www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 7 July 2017, Page No. 22020-22026
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i7.20

Questioning Answering Mechanism Using Natural Language Processing

Shelza[1,] *Lalit*[2], *Priti Aggarwal*[3], *Geetanjali Sharma*[4]

[#1-4]Department of Computer Science and Engineering, [#1-4]Beant College of Engineering and Technology,
Gurdaspur, Punjab, India
[1]Shelzamahajan13@yahoo.in
[2]Lalit2630@gmail.com
[2]Apriti30@yahoo.com
[4]Sharmageetanjali250989@gmail.com

Abstract -- A Question Answering (Qa) System Provides Direct Answers To User Questions By Consulting Its Knowledge Base. This Paper Presents Basic Architecture Of Qa System That Is Based On Information Retrieval System, Modules And Its Various Types. Along With That This Work Also Describes Different Architectures For Various Types Of Question Answering Mechanism Such As Open Domain Qas, Closed Domain Qas, Rule Based And Web Based Qas. The Comparison Among Different Architectures Of Question Answering Mechanism Is Also Drawn.

Keywords*:* Closed domain QAS, Information Retrieval system, Open Domain QAS, Rule based QAS, Web based QAS.

1. INTRODUCTION

Natural language processing (NLP) [1] is a field related to the area of computer science, artificial intelligence, linguistics and human computer interactions by means of which computational mechanisms are investigated and formulated. These mechanisms allow the development of systems that is capable of understanding the knowledge expressed in texts of a given language. Natural Language Processing is a theoretically range of computational techniques for representing and analyzing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

Natural language processing provides both theory and implementations for a range of applications. NLP has been centered on the resource development that provides multiple syntactic and semantic levels. The most common applications utilizing NLP include the following: systems of machine translation, information retrieval (IR), information extraction, question answering (QA), recognition of entities, classification and filtrate of documents, generation of summaries, etc.

Question Answering (QA)[2] is an area of computer science that are designed based on the technique from information retrieval and natural language processing (NLP), which are related to developed systems that automatically answer questions posed by humans in a natural language rather than the keyword based retrieval mechanism. Information Retrieval deals with the representation, storage, organization and access to information items. Informational retrieval is a system that search for precise answers to specific information that is useful for large collection of documents. The basic architecture of Question- Answering systems is based on an information retrieval system that considers the word of the questions as queries to retrieve the appropriate answers.

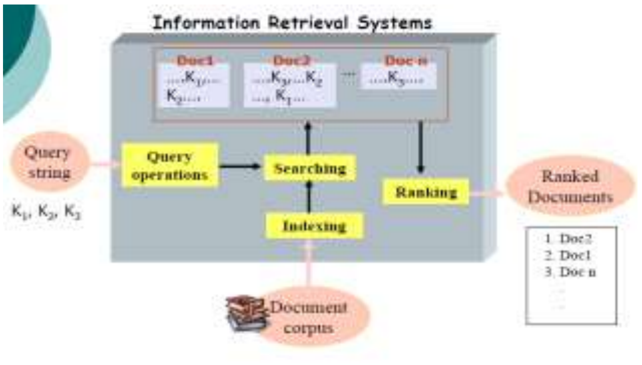


Figure 1: Information Retrieval System (Adapted from [2])

1.1 Modules Of QAS

QAS consists of three modules which plays a vital role:

1. Question Classification Module
2. Information Retrieval Module
3. Answer Extraction Module

1.1.1 Question Classification Module:

Question Classification module [3] plays a primary role in QAS to classify the question based on its type. Questions can be factoid, long answers, definition, how, wh-type questions, semantically-complex and multi-lingual questions, hypo-sort of questions etc. If this classification goes wrong, it will affect the working of the other module. Once this classification is done it extracts keywords and derives expected answer types, and reformulates a

DOI: 10.18535/ijecs/v6i7.20

question into its semantically equivalent multiple questions. Reformulation of a query into similar meaning queries is also known as query expansion and it boosts up the recall of the information retrieval system.

