Carol Peters
Martin Braschler
Paul Clough

# Multilingual Information Retrieval

## From Research To Practice

Springer

# Multilingual Information Retrieval

Carol Peters • Martin Braschler • Paul Clough

# Multilingual Information Retrieval

## From Research To Practice

Springer

Carol Peters
Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie
dell'Informazione
Via G. Moruzzi 1
56124 Pisa
Italy
carol.peters@isti.cnr.it

Martin Braschler
Zurich University of Applied Sciences
Institute of Applied Information
Technology
Steinberggasse 13
8401 Winterthur
Switzerland
martin.braschler@zhaw.ch

Paul Clough
University of Sheffield
Information School
211 Portobello Street
Regent Court
Sheffield, S1 4DP
United Kingdom
p.d.clough@sheffield.ac.uk

ISBN 978-3-642-23007-3 e-ISBN 978-3-642-23008-0
DOI 10.1007/978-3-642-23008-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011944193

ACM Codes: H.3, I.7, H.5, I.5, I.2.7



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*This book is dedicated to our families with thanks for the love, support and encouragement they have given us.*
*It is also dedicated to all our friends and colleagues who have collaborated with us in organizing and supporting the Cross-Language Evaluation Forum (CLEF) over the years.*

# Foreword

This book, written by Carol Peters, Martin Braschler and Paul Clough following nearly a decade of intensive research on Multilingual Information Access, constitutes a pertinent synthesis of a series of achievements but also points in new directions for research in the forthcoming years. As is commonly recognised, we are living in a multilingual world, and this is an aspect that should remain the norm, both now and in the future. Even though hundreds of natural languages will disappear over the next decades, a small set of languages tends to be spoken by an increasing number of people. According to the website Ethnologue,[1] 8 languages are being used by around 39% of humans, while 85 languages are in use by 78%. The most popular languages these days are Chinese, Spanish and English, but even though the latter is labelled 'English', there are many contrasting variants found in Manhattan or in New York City's suburbs, and also in newspapers such as the *The Times* or in blog posts. Within a given language variations in usage can be quite significant and thus require that computer system designers undertake modifications to access them.

This book is also related to the evaluation campaigns carried out by the Cross-Language Evaluation Forum (CLEF) with its focus on multilingual retrieval in European countries. The particular problems and challenges related to multilingual information access are of special importance to the Old Continent, with Europeans having to handle information provided in a variety of languages when carrying out such tasks as searching for national statistics written in a foreign language, booking a hotel or even spending holidays by the sea. Unlike the Far East countries where the communication problems are usually limited to English and another language (e.g. Chinese, Japanese or Korean), Europeans live in a multilingual environment yet many of them possess only a cursory knowledge of one or two other languages. In this context, during its first 10 years, various CLEF campaigns have established a set of pertinent corpora and tools that can be applied to manipulate several of these languages. Thanks to this international effort, many quality tools have now been

[1] http://www.ethnologue.com/

made available to access information written in several language groups, including the Germanic (German, English, Dutch, and Swedish), Romance (Spanish, Portuguese, French, and Italian), Slavic (Russian, Bulgarian, and Czech) family groups, as well as members of the Uralic family (Finnish, Hungarian).

As described in this book, language diversity has generated an extensive variety of challenges to be solved by computer scientists. In German, for example, compound constructions are very frequent but the real concern is that the same concept might be expressed by two (or more) formulations, thus rendering it more difficult to find useful matches between search keywords and target items. The numerous grammatical cases found in Finnish grammar represent another example. Here the real problem involves irregularities imposed by vowel harmonies, which in turn generate complex morphological processing requirements. More research is required in order to promote better automatic processing of such linguistic constructions.

In addition to designing effective and efficient processing techniques for the various natural languages, multilingual information access must make use of some form of automatic translation. To cross these language barriers, various translation strategies have been suggested, implemented, and evaluated. When handling related languages such as French and English for example, we might just consider one language as a misspelled form of the other, so that translating can simply be viewed as a spelling correction problem. Other and more effective methods have of course been developed, and are often based on machine-readable bilingual dictionaries, automatic machine translation systems, as well as statistical translations based on parallel or comparable corpora. In this sense, however, many problems still have to be solved, particularly those involving the less frequently spoken languages.

Multilingual information access is of course not only limited to textual documents, but must be applicable to other media including audio (e.g. interviews), music, images, photographs, videos, movies, etc. As this book demonstrates, cultural heritage consists of a fruitful source of information when exploring multilingual data access. An example would be famous works of art commonly identified under different names, depending on the underlying language (Mona Lisa, Monna Lisa, Mona Liza, La Joconde or La Gioconda).

Finally, unless users accept these more advanced technological solutions, they will be of little value. In this perspective, this book covers pertinent studies and findings applicable to designing multilingual interfaces for various media. Such aspects are of prime importance when designing multilingual applications for websites, mobile phones and other technological devices. In summary, this interesting book covers a range of human communication problems related to media and technology, in efforts undertaken to facilitate understanding and communication between human beings, encourage better dialogue between cultures, and establish bridges between past, present and future cultural heritage challenges.

Jacques Savoy
Department of Computer Science
University of Neuchatel
Switzerland, June 2011

# Preface

This book gives a comprehensive description of the technology involved in designing and developing systems and functionality for multilingual information access and retrieval. Our objective is to provide the reader with a full understanding of the various issues involved in creating systems to make digitally stored information items available, regardless of the language they are written in. The book is aimed at graduate students and practitioners with a basic understanding of 'classical' text retrieval methods. The growing amount of non-English information accessible globally and the increased world-wide exposure of enterprises leaves no doubt that these target groups will need to adapt their existing knowledge of Information Retrieval (IR) methods to new, multilingual settings. Our intention is to close the gap between the material covered by most of the classical IR textbooks and this new operational reality.

The book is divided into six chapters. The purpose is to accompany the reader step by step through the various stages involved in building, using and evaluating Multilingual Information Retrieval (MLIR) systems, concluding with some examples of recent applications of MLIR technology. Some of the techniques that we describe have recently started to appear in commercial search systems, while others have the potential to be part of future incarnations.

The aim of the first chapter is to present the reasons why effective and efficient access to information across language boundaries has become so crucial in our global digital society. The basic concepts are explained and discussed and the major terminology that will be used throughout the book is defined. The Introduction also presents a brief history of academic and commercial activity in this sector, and lists the main challenges that are being faced today.

As is explained in the Introduction, multilingual information retrieval system development is concerned with managing information access and discovery in multiple languages both monolingually and across languages, whereas Cross-Language Information Retrieval (CLIR) refers specifically to those technologies that concern the querying of a multilingual collection in one language in order to retrieve relevant documents in other languages. Effective monolingual information

retrieval has been shown to be one of the preconditions for implementation of MLIR/CLIR systems. Chapter 2 thus presents the mechanisms necessary for building monolingual information retrieval systems, namely indexing and matching strategies and provides details on the adaption of these mechanisms to different languages.

Chapter 3 then presents the various strategies that can be used to support cross-language information retrieval and describes the integration of different types of translation components into the IR system. The difficulties raised by query and/or translation ambiguity, insufficient lexical coverage and the quality of the resources available for different languages are discussed and suggestions for best practices are given.

In MLIR/CLIR systems, the design of an effective search interface and the provision of functionality to assist the user with searching are vital as users cross the language boundary and interact with material in unfamiliar languages. The influences of the user's language skills and cultural background will affect the design. For these reasons, in Chapter 4, the focus is on the user's information-seeking behaviour, the contexts in which users search for information and their cognitive needs and abilities. Guidelines are provided on how to design multilingual search interfaces that support user – system interaction, e.g. in query formulation and translation, in document selection and examination, in browsing and visualisation.

Chapter 5 describes the evaluation of systems for multilingual information retrieval from both system- and user-oriented perspectives and provides the reader with information on how to undertake their own evaluation activity, listing the points that need to be taken into consideration before starting out. The chapter also discusses the importance of the role of evaluation in advancing the state-of-the art and producing systems that give better results not only in terms of performance in retrieving relevant information but also with respect to success in satisfying the expectations of the user.

