peak?
44. Which year left Russian army from Latvia

“Bags of Words”: Skrunda Lpar - Soviet Bmd Ilgonis Upmalis stated on 7 February that two Russian army units -- medical Russian forces left Estonia and Latvia on 31 August 1994, as announced by “Comp. missing w01”: Minelres: Rfe/rli: Russian minorities outside Russia authorities that Ms Slivenko’s husband had to leave Latvia in accordance with the 1994 treaty on the withdrawal of the Russian army from Latvia

“Verbs”: Virac They were taken over by the Latvian Academy of Sciences after withdrawal of the Russian army from Latvia in 1994 On the basis of these antennas and

45. Which Russian’s president did attend’s G7’s meetings in Naples?

46.

47. Which two ethnic group are fighting in Ruanda?

“Bags of Words”: Rwanda - Global Policy Forum - Un Security Council Torn by ethnic conflict between the Tutsis and the Hutus, Rwanda experienced of National Unity, seeking reconciliation between the two ethnic groups

48. Which ship in crash happen New-World tanker of oil?

49.

50. Which flag’s with sail’s tanker of oil Aegean Sea?

“Bags of Words”: The Aegean Mystery She picked it up, unfolded it, and saw that it was a travel brochure for a resort located on one of the many Greek isles which dotted the Aegean Sea

51. How-much oil fell tanker of oil Aegean Sea's in accident?

“Bags of Words”: Environmental disasters The people that have lived in the Chernobyl area during the accident suffer from various 1992 Greek tanker Aegean Sea spills 70,000 ton oil near Galicia

52. How-much happen tanker of oil Aegean Sea's accident's approximately damages?

53. Who gave tanker of oil Aegean Sea's accident's money of damages?

54. Who was tanker of oil Aegean Sea's accident's prosecuting “Bags of Words” attorney of trial?

55. What is World Wild Life fund’s meaning?

“Bags of Words”: Wwf - Species Home Carnivore is derived from two Latin words; caro meaning flesh and vorare meaning The word omnivore comes from two Latin words; omni meaning &apos;all&apos; and
56. How many countries together make up the International Whaling Commission?

57. Which country's above gets an Ozone hole?

*Bags of Words*: Tve's Earth Report: Sky Pirates Scientists looked at the ozone layer above the Antarctic but there wasn't a hole there yet. The scientific establishment accepted the possibility without... (text continues)

58. Where happen International Whaling Commission's yearly meetings?

*Bags of Words*: Daily Excelsior World: The International Whaling Commission (IWC) is deeply divided, High on the agenda at the IWC's yearly meeting, being held in Berlin from June 16 to 19.

59. Which year was made world bank?

60. Which year was made International Monetary Fund?

*Verbs*: Untitled made International Monetary Fund restructuring efforts difficult, and Yemen remains one of the poorest Arab countries; the per capita Gdp is US$800 a year.

61. How many country together make IMF?

62. Who is president of Peru?

*Bags of Words*: Toledo Elected President of Peru: The First Indian American President of Peru Promises Social Change. The 55-year-old Toledo became the first president of Indian descent in Peru.

*Component*: List of Presidents of Peru: Information From Answers.com (used by Bolivar) and "Supreme Protector" (by Santa Cruz). The following table contains a list of the individuals who have served as president of Peru.

*Verbs*: List of Presidents of Peru - Wikipedia, the free encyclopedia. List of Presidents of Peru From Wikipedia, the free encyclopedia (Redirected from President of Peru).

Jump to: navigation, search

63. Which made in conference were WB and IMF?

*Bags of Words*: High-level conference on "trade, growth and poverty" (Bretton Official transcript of the conference to be made available from Dfid Trade-finance coherence, Transparency-government, Transparency-Wb/Imf, Transport.

64. What is Vietnam's currency?

*Bags of Words*: Adventures In Indo-China! Vietnam's currency, the oddly named Dong was 12250 to $US1 with a street "Black Market" rate of about 10% more this. (Bretton Official transcript of the conference to be made available from Dfid Trade-finance coherence, Transparency-government, Transparency-Wb/Imf, Transport.

"Component": Vietnam introduces new 500000 Dong bank note: In the last few days, "reactionary" groups have spread reports on the...
Internet about a change in Vietnam’s currency, Thuy said, adding: “They called on verbs.” Media Lens Message Board: Vietnam’s class war. Before the strikes, the minimum wage of about $40 a month had remained unchanged for six years, while Vietnam’s currency, the dong, lost almost 15 percent.