1.1.2 Information Retrieval Module:

Information retrieval module [3] firstly selects paragraphs that are considered relevant to input questions. In order to narrow the search area, filtering of paragraph will be done. One can also check quality status for these documents so that it can be easy to check whether the selected paragraphs or documents contain correct answer. For ordering the paragraphs one can use radix sort that will give the appropriate paragraph where the exact answers for the questions are available. Then move to answer extraction module for correct answers. If no correct answers are present in a document, no further processing can be done.

1.1.3 Answer Extraction Module:

Answer Extraction is the last module [3] in QAS, which is responsible for identifying, extracting and validating answers from the set of ordered paragraphs passed to it from the information retrieval module.

1.2 Types of QA system

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. There are two type of questioning answering system:

1.2.1 Closed Domain Question:

Closed Domain system [4] 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 frequently formalized in ontologies. Closed-domain might refer to a situation where only limited types of questions are accepted, such as questions asking for descriptive rather than procedural information.

1.2.2 Open Domain Question:

Open Domain systems [4] rely on word knowledge and general ontologies and deals with questions about nearly anything. These systems usually have much more available data from which to extract the answer.

2. DIFFERENT ARCHITECTURES FOR VARIOUS TYPES OF QUESTION ANSWERING SYSTEM

There are four different architectures of Question answering system:

1. Closed domain QAS
2. Open domain QAS
3. Web based QAS
4. Rule based QAS

2.1 An automatic Answering System with template matching using Closed Domain QAS:

To answer natural language questions using computer is an interesting and challenging problem. These problems are handled under two categories: a) open domain system and b) closed domain system. This work proposed a system that attempt to solve close domain problem. In close domain system, answer to question has to be stored in a database by expert i.e. They are not available in public domain. They cannot be searched using any search engine. The first step is to understand natural language questions so that solution could be matched to the respective answer in the database. For this template matching technique [4] is used to perform matching, in addition using this technique, one can overcome mismatches that might occur due to spelling mistakes.

The system architecture for questioning answering is divided into three modules:

1. Pre Processing module
2. Question Template matching module
3. Answering Module

2.1.1 Preprocessing module*:*

This module mainly consists of three steps:

i. Converting SMS abbreviations into general English word
ii. Removing Stop words. Stop words are the word that adds no effect to the meaning of sentence even if they are removed.
iii. Removing Vowels.

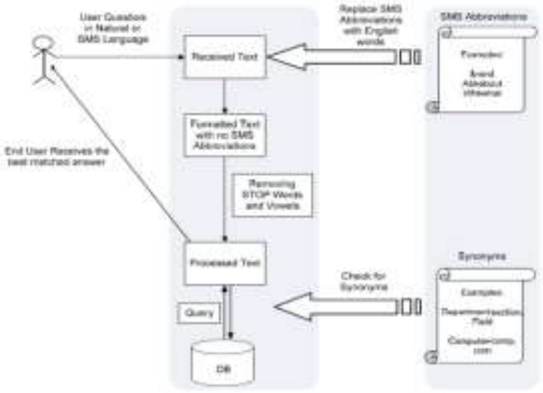

User Question in Natural or SMS Language
Received Text
Replace SMS Abbreviations with English words
SMS Abbreviations
Formatted Text with no SMS Abbreviations
End User Receives the best matched answer
Removing STOP Words and Vowels
Processed Text
Synonyms
Check for Synonyms
Query
Examples
Department-section Field
Computer-comp. com
DB

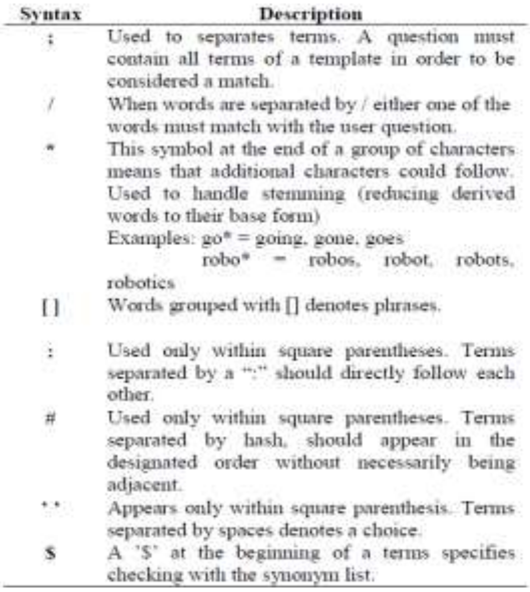
Figure 2: System architecture of Question answer mechanism [4]