Information retrieval is no longer just about text and document retrieval; today's content is overwhelmingly multimedia, and the interests and needs of the user are rapidly changing. For this reason, in the final chapter, we discuss ways in which the technologies developed to implement systems for multilingual and cross-language information retrieval are adopted in areas that go beyond textual document search such as multimedia retrieval and information extraction over language boundaries. We also describe a range of practical application domains which employ these technologies and present some of the challenges raised by the transition of the research prototypes to real-world operational systems.

The topic covered is a wide and complex one; we have tried to provide a comprehensive overview. Our aim has been to offer guidelines and information on all the aspects that need to be taken into consideration when building an MLIR/CLIR system, while avoiding too many 'hands-on details', that rapidly become obsolete as software is in continual evolvement. Each chapter is furnished with a carefully compiled list of references and Chapters 2–5, which provide the main details on developing and evaluating MLIR/CLIR systems, also conclude with a list of suggested reading so that more detailed information can be found as and when

necessary. It is our hope that this book will prove interesting reading and provide a useful source of information for students, scholars, system developers and all those interested in ways to satisfy their need for the acquisition and dissemination of information, however, wherever, and in whatever language it is stored.

Carol Peters
Martin Braschler
Paul Clough

# Acknowledgements

This book would not have been possible without the experience and knowledge gained from our involvement over many years in the activities of the Cross-Language Evaluation Forum (CLEF). We are indebted to those who worked with us, both organisers and participants. In particular we are grateful to a number of individuals who have shaped our thinking and taught us many things: Maristella Agosti, Donna Harman, Noriko Kando, Mark Sanderson, Jacques Savoy, Peter Schäuble, Costantino Thanos, and Ellen Voorhees. In addition, we thank the European Union and the various national and international funding agencies that have helped to stimulate and promote developments in multilingual information access and have supported work that has contributed to this book. Funded projects that should be mentioned include ACCURAT, CLEF, Eurovision, DELOS, MultiMatch, and TrebleCLEF. We also thank our colleagues at the Information School (University of Sheffield), the Institute for Information Science and Technologies (Italian National Research Council), and the Institute of Applied Information Technology (Zurich University of Applied Sciences).

Finally we express our gratitude to those who have reviewed versions of this book in one way or another. Last but not least, we thank Ralf Gerstner from Springer Verlag for his encouragement and support, and David Clough for his valuable assistance in proofreading the chapters and producing the final version of the book.

# Contents


**1 Introduction** ..... 1
1.1 The Growth of the Digital Universe ..... 1
1.2 The Terminology ..... 5
1.3 A Brief History ..... 5
1.3.1 Enabling Technologies and Standards ..... 6
1.3.2 Publicly-Funded Research Initiatives ..... 7
1.3.3 Conferences and Evaluation Campaigns ..... 9
1.3.4 Commercial Products ..... 10
1.4 The Current Research Challenges ..... 13
References ..... 15

**2 Within-Language Information Retrieval** ..... 17
2.1 Introduction ..... 17
2.2 The Retrieval Problem and Its Consequences ..... 20
2.3 Implementation of a Within-Language Information Retrieval System ..... 22
2.4 Indexing Phase ..... 24
2.4.1 Pre-processing (Step 1) ..... 25
2.4.2 Language Identification (Step 2) ..... 26
2.4.3 Document Formation (Step 3) ..... 27
2.4.4 Segmentation, Tokenisation, Parsing (Step 4) ..... 28
2.4.5 Feature Normalisation (Step 5) ..... 31
2.4.6 Enrichment (Step 6) ..... 35
2.5 Matching Phase ..... 36
2.5.1 'Bag of Words' Paradigm ..... 36
2.5.2 Inverted Index ..... 38
2.5.3 Basic Matching Algorithm ..... 39
2.5.4 Vector Space Model ..... 41
2.5.5 The tf.idf-Cosine Weighting Scheme ..... 43
2.5.6 Relevance Feedback ..... 46

2.5.7 Probabilistic Weighting Schemes ..... 48
2.5.8 Ranking Using Language Models ..... 50
2.5.9 Off-Page Information: Page Rank ..... 51
2.6 Summary and Future Directions ..... 52
2.7 Suggested Reading ..... 53
References ..... 54

**3 Cross-Language Information Retrieval** ..... 57
3.1 Introduction ..... 57
3.2 Implementation of Cross-Language Information Retrieval ..... 58
3.2.1 Query Translation and Document Translation ..... 58
3.2.2 No Translation ..... 59
3.2.3 Different Types of Translation Resources ..... 60
3.2.4 Term Ambiguity ..... 61
3.3 Translation Approaches for Cross-Language Information Retrieval ..... 62
3.3.1 Machine-Readable Dictionaries ..... 62
3.3.2 Statistical Approaches ..... 65
3.3.3 Pre-translation and Post-translation Query Expansion ..... 69
3.3.4 Machine Translation ..... 71
3.3.5 Combination Approaches ..... 71
3.4 Handling Many Languages ..... 73
3.4.1 CLIR Flows ..... 73
3.4.2 Merging Across Languages ..... 77
3.4.3 Document Translation ..... 79
3.4.4 Indirect Translation ..... 79
3.5 Summary and Future Directions ..... 80
3.6 Suggested Reading ..... 82
References ..... 83

**4 Interaction and User Interfaces** ..... 85
4.1 Information Seeking and User Interaction ..... 85
4.2 Users' Information Needs and Search Tasks ..... 89
4.3 Users' Language Skills and Cultural Differences ..... 92
4.4 Supporting Multilingual User Interaction ..... 94
4.4.1 Query Formulation and Translation ..... 95
4.4.2 Document Selection and Examination ..... 102
4.4.3 Query Reformulation ..... 108
4.4.4 Browsing and Visualisation ..... 109
4.5 Designing Multilingual Search User Interfaces ..... 114
4.5.1 User-Centred Design ..... 116
4.5.2 Internationalisation and Localisation ..... 118
4.5.3 Case Study: CLIR in Google's Web Search ..... 121
4.6 Summary and Future Directions ..... 123
4.7 Suggested Reading ..... 124
References ..... 125

**5 Evaluation for Multilingual Information Retrieval Systems** ..... 129
5.1 Introduction ..... 129
5.2 System-Oriented Evaluation ..... 130
5.2.1 The Cranfield Tradition ..... 131
5.2.2 Evaluation Campaigns ..... 132
5.2.3 Building a Test Collection ..... 133
5.2.4 Promoting Research into Multilingual and Multimedia System Development via Evaluation ..... 141
5.2.5 Alternative Methodologies for Test Collection Construction ..... 143
5.2.6 Performance Measures ..... 145
5.2.7 Statistical Significance Testing ..... 154
5.2.8 System Effectiveness and User Satisfaction ..... 155
5.3 User-Oriented Evaluation ..... 156
5.3.1 Experimental Design ..... 157
5.3.2 Evaluating Interactive CLIR Systems at CLEF ..... 159
5.3.3 Alternative Performance Measures ..... 160
5.4 Evaluating Your Own System ..... 161
5.5 Summary and Future Directions ..... 164
5.6 Suggested Reading ..... 165
References ..... 166

**6 Applications of Multilingual Information Access** ..... 171
6.1 Introduction ..... 171
6.2 Beyond Multilingual Textual Document Retrieval ..... 172
6.2.1 Image Retrieval ..... 173
6.2.2 Speech Retrieval ..... 177
6.2.3 Video Retrieval ..... 180
6.2.4 Question Answering ..... 183
6.3 Multilingual Information Access in Practice ..... 188
6.3.1 Web Search ..... 189
6.3.2 Digital Libraries and Cultural Heritage ..... 191
6.3.3 Medicine and Healthcare ..... 195
6.3.4 Government and Law ..... 196
6.3.5 Business and Commerce ..... 199
6.4 Summing Up ..... 201
References ..... 202

**Glossary of Acronyms** ..... 209

**Index** ..... 213

# Chapter 1
# Introduction

**The Grand Challenge**

*"Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified."*

Douglas W. Oard and David Hull, AAAI Symposium on Cross-Language IR, Spring 1997, Stanford, USA

**Abstract** Multilingual information access and retrieval is a specific area of the academic domain of information access and retrieval; the main focus is the development of systems for information discovery in multiple languages, both monolingually and across languages. There is both a social and an economic need for such systems and there is ample evidence that this need will grow substantially over the coming years. In this introduction, we describe the range and intentions of research and development in this area from its recognition as an independent discipline in the mid-1990s to the challenges that it is now facing today.