65. Who is writer of Brady-Plan?

“Bags of Words”: Building a system for answering Dutch person questions. For example “writer” as a query term yields better results than “wrote.”

“Comp. missing w01”: National Review: No banking on Brady - Nicholas F Brady, Third. Like the Baker Plan before it, the Brady Plan has fallen victim to the hard realities of skeptical bankers and wayward debtor nations.


66. Which was in office Brady in US-Treasury?


Nicholas F Brady, former US Treasury Secretary whose department was.

67. Which company made Nicholas-Brady did?

68.

69. Which French ex-minister went jail for Corruption's on charges?

70. Which day resignation gave Agriculture, Fisheries and Food’s Spanish minister Vicente Albero did?

71.

72.

73.

74.

75.

76. How many crew's death happened in Submarine-Emeraude?

“Bags of Words”: Wise Nc; Ten Die In Nuclear Sub Accident. The accident happened as the vessel with a crew of 66, was deep under-water on a routine exercise with other navy ships 45 miles off the port of Toulon.

77. When happen accident of Chernobyl?
“Bags of Words”:

Cbbc Newsround | Guides | Tech | Chernobyl | What happened at The World's worst nuclear accident happened at Chernobyl on 26 April 1986 A safety test at a nuclear power plant in Ukraine went disastrously wrong and big

“Comp. missing w0”:

After the nuclear accident of Chernobyl in 1986 the need for a system that supported the decision makers (DMs) in taking appropriate countermeasures was

“Comp. missing w01”:

Discovery Channel - Battle of Chernobyl Battle of Chernobyl The accident that occurred at Chernobyl on 26th April 1986 was the most disastrous reactor malfunction in the history of nuclear power

“Verbs”:

After the nuclear accident of Chernobyl in 1986 the need for a system that supported the decision makers (DMs) in taking appropriate countermeasures was

78. How many divided in countries happen USSR?

79. How many people's are Chernobyl's dieing from disaster of risk?

80. What is world's rate of suicide?

“Bags of Words”:

Suicide around the world every 40 seconds - Mental Health - Msnbc com A suicide takes place somewhere around the world every 40 seconds, or nearly one million a year, and the rate looks set to surge over the next two decades

“Comp. missing w0”:

The Merck Manual, Sec 15, Ch 190, Suicidal Behavior The rate among physicians is high largely because of women physicians, whose annual rate of suicide is 4 times that of a matched general population

“Verbs”:

Suicide, Facts - Ncipc The overall rate of suicide among youth has declined slowly since 1992 American Indian and Alaskan Natives have the highest rate of suicide in the 15 to

81. How high is Everest?

“Bags of Words”:

#39;Surprise!#39; says Everest climber given up for dead - Yahoo! News An aerial view of the world's tallest 8848 metre (29028 feet) high Mt Everest &middot; Afp/file Photo: An aerial view of the world's tallest 8848 metre (29028

“Component”:

Human Edge Tech - Expedition communication and technology

To Everest fans, the immediate question that comes up is, How high is Everest? There is much debate over this, between the 8848m traditionalists,

“Comp. missing w0”:

mequz How tall is Everest? a) 28024 ft above sea level b) 24028 ft above sea level c) 52890 ft above sea level 3 Where is Everest? a) Between Tibet and Nepal

“Comp. missing w01”:

Everest News, Where Everest Climbers Come for News News and expedition dispatches from Everest Discussion forums Information on the 8000m peaks and
Appendix II

“Verbs”: mequz How tall is Everest ? a) 28024 ft above sea level b) 24028 ft above sea level c) 52890 ft above sea level 3 Where is Everest ? a) Between Tibet and Nepal

82. What meaning are PLO’s?
“Bags of Words”: Palestine Liberation Organization: Definition and Much More From Meaning , Category & Orthodox Liberal Party; International- guatemalan After the Palestinian Declaration of Independence the Plo’s representation was
“Comp. missing w0” Strategic Issues The classic examples are Plo’s Yasser Arafat and Israel’s Menachem Begin) But whether this would have spilled over into a successful national liberation
“Comp. missing w01” The Plo’s struggle for Palestinian autonomy (from Palestine) Although generally recognized as the—v: The Plo’s struggle for Palestinian autonomy (from Palestine) Although generally recognized as the symbol of the Palestinian national movement, symbol of the Palestinian national movement,

83. How many countries do attend are MED 21 in conference?

84. Who is Palestinian National Authority’s president?
“Bags of Words”: Yasser Arafat Winner of the 1994 Nobel Prize in Peace President of the Palestinian National Authority Background 1929-2004 Residence: Palestine Book Store Books about Yasser Arafat Featured Internet Links
“Verbs”: Jmcc / Public Opinion Polls Q9 Do you support or oppose the method that the President Yasser Arafat administers his work as the National Authority’s President?