The system is designed such that processing of text can be done with both natural and SMS languages. So, before processing user questions it is necessary to convert abbreviations of SMS into general English words. This is done by referring to pre-stored frequently used SMS abbreviations. Then stop words are removed that increase the effectiveness and response of system by saving time and disk space. Examples of stop words are the, a, and, etc .Last step of this module is to remove vowels from text to handle spelling mistake. This process is called disemvoweling [5]. Disemvoweling is also done in our templates as a means of accounting for spelling mistakes in user queries and for easy matching of the template.

2.1.2 Question Template matching module:

The pre-processed text is matched against each and every Pre stored template until it finds the best matched template with the received text. For this, templates are created according to a specific syntax and shown in table 1:

TABLE 1
Syntax Used for Template [4]



Syntax	Description
;	Used to separates terms. A question must contain all terms of a template in order to be considered a match.
/	When words are separated by / either one of the words must match with the user question.
*	This symbol at the end of a group of characters means that additional characters could follow. Used to handle stemming (reducing derived words to their base form) Examples: go* = going, gone, goes robo* = robos, robot, robots, robotics
[ ]	Words grouped with [] denotes phrases.
:	Used only within square parentheses. Terms separated by a ":" should directly follow each other.
#	Used only within square parentheses. Terms separated by hash, should appear in the designated order without necessarily being adjacent.
* *	Appears only within square parenthesis. Terms separated by spaces denotes a choice.
$	A '$' at the beginning of a terms specifies checking with the synonym list.

The syntax of template is defined as a single template could match many different varieties of questions. Using this syntax, templates can be constructed. In this module words are considered to have synonym and referred in synonym list or file. It is necessary to list synonym for each term, since users use queries using different terminology rather than the person who produces the FAQ. This list contains phrases (nested within each other) have same meaning as a single word. The synonym file can be altered or modified from a standard database WordNet [6] according to relevant domain. The main purpose of this system is to identify the template that matches the question we have received from user.

2.1.3 Answering Module:

Each and every template that represents a question is pre stored in a database with its relevant answers. When the best template is matched and found, the corresponding answer will be returned to the user**.**

2.1.3.1 An Open Domain Question Answering System:

The aim of open domain QAS is to respond questions of user. In this case, the reply is a short texts rather than a lengthy list of relevant documents. This type of systems makes use of multiple techniques from computational linguistics, information retrieval and knowledge representation for searching answers. The most important challenge of an open domain system is its database. The efficiency of any system depends on how well the database is arranged and maintained. Especially for open domain QAS as it aims to answer merely for everything.

For example: INSUN05QA [7] is an open domain question answering system and composed of 4 parts as shown in below figure:

1. Preprocessing
2. Question Analysis
3. Retrieval
4. Answer Extraction

Preprocessing for documents and questions: The preprocessing module is developed to accomplish the preprocess task for both documents and questions.
According to documents, the first step is sentence boundary detection and then writes the text of documents in the form of one sentence a line. After that tokenization, part of speech, name entity reorganization, stemming, and co-reference resolution, are applied orderly. According to questions, detection of sentence boundary is not used and rest process is same as documents. GATE is a bridge between question and document preprocessing, applied to accomplish name entity reorganization.

Question Analysis: The question analysis module has two functions: key generation and answer type prediction. To realize these two functions external tools are used i.e. WordNet and Minipar [8].