## 1.1 The Growth of the Digital Universe

The term 'global information society' is often used to describe the environment in which we live at the beginning of the twenty-first century, the term meaning different things to different people. Generally speaking, there is agreement that there is an ever greater amount of information at one's disposal. The major sources of knowledge and reference are increasingly digital. As a result of the diffusion of the Internet and the World Wide Web, vital information has never before been this available to an increasingly wider public, breaking a former 'information monopoly' of select circles. If this information is successfully made accessible, it has the power to transform society in a profound way. However, a major obstacle to the worldwide dissemination and access to information is the boundary posed by language diversity. Information is published digitally every day in a myriad of the world's languages. The challenge is to provide tools that enable users of global

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_1, 

networks to find, retrieve and understand information of interest in whatever language it has been stored.

At the beginning this was not an apparent problem. The first websites were almost entirely dedicated to provision of information in English and the first search services in the mid-1990s (e.g., Lycos, AltaVista, Yahoo!) were implemented to meet the needs of an English-speaking community. The users of these services had mainly academic backgrounds and had sufficient English language skills to formulate meaningful queries in English and to understand the documents retrieved. However, in the last few years of the twentieth century, the World Wide Web expanded rapidly in the more highly developed countries reaching a mass audience and impacting on many aspects of daily life, changing the ways people communicate, shop and plan travel. From this moment on, the percentage of English content started to decline and monolingual search services began to be available in some of the major languages.[1]

Nowadays, in the twenty-first century, the Internet and the World Wide Web are used throughout the world for communication, business and leisure activities, and the dissemination of information, and the number of languages in which electronically accessible material is available is in continual growth. Tables 1.1 and 1.2 give a good idea of the growth of the digital universe in the first decade of this millennium. Table 1.1 shows that while the percentage of the population that uses the Internet is still much higher in the more developed parts of the globe (North America, Australasia and Europe), there was a very strong spurt of growth in the period 2000–2010 in the lesser developed regions. This trend is expected to continue.

While Table 1.1 shows where and to what extent the Internet is being used globally, Table 1.2 lists the ten most used languages on the Web as of 2010. Although English still maintains an important position as a 'global' language, the table shows that the number of internet users speaking Chinese has grown more than a 1,000-fold in the period 2000–2010. Judging from this trend, within a few years Chinese will be the predominant web language, both for users and for content.[2] The 2,500% growth of Arabic in the same period is similarly impressive and indicative of future trends.

From these tables, it is clear that the position of English as the dominant language is declining and the Web is becoming a truly global information resource. The question is: How much information is lost or remains hidden because it is

[1] In this period, an increasing proportion of new users coming online were individuals and small businesses chiefly interested in using the Internet for local communication. In non-English speaking countries, large firms or public institutions may have an incentive to also post their web pages in English, but a small local business does not. As more people in a language community come online, content and service providers have a strong interest in accommodating them in their own language.

[2] In 2009 at the Gartner Symposium, Orlando, Eric Schmidt, CEO of Google, predicted that within 5 years the Internet will be dominated by Chinese-language content.

**Table 1.1** World Internet users and population statistics, June 2010[a]

| World regions | Population (2010 est.) | Internet users Dec. 2000 | Internet users June 2010 | % of population | Growth 2000–2010 (%) | Internet users % of total |
|---|---|---|---|---|---|---|
| Africa | 1,013,779,050 | 4,514,400 | 110,931,700 | 10.9 | 2,357.3 | 5.6 |
| Asia | 3,834,792,852 | 114,304,000 | 825,094,396 | 21.5 | 621.8 | 42.0 |
| Europe | 813,319,511 | 105,096,093 | 475,069,448 | 58.4 | 352.0 | 24.2 |
| Middle East | 212,336,924 | 3,284,800 | 63,240,946 | 29.8 | 1,825.3 | 3.2 |
| North America | 344,124,450 | 108,096,800 | 266,224,500 | 77.4 | 146.3 | 13.5 |
| Latin America | 592,556,972 | 18,068,919 | 204,689,836 | 34.5 | 1,032.8 | 10.4 |
| Oceania/ Australia | 34,700,201 | 7,620,480 | 21,263,990 | 61.3 | 179.0 | 1.1 |
| **Total** | 6,845,609,960 | 360,985,492 | 1,966,514,816 | 28.7 | 444.8 | 100.0 |

[a] Source: Internet World Stats: http://www.internetworldstats.com/stats.htm

**Table 1.2** Top ten languages used in the Web, June 2010[a]

| Top ten languages on Internet | Internet users by language[b] | Internet penetration by language[c] (%) | Growth in Internet 2000–2010 (%) | Internet users % of total | World population for this language 2010 estimate |
|---|---|---|---|---|---|
| English | 536,564,837 | 42.0 | 281.2 | 27.3 | 1,277,528,133 |
| Chinese | 444,948,013 | 32.6 | 1,277.4 | 22.6 | 1,365,524,982 |
| Spanish | 153,309,074 | 36.5 | 743.2 | 7.8 | 420,469,703 |
| Japanese | 99,143,700 | 78.2 | 110.6 | 5.0 | 126,804,433 |
| Portuguese | 82,548,200 | 33.0 | 989.6 | 4.2 | 250,372,925 |
| German | 75,158,584 | 78.6 | 173.1 | 3.8 | 95,637,049 |
| Arabic | 65,365,400 | 18.8 | 2,501.2 | 3.3 | 347,002,991 |
| French | 59,779,525 | 17.2 | 398.2 | 3.0 | 347,932,305 |
| Russian | 59,700,000 | 42.8 | 1,825.8 | 3.0 | 139,390,205 |
| Korean | 39,440,000 | 55.2 | 107.1 | 2.0 | 71,393,343 |
| **Top ten languages** | 1,615,957,333 | 36.4 | 421.2 | 82.2 | 4,442,056,069 |
| Rest of languages | 350,557,483 | 14.6 | 588.5 | 17.8 | 2,403,553,891 |
| **World total** | 1,966,514,816 | 28.7 | 444.8 | 100.0 | 6,845,609,960 |

[a] Source: Internet World Stats: http://www.internetworldstats.com/stats7.htm

[b] Although many people are competent in more than one language, the table assigns just one language per person.

[c] Internet Penetration is the ratio between the sum of internet users speaking a language and the total population estimate that speaks that specific language.

published in one language rather than another and to what extent is this important? Foreign language skills vary considerably according to geographical location, educational and cultural backgrounds. How many people are willing or able to search for information in languages other than their own?

At the same time it must not be forgotten that the World Wide Web is just one, even if the most highly visible, part of the so-called digital universe. The populations of highly developed countries are nowadays often described as forming

'information societies' as the manipulation of information has become a central economic activity. Businesses that need to strive for a competitive advantage in this environment are dependent on effective and efficient ways to access large amounts of information. The intranets of many large international public and private organisations increasingly contain multilingual information as interests and activities transcend national boundaries and the use of a single common language is not always acceptable.

Thus, as the digital universe expands, situations where a user is faced with the task of querying a multilingual document collection are increasingly common. Sectors where facilitating access to information in multiple languages is becoming important include: international legal studies and practices, multilateral anti-terrorism and criminal justice activities, digital libraries, tourism, global market research, international banking and investment, journalism, medical research.

Examples of tasks involving cross-language searching are:

- Journalists wanting to search for news stories in other countries, and languages;
- Patent lawyers looking for patent infringements within multilingual databases;
- Business analysts wishing to gather foreign business information and provide services to different countries;
- Immigrants having poor local language skills scanning web pages for information about their new environment;
- Investors interested in examining new markets seeking news reports or web documents about foreign companies;
- Patients or caregivers finding medical treatment information from other countries and languages;
- Foreign travellers searching for local information, such as events or services, *en route*.

These users could all benefit from having the assistance of some kind of multilingual retrieval functionality. Language skills vary considerably according to geographical location, educational and cultural backgrounds. For users with a good passive knowledge of a second language but unable to formulate queries that adequately express their information need in that language, a system that translates their queries and finds relevant documents in the target language will be sufficient. However, users looking for information in an unfamiliar language need a system that includes translation aids to help them understand their search results.