85. Which day happen start of Intifada?
“Bags of Words”: Palestine-un org the start of the intifada , which gloriously began on that day in 1987 It is our strong hope that this will happen through the current peace process
“Comp. missing w01” 1986 1986 Start of intifada Hypercard introduced Van Gogh whereas sold for $49 million World population 5 billion Iran”Component”ontra Scandal
“Verbs”: 1986 1986 Start of intifada Hypercard introduced Van Gogh whereas sold for $49 million World population 5 billion Iran”Component”ontra Scandal

86.

87. How many people were died Nagoya’s airport’s accident of Airbus in?

88. What is world’s population?
“Bags of Words”: World’s 50 Most Populous Countries: 2005 Population Hits Six Billion - World Population Reaches New Milestone by Borgna Brunner Related Links Census 2000 Most Populous
“Component”: Beyond Six Billion: Beyond Six Billion: Forecasting the World’s Population accessible introduction to the state of the world’s population and likely future trends

89. When formed International Socialist Organization?
“Bags of Words”: Socialist Appeal: Facts and details from Encyclopedia Topic International Socialist Organization International Socialist Organization quick summary: (formed in 1953 by groups who disagreed with the course “Component”: Message from &lt;i&gt;Socialist Worker&lt;/i&gt; Sw first appeared in April 1977 as the newspaper of the newly formed International Socialist Organization We’ve come out regularly ever since--first as a

90. Which Nobel Prize get Willy Brandt to?
“Bags of Words”: Willy Brandt - Nobel Lecture The Nobel Peace Prize for 1971 has been awarded to a man still active in to preserve the purity of their doctrine, or to get the better of other

91. Where happen 1996’s Olympic Games?
“Bags of Words”: Newyorkgames org: June 05, 2005 - June 11, 2005 Archives It could be built to Olympic standards, then altered after the Games to baseball standards The precedent is 1996’s Olympic stadium where the Atlanta Braves
“Comp. missing w0”: :: If an event isn’t on Atlantahapps Com, it’s not happening :: Atlanta has been host to a number of events that has been receiving the masses from all over-- 1996’s Olympic Games, in my opinion, was the Atl’s jumpoff to
“Comp. missing w01”: The Ancient Olympics Members of the Perseus Project created this exhibit on the ancient Olympics in 1996, as a tribute to the Centennial Olympic Games held in Atlanta, Georgia
“Verbs”: :: If an event isn’t on Atlantahapps Com, it’s not happening :: Atlanta has been host to a number of events that has been receiving the masses from all over-- 1996’s Olympic Games, in my opinion, was the Atl’s jumpoff to

92. Which is from country Painter Piet Mondrian?
“Bags of Words”: Piet Mondrian - Mr Boogie-woogie Man - Dvds &amp; Vhs - Moviemail Uk The Dutch painter Piet Mondrian (1872–1944), who began his career painting delicate landscapes, became one of Country, Uk, Language, English, Year, 1995

93. Which is in city Prado Museum?
“Bags of Words”: Skyscraper City - Prado Museum -Madrid News, photo’s and discussions on skyscrapers, skylines, cities, architecture and urbanity
“Comp. missing w01”: Skyscraper City - Prado Museum -Madrid News, photo’s and discussions on skyscrapers, skylines, cities, architecture and urbanity
“Verbs”: Skyscraper City - Prado Museum -Madrid News, photo’s and discussions on skyscrapers, skylines, cities, architecture and urbanity
94.