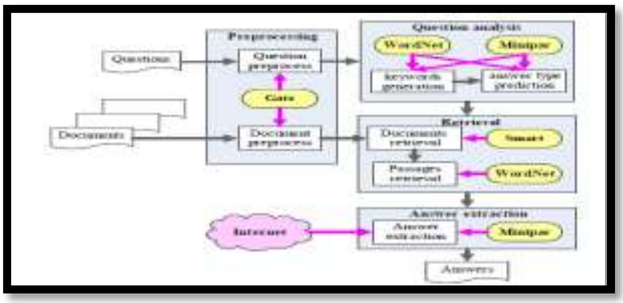


Figure 4: Architecture of INSUN05QA[7]

Minipar is used to analyze the sentence structures of questions, and based on their analysis constituent information's are extracted from questions.

By using the WordNet, a dictionary of synonymous words is constructed, and with the help of this dictionary, the constituents extracted from questions are explored so as to accomplish keyword generation.
In the answer type prediction step, a rule-based algorithm is adopted to classify the answer type of inputQuestions.

Retrieval: Retrieval module is composed of documents retrieval and passages retrieval. Documents are related to each questions are extracted from corpus by using document retrieval and related passages are returned by passages retrieval. Document retrieval module is built based on SMART [9] retrieval system.

Answer Extraction: Based on the results of above steps, the answer extraction module returns the exact answers and the respective documents numbers with the help of Web retrieval.

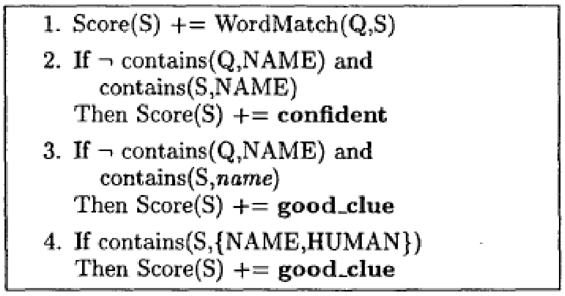
Answer extraction is implemented by inspecting the candidate passages which is full of information including POS and NE tags. If the named entity in the candidate passages corresponds to an expected answer type, the entity will be picked out as a candidate answer. When multiple candidate answers exist, the selection of the best candidate answer is performed with an answer ranking scheme that relies on heuristics method.

2.1.3.2 A Rule Based Question Answering System for reading Comprehension Tests:

Rule based QAS is one of the most important and efficient system. Its basic application is comprehension reading.

For example: Quarc (Question Answering For Reading Comprehension) [*10]* is a rule based question answering system that read a story and look for evidence that sentence contain best answers to questions using lexical and semantic heuristics. Each type of WH questions look for different types of answers, so Quarc uses separate set of rules for each question types e.g. why, where, when, who, what.

First of all Quarc parses all of the questions and sentences in the comprehension (story) using parser technique and syntactic analysis is optional. Then Quarc uses NLP techniques like morphological parsing, POS tagging, semantic entities etc. The parser recognizes mainly two types of semantic entities: proper_nouns and names. Proper_nouns is defined as noun phrase in which all words are capitalized. Names are defined as proper_nouns that contains at least one human word. The rules are applied to each sentence as well as title of story except that title is not considered for Why questions. Then each rule awards a certain number of points to each sentence. The rules like dateline (for WHEN & WHERE type questions), wordMatch function (count the number of words that appear in both question and sentence) etc can be applied on the sentences as per required. A rules can assign four possible point values: clue (+3), good_clue (+4), confident (+6) and slam_dunk(+20).These point values were based on our intuitions and worked well but they are not well justified. The main purpose of these values is to assess relative importance of each clue. **For example:**



```
1. Score(S) += WordMatch(Q,S)
2. If ¬ contains(Q,NAME) and
      contains(S,NAME)
   Then Score(S) += confident
3. If ¬ contains(Q,NAME) and
      contains(S,name)
   Then Score(S) += good_clue
4. If contains(S,{NAME,HUMAN})
   Then Score(S) += good_clue
```

Figure 4: WHO RULES[10]

WHO rules uses three general heuristics as well as wordMatch function (Rule #1). If the question Q does not contain any name, the rule #2 and rule #3 assume that question is looking for a name. Rule #2 rewards sentences that contains a recognized NAME , Rule #3 contains the word ”name” and rule #4 awards points to all sentences that contain either a name or reference to humans. In the event of a tie, a WHY question chooses the sentence that appears latest in the story and all other question types chooses the sentence that appears earliest in the story. If no sentence receive a positive score, then when and where question return the dateline by default.