In summary, there is a widely recognised need for technologies that enable users to search for and discover digital information, wherever and however it is stored and in whatever language. This need encompasses both the private and the public sectors, involves government, academia and industry, and includes most areas of society, e.g., education, commerce, leisure, tourism, etc. If the goal is to be fully achieved, then the objective must be not just to find relevant information, in any media and any language, but to be able to understand, interpret and reuse it. This is what multilingual information access and retrieval is all about.

## 1.2 The Terminology

Multilingual information access and retrieval is a specific (and very multidisciplinary) area of the academic domain of information access and retrieval. The focus on aspects that regard language understanding and processing means that it combines strategies and technologies used in classical Information Access (IA) and Information Retrieval (IR) with methodologies, tools and resources coming from the Computational Linguistics (CL) and Natural Language Processing (NLP) sectors. Three terms are commonly used when discussing research in this area: Multilingual Information Access, Multilingual Information Retrieval, and Cross-Language Information Retrieval.[3] In the literature, at times, the meaning of these terms may overlap. It is thus important to define them clearly here.

We use the term Multilingual Information Access (MLIA) in its broadest possible sense. MLIA addresses the problem of accessing, querying and retrieving information from collections in any language at any level of specificity. It covers the most basic enabling techniques ranging from those that regard the overall management of scripts in any language, e.g., language identification, character encoding, visualisation and display, up to the overall access and retrieval of multilingual information.

More specifically, systems that process information in multiple languages (either queries, documents, or both) are called Multilingual Information Retrieval (MLIR) systems, whereas Cross-Language Information Retrieval (CLIR) is used to refer precisely to those technologies that concern the querying of a multilingual collection in one language in order to retrieve relevant documents in other languages and concerns issues of translation, merging, summarisation and presentation of the results. MLIR is thus a more general term and can embrace the concept of CLIR as a MLIR system is concerned with managing information access and discovery in multiple languages both monolingually and across languages. In this book, we do not describe any of the basic MLIA enabling technologies in any detail, but pose the main focus on issues that regard MLIR and CLIR as this is where current research and development activities are focused.

## 1.3 A Brief History

Although the very first experiments in cross-language text retrieval were made by Gerard Salton in the 1970s (Salton 1971) using a carefully constructed multilingual thesaurus, research in this field did not really take off until the mid-1990s when the

[3] Other terms that have been used are Translingual and Cross-Lingual IR. 'Translingual' was made popular for a short period by the TIDES project in the US but now seems to have fallen into disuse; 'cross-lingual' can still be found but 'cross-language' is generally the preferred choice.

growth in popularity of the multilingual Web meant that it became an important topic. We can identify four main activities which have contributed to promoting the creation of MLIR/CLIR systems in both the academic and commercial sectors: the development of basic enabling technologies and standards; the public funding of research activities; the promotion of experimentation by international conferences and evaluation initiatives; the marketing of commercial tools.

### *1.3.1 Enabling Technologies and Standards*

Instrumental in the rise in interest was the development of some of the basic enabling technologies and standards. For example, ISO Standard 5964 providing guidelines for the establishment of multilingual thesauri was first released in 1978, and a revised version was published in 1985 (ISO 1985). Multilingual thesauri are an important resource when building domain-specific MLIR systems and were employed in many of the first experimental prototypes. This was recognised in April 2005 when the International Federation of Library Associations (IFLA) presented their Guidelines for Multilingual Thesauri, with the objective of adding to and extending ISO-5964-1985. However, a real breakthrough was the introduction of Unicode. The Unicode Standard, Version 1.0, was published in 1991 with the aim of promoting a universal, uniform, unique, unambiguous worldwide character encoding standard. Since then Unicode Standards have been released at varying intervals. Unicode Standard 6 was released in 2010.[4] In 1993 ISO/IEC 10646 was released as the 'Universal Multiple-Octet Coded Character Set' (UCS). Unicode-compatible UCS aims at eventually including all characters used in all the written languages in the world (ISO/IEC 1993). Nowadays UTF-8, an 8-bit variable length character encoding for Unicode, is commonly employed. UTF-8 can represent every character in the Unicode character set and is also backward-compatible with ASCII. Another important set of standards are the language code schemes which attempt to classify human languages and dialects. The most commonly used are ISO 639-1, introduced in 2002, and ISO 639-2, first released in 1998. The former is a two letter code system covering 136 major languages, whereas the latter is a more extensive three-letter system of 464 codes. ISO 639-3 is an extension which attempts to cover all known spoken or written languages in 7,589 entries. The existence and wide-spread acceptance of these various standards has been important in the internationalisation and localisation of websites, i.e., the linguistic and cultural adaptation of the sites of an organisation or company to meet the requirements of a particular target area.[5]

[4] See the Unicode web page http://www.unicode.org/ for Unicode standards and updates.

[5] Internationalisation and localisation are discussed in the section on implementing multilingual user interfaces in Chapter 4.

### 1.3.2 Publicly-Funded Research Initiatives

Since the mid-1990s there have been many research activities in the MLIA domain sponsored by various types of public funding. In particular, the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA) in the US and the European Commission (EC) in Europe, have funded a number of initiatives. While a major interest in the US is the development of systems that provide access to content in languages other than English (often for defence purposes), the European Union (EU) is a truly multilingual environment with 23 official languages in 2010, and more will be added as new countries join. Thus the EU is committed to promoting tools for the dissemination and access of information in many languages in order to encourage communication and sharing of information across language boundaries while preserving and protecting the status of national languages. Since 1990, the Information Society and Media Directorate General of the EC has funded many research initiatives aimed at promoting the development of language technologies and tools with particular emphasis on machine translation (MT) and language resources such as machine-readable general purpose dictionaries and domain-specific lexicons. Over the years, the focus has shifted from technologies just interested in text to include other media such as speech and video.[6] India is another geographic area that can be compared to Europe with respect to the number of languages and political commitment to language preservation. Since 1991, the Indian government is funding research activities in this field, partly through the programme for Technology Development for Indian Languages (TDIL) which aims at "*developing information processing tools to facilitate human machine interaction in Indian languages and to create and access multilingual knowledge resources*".[7] Here below we just mention a few of the most significant publicly-funded projects and activities which have helped to advance the state-of-the-art.

In 1994 the final prototype of EMIR (European Multilingual Information Retrieval) was released. EMIR was an EC project and one of the first general purpose cross-language systems to be implemented and evaluated (EMIR 1994). Since then the Commission has sponsored a number of information retrieval projects that have involved the development of MLIR/CLIR functionality.[8] In 1995, SYSTRAN Software Inc. received funding from US Government to develop a CLIR system based on NLP and MT technology. In 1997, the EU-NSF Working

[6] Most of these initiatives have been funded by the Directorate for Digital Content and Cognitive Systems and the Language Technologies programmes.

[7] http://tdil.mit.gov.in/

[8] Two of these projects which have had considerable impact and are cited several times in this book are the Clarity and the MultiMatch projects. The objective of Clarity was to develop general purpose CLIR techniques which would work with minimal translation resources; MultiMatch aimed at providing personalised access to cultural heritage information over both language and media boundaries.

Group on Multilingual Information Access was given mandate to identify and prioritise the major open research issues and propose a short and medium term research agenda (Schäuble and Smeaton 1998). In 1999 the NSF/EC/DARPA report on Multilingual Information Management was released. The aim of this study was to identify how technologies developed in the areas of computational linguistics and information retrieval can be integrated to address problems of handling multilingual and multi-modal information (Hovy et al. 1999).

From 2000 to 2004, DARPA, the US Defense Advanced Research Projects Agency, supported the TIDES programme for Translingual Information Detection, Extraction and Summarization with the goal of "*enabling people to find and interpret needed information, quickly and effectively, regardless of language or medium*". The TIDES programme's ultimate objective was to enable the US to be able to quickly and accurately develop a comprehensive understanding of unfolding international situations.[9] Much work was done within TIDES aimed at developing translation resources and machine translation and document understanding systems. In 2003 the programme developed a test scenario called the 'TIDES Surprise Language Exercise'. The goal was to test the Human Language Technology community's ability to rapidly create language tools for previously un-researched languages. The surprise language chosen for a practice exercise was Cebuano, the *lingua franca* of the southern Philippines. The test language was Hindi. Each language presented special challenges: Cebuano because of the scarcity of electronic resources and Hindi because of the multiplicity of encodings of Hindi texts found on the Web. By the end of the exercise a great deal had been learnt and translation resources had been developed for both languages (Oard 2003).