95. Which is in city Broadway?
   “Bags of Words”: New York City Broadway Plays Broadway Plays Broadway Readio Com New York City, Broadway, Plays, Theatre, Theater, Broadway Plays, Broadway Shows, Broadway Theater, Broadway Theatre, Listings, Broadway, Reviews,
   “Comp. missing w01”: New York City Broadway Plays Broadway Plays Broadway Readio Com New York City, Broadway, Plays, Theatre, Theater, Broadway Plays, Broadway Shows, Broadway Theater, Broadway Theatre, Listings, Broadway, Reviews,
   “Verbs”: New York City Broadway Plays Broadway Plays Broadway Readio Com New York City, Broadway, Plays, Theatre, Theater, Broadway Plays, Broadway Shows, Broadway Theater, Broadway Theatre, Listings, Broadway, Reviews,

96. What is Yugoslavia's president Tito's real name?
   “Bags of Words”: Josip Broz Tito: Definition and Much More From Answers com Tito, whose real name was Josip Broz, led the resistance in Yugoslavia to the On April 17, representatives of Yugoslavia's various regions signed an
   “Comp. missing w01”: Tito and his People by Howard Fast (2) Tito's real name is Josip Broz He was born about 65 years ago in the village of Kranak in the hills of Croatia
   His father was a poor peasant who,
   “Verbs”: Tito and his People by Howard Fast (2) Tito's real name is Josip Broz He was born about 65 years ago in the village of Kranak in the hills of Croatia
   His father was a poor peasant who,

97. What is capital of Croatia?
   “Bags of Words”: Zagreb - capital of Croatia Virtual tour to the capital of Croatia Includes list of city streets, squares, parks and vistas
   “Component”: Zagreb - capital of Croatia Virtual tour to the capital of Croatia Includes list of city streets, squares, parks and vistas
   “Comp. missing w0”: Croatia Part of Croatia is a barren, rocky region lying in the Dinaric Alps The Zagorje region north of the capital, Zagreb, is a land of rolling hills,
   “Comp. missing w01”: Us Embassy Zagreb, Croatia The Home page of The United States Embassy in Zagreb, Croatia, Us Embassy, Zagreb, Republic of Croatia, provides information on the Us, with an emphasis on
   “Verbs”: Croatia Part of Croatia is a barren, rocky region lying in the Dinaric Alps The Zagorje region north of the capital, Zagreb, is a land of rolling hills,

98. What is Belgium's population?
   “Bags of Words”: Belgium Population The population of Belgium in 2003 was estimated by the United Nations at 10318000, Belgium's population has distinctive language and ethnic divisions
Appendix II

“Component”: Belgian Internet Users: Broadband or Offline
Belgium’s population is 10.36 million. The Ispa cites a low household PC penetration for the relatively low number of users. Initiatives from the Federal

“Verbs”: Agri-food Past, Present and Future Report - Belgium In 2004, Belgium’s population was approximately 10.3 million. About 65.7% of Belgians are between the ages of 15-64, 17.4% are 65 years or older and

99. Who is biggest in world consumer of alcohol?

“Bags of Words”: Companies news – Consumer industries – Ft.com. But the global nature of the football World Cup led to the horrifying prospect of

only foreign alcohol being available before German lobbying started

100. Which is in country Tegucigalpa?

“Bags of Words”: Country of Honduras - Honduras Com. Honduras is a vibrant country, brimming with clear turquoise waters. Click for La Ceiba, Honduras Forecast; Click for Tegucigalpa, Honduras Forecast

“Comp. missing w0” Opic Links: Honduras Foreign Service Post Report &middot; Diplomacy: State Dept Foreign Service Post Report 82 Diplomacy: Us Embassies/consulates in Country - Tegucigalpa

“Comp. missing w01” Wec Graduate Students (PS) Jabiru Stork Alejandro J. Paredes Borjas, Masters Student Hometown/ Country: Tegucigalpa, Honduras Previous Education: Bachelors of Science in Biology,

Books:

Information Retrieval
C.J. Van RIJSBERGEN- Second Edition

Modern Information Retrieval
Ricardo Baeza Yates
Berthier Ribeiro Neto

Foundation of Statistical Natural Language Processing
Christopher D. Manning
Hinrich Schutze

Natural Language Question Answering System
Boris Galitsky

Links:
http://en.wikipedia.org
http://computer.howstuffworks.com

Published papers:

“What they are, How they Work and Practical suggestions for getting the must out of them”
By: Bruce Grossan

“Web-Based Question Answering: A Decision-making Perspective”
By: David Azari
Eric Horvitz
Susan Duman
Eric Brill

“QA on the Web: A Preliminary Study for Spanish Language”
By: Alejandro Del-Castillo-Escobedo
Manuel Montes-y-Gómez
Luis Villaseñor-Pineda

“Cross-language Question Answering: The Key Role of Translation”
By: S. Larosa1, J. Peñarrubia
P. Rosso
M. Montes-y-Gomez