After all of the rules have been applied to every sentence in the story, the sentence (or dateline) that obtains highest score is returned as best answer.

2.1.3.3 A Web Based Question Answering System:

The Web is apparently an ideal source of answers to a large variety of questions, due to the tremendous amount of information is available online. The most important property of any web based QAS is “snippet-tolerant” that allows it to provide correct responses to the QAS while searching through any search engines like Google, yahoo.

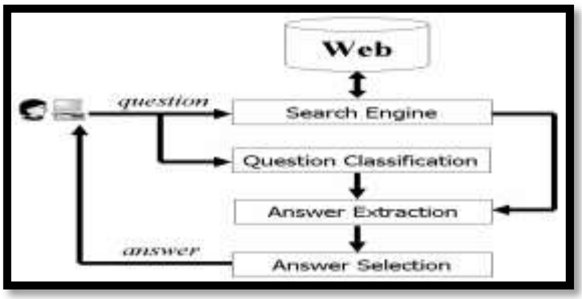


Figure 5: System Architecture of LAMP([11])

For Example: LAMP [11] is a web based question answering system and publically accessible. A characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google.

LAMP system mainly consists of 4 modules as shown in above figure:

Search Engine: The system submits user natural language questions to any search engine like yahoo, Google and grabs its top 100 search results. Each search result usually contains title, URL and some strings of the related web document and that is called “snippets”. The system take advantage of these snippets in the search results because it is time consuming to download and analyze original web documents.

Question Classification: In order to answer the questions correctly, one need to understand what type of information the question asks for e.g. “Who was the first American in space?” asks for a person name. The system may use a Support Vector Machine (SVM) [12] to classify the questions

Answer Extraction: After the identification of question type has been done, then the system extracts all information from the snippets as plausible answers using HMM based entity and some heuristics rule.

Answer Selection: For each plausible answer, the system constructs a snippet cluster which is composed of all the snippets containing that answer. Moreover, the snippet clusters of different answers referring to the same entity should be merged into one. A vector space model in an information retrieval area is a kind of model which can be used to classify the candidate answers. After using the standard Vector Space Model it has been observed that the count for correct answer to the question is usually greater than the incorrect ones on the search results of that question and finally the evaluation of the final answer will be done and returns to the system.

3. *Comparison Between Different Architectures Of Question Answering System*

The comparison between different architectures of Question answering system

TABLE 2
Comparison between different architectures of Question Answering Systems

Characteristics	**Closed Domain QAS**	**Open Domain QAS**	**Rule based QAS**	**Web Based QAS**
Feature	In close domain system, answer to question has to be stored in a database by expert i.e. They are not available in public domain.	This system aims at returning an answer in response to user's question; The reply is a short texts rather than lengthy list of relevant documents.	This system is mainly used for reading comprehension in which one can read a story and find sentence in story that best answers given question.	The most important feature of web based QAS is "snippet tolerant" that allows it to provide correct responses to the QAS while searching through any search engines like Google, yahoo.
System Used	Template Matching so that solution could be matched to the respective answer in the database	INSUN05QA	Quarc (Question Answering For Reading Comprehension)	LAMP
Approaches	Automated FAQ answering system that replies with pre stored answers to user questions asked in ordinary English rather than keyword using some mechanisms like disemvoweling, matching synonyms.	The system used to deal with the "list" questions and "Other" questions are seemed simple and arbitrary. In this first preprocessing of question and documents are done then analysis of questions are done along with that Documents are related to each questions are extracted from corpus by using document retrieval and related passages are returned by passages retrieval. And last module extracts the best answers.	The system uses heuristics rule that look for lexical and semantic clues in the question and the story.	The system uses snippets in the search results returned by a search engine like Google. Then extract all plausible answers from the search results according to the type of question identified by the question classification module, Finally select the most plausible answers to return.
Semantic	Yes	Yes	Quarc	Yes