In 2005 the European Commission launched its 2010 Digital Library Initiative. The vision was to "*make Europe's cultural and scientific heritage accessible to all*" and one of the main steps in achieving this was by providing a common multilingual access point. Two major results of this initiative are The European Library (TEL)[10] and Europeana.[11] The European Library, operational since 1994, offers free access to the bibliographical resources of 48 national libraries of Europe in 35 languages. Much digital content is also available (books, posters, maps, sound recordings, videos). Europeana – the European digital library, museum and archive – aims to provide access to many millions of cultural objects,[12] including photographs, paintings, sounds, maps, manuscripts, books, newspapers and archival papers. Currently both TEL and Europeana provide multilingual interfaces, i.e., users can choose their interface language from a wide selection of European languages. The goal is also to offer cross-language query functionality in the near future.

[9] See DARPA policy statement at http://www.darpa.mil/darpatech99/Presentations/scripts/ito/ITOTIDESScript.txt

[10] http://theeuropeanlibrary.org/

[11] http://www.europeana.eu/

[12] Over 15 million at the beginning of 2011.

### 1.3.3 Conferences and Evaluation Campaigns

The very first workshop on cross-language information retrieval was held at the 1996 ACM-SIGIR conference in Zurich.[13] At the workshop, different approaches to the CLIR problem were presented and a research community began to be identified around this area (Grefenstette 1998). This workshop was followed by a second event at the AAAI Spring Symposium in Stanford in 1997. It was at this meeting that the Grand Challenge quoted at the beginning of this chapter was formulated. This is generally felt to mark the beginning of the recognition of MLIR/CLIR as an independent sector of the IR field and the Grand Challenge is still cited today as the ultimate goal. From 1996 on, many workshops have been held on this topic and aspects of the problem now routinely appear at conferences on digital libraries, information retrieval, machine translation, and computational linguistics. In particular, a series of workshops at SIGIR 2002, 2005 and 2009 have been instrumental in assessing the state-of-the-art and in proposing research agendas for future work (Gey et al. 2005, 2006 and 2009).

Evaluation campaigns have also played an important role in promoting the development of MLIR/CLIR functionality and in influencing directions that future research can take. The purpose of an evaluation campaign is to support and encourage research by providing the infrastructure necessary for large-scale testing and comparison of techniques and methodologies and to increase the speed of technology transfer. End products are valuable test collections of resources that can be used for system benchmarking.[14]

Modern information retrieval evaluation began with the first edition of TREC[15] (Text REtrieval Conference) in 1992. TREC is co-sponsored by the National Institute of Standards and Technology (NIST) and the US Department of Defense. Over the years, TREC has introduced many innovative evaluation ideas and approaches (Harman 2003). In particular, it introduced the first evaluation exercises in the field of multilingual and cross-language IR, thus paving the way for later work by the Cross-Language Evaluation Forum (CLEF[16]) for European languages, the NII Text Collection for IR (NTCIR[17]) for Asian languages and the Forum for Information Retrieval Evaluation (FIRE[18]) for Indian languages.

Although the main focus of TREC has always been on experiments on English texts, TREC-3 offered a first foreign language track for Spanish and this was

[13] The actual name was 'Workshop on Cross-Linguistic Information Retrieval', however discussing terminology for this new sector of IR the participants felt that 'cross-language' was a more appropriate term.

[14] The creation of test collections for (ML)IR is described in detail in Chapter 5.

[15] http://trec.nist.gov/

[16] http://www.clef-campaign.org/

[17] http://research.nii.ac.jp/ntcir/

[18] http://www.isical.ac.in/~clia/index.html

repeated in TREC-4 and TREC-5. The TREC-3 and -4 Spanish collections were used for one of the earliest CLIR studies, a widely cited paper on reducing ambiguity in cross-language IR using co-occurrence statistics (Ballestreros and Croft 1998). TREC-5 also introduced a Chinese language track using the GB character set of simplified Chinese. Chinese monolingual experiments on TREC-5 and TREC-6 collections stimulated research into the application of Chinese text segmentation to information retrieval. From 1997 to 1999 TREC organised the first track testing CLIR systems, operating with European languages – first English, French and German, and later Italian (Harman et al. 2001). Following TREC-8, the co-ordination of European-language retrieval evaluation moved to Europe with the creation of the Cross-Language Evaluation Forum (CLEF) (Peters 2001). In TREC-9, CLIR experiments used a target collection of Chinese documents written in the traditional Chinese character set and encoded in BIG5. In 2001 and 2002, the task of the CLIR track at TREC was cross-language retrieval submitting queries in English to an Arabic document collection (Oard and Gey 2003).

NTCIR is supported by the Japanese Society for the Promotion of Science and the National Institute of Informatics. The first two NTCIR Workshops on Text Retrieval System Evaluation for Asian languages included a Japanese-English track for CLIR (Kando et al. 1999, Kando 2001). NTCIR-3 and -4 set multilingual tasks with Chinese, Korean, Japanese plus English target collections (Kando et al. 2008). The availability of the test collections produced by these workshops has contributed greatly to clearer insights into segmentation and search mechanisms for languages using ideograms.

CLEF is partially supported by the European Commission as it has concentrated on European languages. Highly motivated by the Grand Challenge, it has focused on promoting the development of fully multilingual multimedia retrieval systems and, over the years, has built a number of test collections in different media and different languages (Ferro and Peters 2008). After a start-up exercise in CLEF 2007, FIRE, the Forum for Information Retrieval for Indian languages held its first campaign and workshop in 2008. This was followed by a second campaign in 2009–2010 and a third edition in 2011. Test collections have been created for Bengali, Hindi, Marathi, Punjabi, Tamil and Telugu (FIRE 2008, 2010). A recent special issue of ACM TALIP is dedicated to current research in Indian language IR; many of the papers describe experiments using the FIRE dataset (Harman et al. 2010).

The importance of the role played by these initiatives in building and maintaining IR evaluation infrastructures and test collections and in stimulating research in the domain of IR system development is discussed in more detail in Chapter 5.

### *1.3.4 Commercial Products*

While the research focus has been very much on the development of MLIR/CLIR systems – as described in the rest of this book, the market interest so far has mainly

been concentrated on certain specific components: software for internationalisation/localisation, machine translation tools, multilingual web services.

In a commercial setting, the benefit from internationalisation/localisation is access to wider markets. It costs more to produce products for international markets but in an increasingly global economy supporting only one language/market is scarcely a business option. The last decade has thus seen a strong and growing commercial demand for software that enables enterprises to adapt their products and sites for a specific region or language by adding locale-specific components and translating text.

Machine translation has a long and troubled history – from the toy systems available in the 1950s to the various software packages commercially available today. Although there is still no system that can compete with the work of a human translator, language translation software is gaining an increasing important niche in the market. However, the offer tends to be limited to those languages which have the most economic impact. This was evident in a survey of nine of the best known translation software packages by TopTenReviews,[19] which compared the different products for effectiveness, ease of use, supported formats and available languages. While the number of language pairs offered varied considerably from package to package, there is a general tendency to focus on translation to and from English and a second language, and the second languages available are those which are considered to be of major commercial interest.

There have been several attempts to offer multilingual search as a web service. In 1995 ALIS Technologies launched TANGO, the first multilingual web browser, no longer operational now. The best known search engines for multilingual search today are probably Google and Yahoo! although it is not always easy to locate this functionality on their main sites. Yahoo! started to offer this service in a beta version in 2006. Queries in French and German were automatically translated to four other languages – English, Spanish, Italian and French/German. This functionality can now be found under Yahoo! Babelfish[20] and about 40 language pairs are currently offered; translations are either between English or French and a second language. Google began to offer CLIR functionality in 2007. The user must invoke Google Language Tools. The user's query is translated to the selected target language and the documents retrieved are translated back to the query language using an MT system. The number of possible translation pairs is impressive as well over 50 languages are offered both as source and target. The quality of the translations is variable depending on the domain and the language, but as Google is continuously updating its lexical resources, partially on the basis of usage and user input, the quality is destined to improve. In January 2011, Google announced that it is releasing an alpha version of its Google Translate conversations mode, a technology that allows two people to speak in different languages and have their

[19] TopTenReviews is a website which aggregates reviews for software, hardware, and web services, from other sites and publications, see http://translation-software-review.toptenreviews.com/

[20] http://babelfish.yahoo.com/

words translated in near real time. The initial version is limited to English and Spanish but a wider variety of languages is envisaged.

An important area for MLIA technology is enterprise search. Many businesses have offices all over the world with millions of documents in many different languages. There are a number of platforms offering search capabilities in multiple languages but not many are also able to offer cross-language functionality. The most successful products currently available work in domain-specific contexts, e.g., legal, medical, and defence sectors, tuning their system parameters and optimising their lexical resources to meet the demands of the given sector. Google entered into the enterprise search area in 2008 and will probably have the edge over many competitors precisely because its translation software is very powerful and flexible, giving good results in many domains.

Notwithstanding this market interest and in particular the proliferation of localisation software and translation tools there has been little commercial development or success for CLIR. This is an area where the revenue predictions for market trends have proved over-optimistic. For example, although in 2001, IDC[21] predicted that global revenue for general multilingual support software by 2005 would be about $290 million, in 2005, their reported estimate for that year's revenue was actually below $190 million, and they predicted that the revenue for 2009 would be no higher than $260 million (lower than the original prediction for 2005). Of this, the revenue predicted for CLIR-specific products was considered to be negligible.

In a workshop at SIGIR 2006, David Evans[22] commenting on these figures claimed that they were due not so much to a lack of demand in the market-place but mainly to the special requirements of the real world context, not normally addressed by research efforts (Gey et al. 2006). Evans stated that demands on a commercial CLIR system included (a) automatic or semi-automatic adjustment to proper names and domain-specific terms; (b) retrieval of semi-structured information (such as tables); and (c) support for non-retrieval-specific applications such as portals, FAQ systems, and text mining. In addition, there is a greater need for end-user support, reflected in requirements such as translation or summarisation of retrieved information. From his experience as a supplier of enterprise multilingual support platforms, he felt that, at that moment, there was no viable business case for commercial CLIR. The complexity of a complete CLIR system, the difficulty of obtaining sufficient resources and of keeping them continually updated, problems of scalability and slow response times, and the need for intensive customer support meant that the costs of the system are much higher than the price that the customer was willing to pay.

[21] IDC is a global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets, see http://www.idc.com/

[22] CEO of Clairvoyance Corporation.

His conclusions were that:

- The market for multilingual globalisation support was *still* "not there yet";
- Quality and scope of MT is a *major* gating factor;
- The demand for CLIR, per se, is low. To be successful today, CLIR systems (already very complex) must be fashioned around 'solutions' – integrated into systems that may need CLIR functionality only as a means to other ends.

However, despite the slow growth of the CLIR market and the evident problems, in 2009, IDC made the following prediction "*Machine translation, globalization, and multilingual/cross-language applications and tools will grow. The growth of tools to address one of the information access and integration barriers — language — will be fueled by the need for the industrialized world to move into the emerging economies. Government investment in these technologies for terrorist and fraud detection will also spur new developments that will result in new enterprise and consumer uses as well.*"[23]

This expected demand will provide a major stimulus to research and development in the MLIR/CLIR area in the next decade or so, and is a primary motivation for this book.

## 1.4 The Current Research Challenges

There are two main challenges now facing our domain. (ML)IR is no longer just about text, today's content is increasingly multimedia and search paradigms are changing. The user today has different expectations and makes high demands; the tendency is no longer passive information seeking but rather dynamic interaction with content. Queries can be formulated using images and/or sound – not just text, and retrieved information may be in several media formats and in several languages. Future research must aim at satisfying these new requirements.[24] At the same time, we need to develop functionality and systems that are capable of meeting the demands of the market, i.e., facilitate transition from research prototype to operational system. In this section, we examine these two challenges, focusing on questions that concern CLIR as this is where the difficulties lie.

So far research has focused very much on the search problem, i.e., access and retrieval, from the technology viewpoint. To a large extent it can be claimed that

[23] IDC Predictions 2009.

[24] Think of an English tourist visiting south-east Asia and interested in traditional music and dance. An initial query in English finds preliminary information on dances in Cambodia, Vietnam and Laos. Some of the documents returned have pictures and music associated. The tourist uses these to find similar images and music and also reformulates the query in CLIR mode, specifying that they are interested in target documents in these three languages. The documents returned are no longer in English but are in the national languages accompanied by an MT gist in English.

this part of the CLIR problem is understood and (to a fair degree) solved. We know how to set about processing and indexing multiple languages, and we know the mechanisms that need to be deployed in order to match queries to documents over languages. Thus, at the search level, it is not so much the inherent difficulty of the problem that constitutes an obstacle but rather its vastness. There are a little over 2,000 languages which have a writing system,[25] although only about 300 have some kind of language processing tools. Clearly the implementation of a system that would accept queries in any of these languages and match them against documents in any other language(s) would require the deployment of an impossibly large number of language processing tools and translation resources of some type.[26]

Where research has been lacking so far is in the study of the implementation of CLIR technology from the user and the usage viewpoints. In order to produce better systems, we need better understanding of how the user addresses the cross-language information seeking task and what the real requirements are. We must implement systems that provide personalised search assistance according to the user's cultural expectations and language competence. We should also examine the possibility of faceted search and browse capabilities to provide better interaction with multilingual content. In addition, we need to work far more on the end results, on the presentation of the retrieved information in a form that is useful to and exploitable by the user. This last problem represents a serious obstacle to the take-up of MLIR/CLIR by the application communities. Although there has been an enormous improvement in MT systems in the last decade, performance levels can vary greatly and are still a long way from the style and accuracy achieved by a human translator. As has already been stated, for many languages there are still no good MT systems available.

Finally, we need to remember that a MLIR/CLIR system is never an end in itself but a component within a particular information seeking application – and the application is most probably multimedia. Thus much more research is needed on how to develop/engineer commercially viable search systems that meet the typical requirements of the average enterprise user:

- Search system must run on a single 'off-the-shelf' server;
- System must be easily integrated into the client's platform;
- The response times even for complex queries must be fast (<2 s);
- Scalability problems must be resolved (CLIR queries are typically several times larger than in monolingual search);
- Easy tuning of parameters to achieve precision;
- High quality translation of results and presentation according to the customers' requirements;
- The expected costs for customer support, integration and maintenance must be low.

[25] There are approximately 6,800 known languages in the world.

[26] If this problem is ever to be overcome, it implies a rethinking of the current mechanisms for CLIR and increased study of language-independent or conceptual mapping systems.

In addition, the necessary lexical and translation resources must be easy to acquire and easy to optimise to meet the demands of the domain to be covered. And last, but certainly not least, the cost of the system must be within the limits of the budget specified by the client.

## References

Ballestreros L, Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Proc. 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1998). ACM Press: 64–71

EMIR (1994) Final report of the EMIR project number 5312. Commission of the European Union, Brussels

Ferro N, Peters C (2008) From CLEF to TrebleCLEF: the evolution of the cross-language evaluation forum. In: Proc. NTCIR-7 Workshop Meeting, December 16–19 2008, NII, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/

FIRE (2008) First workshop of the forum for information retrieval evaluation. http://www.isical.ac.in/~fire/2008/working_notes.html

FIRE (2010) Working notes FIRE 2010, 19-21 February 2010, DAIICT, Gandhinagar. http://www.isical.ac.in/~fire/2010/working_notes.html

Gey FC, Kando N, Peters C (2005) Cross-language information retrieval: the way ahead. J. Inf. Process. & Manag. 41(3): 415–431

Gey FC, Kando N, Lin C-Y, Peters C (2006) New directions in multilingual information access. SIGIR 2006 workshop report. ACM SIGIR Forum 40(2): 31–39

Gey FC, Kando N, Karlgren J (2009) Information access in a multilingual world: Transitioning from research to real-world applications. ACM SIGIR Forum 43(2): 24–28