Features			System uses lexical and semantic heuristics to look for evidence that a sentence contain the answer to questions.	
Syntactic Features	In Template Matching, templates are created according to a specific syntax. The syntax of the templates is defined so that a single template could match many different variants of the same question.	Yes	Syntactic analysis is not used.	Yes
Performance	The system developed is a smart, user friendly automatic Answering with the ability of detecting and answering questions asked in English or SMS language.	The system obtaining a very Perfect performance in the test of 2005 TREC(Text Retrieval Conference) QA tracks.	Quarc found the correct sentence 40% of time, which is encouraging given the simplicity of its rule.	LAMP performs very well on some types of questions such as PERSON, LOCATION, and DATE. Performance of LAMP is evaluated using MRR(Mean Reciprocal Rank) metric.
Advantages	(1) Precision of the retrieval is high because the keywords are selected using human intelligence; (2) It is an evolving system, because its question answering ability improves as more questions are asked, and new FAQ entries are added to the database.	The system is a domain-independent question-answering which is based on information retrieval in a large-scale collection of text and improve the accuracy of system.	The system does not use deep language understanding or sophisticated techniques. It uses hand crafted heuristics rule. The system performed best on WHEN question achieving 55% accuracy.	The system take advantage of snippet because it is time consuming to download and analyze original web documents.
Disadvantages	Templates need to be written manually for all questions.	Documents retrieval module is built based on SMART. In SMART, user interaction is needed and based on web query form.	The system performed worst on WHAT and WHY questions, reaching only 28% accuracy.	The problem of scaling question answering techniques in the information retrieval and information extraction to the web.

4. Conclusion

Search engines can return ranked documents as a result for any query from which the user struggle to navigate and search the correct answer. This process wastes user's navigation time and due to this the need for automated question answering systems becomes more urgent. So we need such a system which is capable of replying the exact and concise answer to the question posed in natural language. The best way to address this problem is use of Question answering systems (QAS). The basic aim of QAS is to provide short and correct answer to the user saving his/her navigation time. The concept of Natural Language Processing plays an important role in developing any QAS. Different architectures developed to extract best answers posed in natural languages such as open domain QAS, closed domain QAs, rule based QAS, and web based QAS.

New approaches and methods are developed for open domain question answering system can be applied to web based and rule based system and evaluated.

References

[1] Chowdhury, Gobinda G. "Natural language processing." *Annual review of information science and technology* 37.1 (2003): 51-89

[2] Benamara, Farah. "Language and reasoning for question answering: state of the art and future directions." *Proceedings of the workshop KRAQ06: Knowledge and Reasoning for Language Processing, Trento, Italy*. 2006.

[3] Walke, P. P., and S. Karale. "Implementation approaches for various categories of question answering system." *Information & Communication Technologies (ICT), 2013 IEEE Conference on*. IEEE, 2013.

[4] Gunawardena, T., Lokuhetti, M., Pathirana, N., Ragel, R., & Deegalla, S. (2010, December). An automatic answering system with template matching for natural language questions. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on* (pp. 353-358). IEEE

[5] Maxwell, Kerry (2007, August 13). “Disemvowelling or disemvoweling” [Online]. *Word of the Week Archive*. Macmillan Available:http://www.macmillandictionaries.com/wordoftheweek/archive/070813-disemvowelling.html

[6] Fellbaum, Ch. (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

[7] YU-MING ZHAO et al “*AN OPEN DOMAIN QUESTION ANSWERING SYSTEM BASED ON IMPROVED SYSTEM SIMILARITY MODEL*” Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

[8] D. Lin, “A Dependency-based Method for Evaluating Broad-Coverage Parsers”, Proceedings of IJCAI-95,pp. 1420-1425, 1995.

[9]Buckley, C., Singhal, A., and Mitra, M, “Using Query Zoning and Correlation with SMART: TREC-5”, In Proceeding of the 5th Text Retrieval Conference, NIST, pp. 105-119, 1996.

[10] Riloff, Ellen, and Michael Thelen. "A rule-based question answering system for reading comprehension tests." *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems-Volume 6*. Association for Computational Linguistics, 2000.

[11] Zhang, Dell, and Wee Sun Lee. "A web-based question answering system." (2003)

[12] C. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.