Grefenstette G. (ed.) (1998) Cross-language information retrieval. The Kluwer International Series on Information Retrieval, Kluwer Academic Publishers, Boston

Harman D (2003) The development and evolution of TREC and DUC. In Proc. 3rd NTCIR workshop on research in information retrieval, question answering, and summarization. NII, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/

Harman D, Braschler M, Hess M, Kluck M, Peters C, Schäuble P, Sheridan P (2001) CLIR evaluation in TREC. In Peters C (ed.) op.cit.: 7–23

Harman D, Kando N, Majumder P, Mitra M, Peters C (eds.) (2010) Special issue on Indian language information retrieval. ACM Trans. Asian Lang. Inform. Process. 9(3)

Hovy E, Ide N, Frederkin R (eds.) (1999) Multilingual information management: current levels and future abilities, NSF/EC/DARPA, http://www.cs.cmu.edu/~ref/ mlim/index.html

ISO (1985) ISO Standard 5964-1985: Guidelines for the establishment and development of multilingual thesauri. First edition 1985-02-15. International Organisation for Standardisation, Technical Committee ISO/TC 46

ISO/IEC (1993) ISO/IEC International Standard 10646-1:1993(E): Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and basic multilingual plane. International Organization for Standardization, Geneva 1993

Kando N (2001). Overview of 2nd NTCIR workshop. In: Proc. 2nd NTCIR workshop on research in Chinese and Japanese text retrieval and text summarization, Tokyo, May 2000–March 2001. NII, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/overview-kando.pdf

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S, Adachi J (1999) The NTCIR workshop: the first evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. In: Proc. 4th international workshop on information retrieval with Asian languages (IRAL'99), Nov. 11-12, 1999, Taipei, Taiwan

Kando N, Mitamura T, Sakai T (2008) Introduction to the NTCIR-6 special issue. ACM Trans. Asian Lang. Inform. Process. 7(2): 1–3

Oard DW (ed.) (2003) The surprise language exercises. ACM Trans. Asian Lang. Proc. 2(3-4): 79–84

Oard DW, Gey FC (2003) The TREC-2002 Arabic-English CLIR track. In: The eleventh text retrieval conference. TREC 2002. NIST special publication 500-251: 17–26

Peters C (ed.) (2001) Cross-language information retrieval and evaluation. 1st workshop of cross-language evaluation forum, CLEF 2000. Springer LNCS 2069

Salton G (1971) Automatic processing of foreign language documents. Prentice-Hill: Englewood Cliffs, NJ

Schäuble P, Smeaton A (1998) An international research agenda for digital libraries: Summary report of the series of joint NSF-EU working groups on future directions for digital libraries research, 1998. http://www.ercim.eu/publication/ws-proceedings/DELOS-B/dl_sum_report.pdf

# Chapter 2
# Within-Language Information Retrieval

*"The idea that you can index billions of pages and look for a word and get what you want is quite a trick."*

Terry Winograd, Ubiquity 2002

**Abstract** The information retrieval system stands at the core of many information acquisition cycles. Its task is the retrieval of relevant information from document collections in response to a coded query based on an information need. In its general form, when searching unstructured, natural language text produced by a large range of authors, this is a difficult task: in such text there are many different valid ways to convey the same information. Adding to the complexity of the task is an often incomplete understanding of the desired information by the user. In this chapter, we discuss the mechanisms employed for matching queries and (textual) documents within one language, covering some of the peculiarities of a number of widely spoken languages. Effective within-language retrieval is an essential prerequisite for effective multilingual information access. The discussion of within-language information retrieval or monolingual information retrieval can be structured into two main phases: the indexing phase, commonly implemented as a pipeline of indexing steps, producing a representation that is suitable for matching; and the matching phase, which operates on the indexed representations and produces a ranked list of documents that are most likely to satisfy the user's underlying information need.

## 2.1 Introduction

Information Retrieval (IR) systems are part of a larger information acquisition cycle. Systems that process information in multiple languages (either queries, documents, or both) are called Multilingual Information Retrieval (MLIR) systems. It is fair to assume that interaction with the IR system proper is only part of a larger process initiated by the user to acquire information necessary to solve a problem or satisfy an information need. A software solution that covers this whole process of accessing information in response to an information need is called an *Information Access application*, or *Multilingual Information Access (MLIA) application*, when multiple languages are involved.

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_2, 

In information retrieval, users try to satisfy their information needs by querying an unstructured data collection, potentially fed with information from very disparate sources.[1] Obvious examples include the World Wide Web, where search services cover content stored on millions of different servers, but also enterprise search applications that interface with a wide range of content management systems and databases inside a company. Nowadays, such services and applications often offer access to information written in many different languages.

Typically, the user's understanding of a problem leading to an information need is incomplete at the outset (looking for a 'solution to a problem', the user can normally only formulate their understanding of the problem, not the outlines of the solution itself). In contrast with query mechanisms in (relational) databases, no schema guides the user in understanding the structure and content of the document collection that is made accessible through the IR system. The system is thus used in an iterative process, either directly by the user, or through a larger information access application. During this iterative process (illustrated in Figure 2.1), users gain a better insight into possible solutions, influencing their ability to clearly state their needs.

Let us consider a small example in order to better identify how we define the role of the information retrieval system (see also Braschler and Gonzalo (2009)). Suppose the user wants to know how to best archive digital photographs. Approaching the information access application, the user faces a dilemma when trying to express the information need: relevant documents will present a solution to the problem, e.g., descriptions of suitable archival tools, but they do not necessarily contain descriptions of the questions that may lead to this solution. It would thus be preferable for the user to provide the system with search terms that are associated with the solution, instead of terms that are contained in possible questions.[2] This is a contradiction to the assumption that the user searches precisely because the solution is unknown. The task of query formulation is even more difficult for the user if forced to use an unfamiliar language. Users should thus attempt to put whatever prior knowledge of a possible solution they have into words; this is referred to as verbalisation (Tague-Sutcliffe 1996). We assume that the user realises from previous experience they want to look into a 'CD burner' as an archival solution. The system will present some form of user interface where the user can enter queries. While systems built according to the mechanisms described in this chapter typically

[1] This is what distinguishes information retrieval systems from Database Management Systems (DBMS), where correctly structured data is stored according to a well-defined database schema and where users select subsets of this data according to exactly specified criteria.

[2] This issue is especially important on comparatively 'smaller' document collections. The World Wide Web is a special case in that its massive size and the high redundancy of its content helps with locating some relevant documents for nearly all queries. In this sense, searching on smaller collections is often harder – the task of the user to pick good search terms is considerably more complex, and the IR system should thus be able to match on a wider range of possible query formulations.

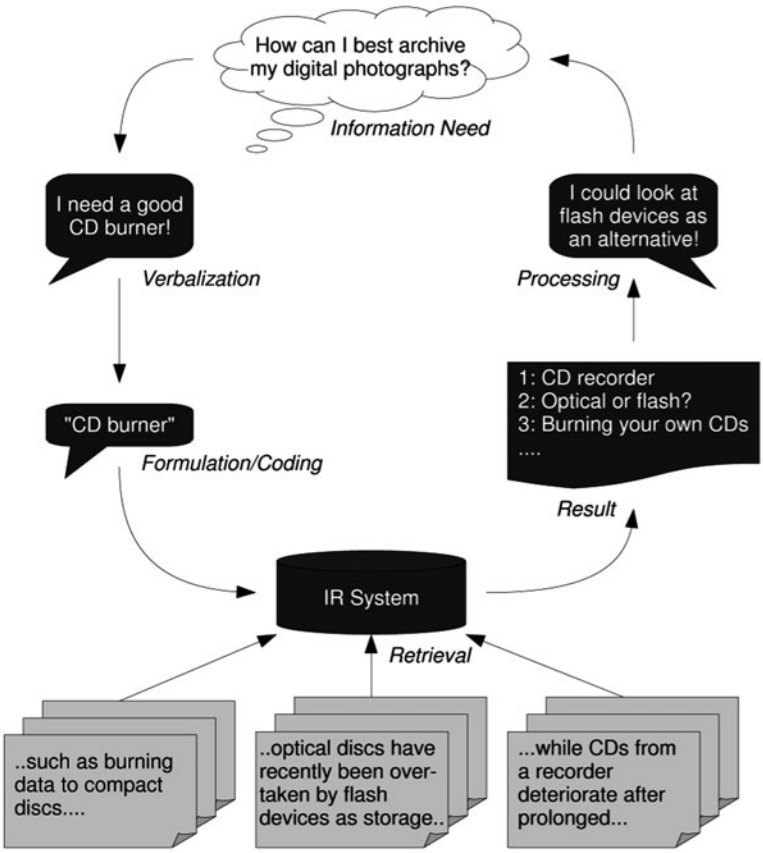


**Fig. 2.1** The information retrieval system is part of a larger information acquisition cycle

allow the use of full natural language queries, many users are conditioned to input a (short) sequence of keywords.[3] Our user types 'CD burner' and starts the search.

The output of the information retrieval system is now judged by the user on how useful it is to find a CD burner, but ultimately also on how it matches the underlying desire to find a good solution to archive the photographs. The task of the information retrieval system is thus not simply to match documents to the query, but to satisfy the information need. The query itself may have been wider, narrower or even differently focused than the task that the user really wants to solve. In our example, possible solutions present in the document collections may be formulated in a variety of ways: documents can refer to 'CDs', 'compact discs', 'optical discs',

[3] One likely reason for the use of very short queries is the prevalent use of web search services such as Google and Yahoo, which by default narrow search results by using an implicit 'AND' operator for query terms.

maybe also to 'DVD'. The 'CD burner' may be called 'CD recorder' or 'CD-R drive', among other possibilities. And of course, there may be relevant information in documents that do not match the specific query chosen by the user at all. For example, documents may talk about 'flash memory' devices as an alternative for data storage. Lastly, for some queries, cultural differences when crossing language boundaries may also contribute to a disconnect between the query formulation and the content of relevant documents. Processing the list of results compiled by the system, the user will gain a better understanding of the problem, and will be able to re-verbalise and re-formulate the information need – thus enabling a new cycle through the information acquisition process.

The remainder of the chapter will focus exclusively on the (multilingual) information retrieval system proper, i.e., the component that accepts queries as input and produces a set of documents that best match the query formulation according to well-defined matching criteria. In our discussion, we will mainly address systems that accept the query as a sequence of terms (words), and produce the result set in the form of a list of documents, ranked by an estimate of the probability of their relevance with respect to the query. In this chapter, we will present the mechanisms necessary for building a within-language (monolingual) retrieval system, namely indexing and matching. Effective monolingual information retrieval in all languages to be handled by the system has been shown to be one of the preconditions for implementation of successful multilingual information retrieval systems (Braschler and Peters 2004). The discussion will provide details on the adaption of the mechanisms to different languages where appropriate. Information on how to integrate translation components into the IR system in order to support Cross-Language Information Retrieval (CLIR) can be found in Chapter 3. The chapter on user interfaces (Chapter 4) addresses many of the other main points of the information acquisition cycle and information access applications.

## 2.2 The Retrieval Problem and Its Consequences

As stated, the task of the retrieval system is the retrieval of (all) relevant items (often in the form of providing the user with a ranked list of documents) in response to a coded information need (the 'query'). Complicating matters, 'relevance' is a subjective notion that can be influenced by factors not explicitly available in either the document or the query (such as user's preferences, prior knowledge, etc.).[4] This task is often referred to as the (document) retrieval problem.[5]

[4] We discuss the concept of relevance and its implications in more detail in Chapter 5.

[5] One of the early definitions of the 'document retrieval problem' is given by Robertson et al. (1982): *"…. the function of a document retrieval system is to retrieve all and only those documents that the inquiring patron wants (or would want)."*

Two main issues need to be addressed:

1. How to match an often incomplete, or even inadequate formulation of the underlying information need with a wide variety of different possible paraphrasings by different authors of information that is relevant to this need; and
2. How to weigh and rank the different (sometimes partial) matches that are obtained.

Considering the case of document or text retrieval, simply returning all documents that contain exact representations of all the search terms in a query[6] is clearly not sufficient, as neither of the two aspects is properly addressed: the query supplied by the user could easily under-specify (miss essential search terms) or over-specify (contain ill-suited search terms) the desired information need. Nor is it guaranteed that relevant documents would necessarily contain all or even any of the search terms as provided. In the multilingual case, translation ambiguities will be another source of complication. A probabilistic view addresses these concerns: the system calculates an estimated probability of relevance for every document with respect to the query supplied. Consequently, there are two main goals for an IR system:

1. To maximise the likelihood that the query matches relevant items; and
2. To produce the best possible ranking of items according to their estimated probability of relevance.

The two goals show a clear parallelism to the two issues identified through the definition of the retrieval problem. A theoretical grounding can be found in the *Probability Ranking Principle (PRP)*.[7]

An information retrieval system has to cope with partial matches, both in terms of allowing inexact matches based on single query terms, and by ranking documents that only contain a subset of the query terms. While both these steps directly address point 1, i.e., the maximisation of the likelihood of actually retrieving a relevant document, they also often lead to increased retrieval of items that are not relevant to the user formulating the query (irrelevant items, 'noise'). Ranking addresses this problem to some degree, by sorting the set of retrieved documents according to the estimated probability of relevance. However, considering the possible difference between the underlying information need and the actual query formulation, and considering the probable wide range of alternative formulations for the same information contained in the documents, no ranking mechanism can in practice guarantee that only relevant items will be presented to the user. The question as to how to measure the effectiveness of an information retrieval system

[6] As could be done with a database query.

[7] Defined as: "*If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data*." (Robertson 1977).

is thus a central one, and all steps taken to ensure retrieval effectiveness need to be properly evaluated. The chapter on evaluation (Chapter 5) addresses these questions in detail. For the time being, it is sufficient to introduce the two evaluation measures that are most widely used to assess retrieval effectiveness: *precision* and *recall*. These two measures are based on the assumption that the user wants to retrieve as many relevant documents as possible, while at the same time retrieving as few irrelevant documents as possible, and are thus directly compatible with our understanding of the retrieval problem.

To compute precision and recall for any given set of documents, the following definitions are used:

Precision = number of relevant documents in the set/total number of documents in the set;
Recall = number of relevant documents in the set/total number of relevant documents in the collection.

To adapt these two set-based measures for use on ranked list output by IR systems, they can be computed continuously after each rank, starting with the top-ranked document (set of size 1), and then expanding the size of the set by one document in every step of the computation. It is often desirable to optimise for both measures, but the measures in practice are often contradictory: optimising for recall equates to maximising the likelihood of a match between query and (relevant) document, which often requires increasing tolerance towards partial matches. This in turn leads to more noise in the form of irrelevant documents introduced into the retrieval results. Optimising for precision conversely indicates a conservative matching strategy, increasing the danger of not retrieving some relevant documents. For all further discussion of these issues, see Chapter 5, and specifically Section 5.2.6 for details on the computation.

## 2.3 Implementation of a Within-Language Information Retrieval System

Structurally, the discussion of the implementation of a within-language (monolingual) IR system, taking coded queries as input, and providing lists of ranked documents as output, can be separated into two distinct main phases:

1. Indexing phase,
2. Matching phase.

A third, additional phase
3. Translation phase

is normally crucial for the cross-language case, and is described in Chapter 3.

This separation of the discussion is motivated from the considerations outlined in Section 2.2 as follows: the *indexing phase* prepares the retrievable items

(documents) and the formulation of the information need (query) for matching, addressing the desire to maximise the likelihood of obtaining a match between query and relevant documents (producing an indexed document representation and query representation, respectively); the *matching phase* (also called retrieval phase) ranks the documents according to a function of the estimated probabilities of relevance (the 'retrieval scores', or 'Retrieval Status Values' (RSV)) and collates result lists, addressing the desire to optimise these lists for high precision and recall (see Figure 2.2). There are different options that can be adopted when including the translation phase for cross-language systems; these will be described in Chapter 3.

The indexing phase derives its name from its role in preparing the retrievable items (documents) for their inclusion in a searchable index.[8] All retrieval systems designed for ranked retrieval on large datasets use some form of an index which is pre-built and is continuously or periodically updated. Searching large datasets is not practicable without an index, as the search would require linear scans at query execution time, with the corresponding run time for such a scan quickly becoming prohibitive. Once an initial index has been built, IR systems can then use the

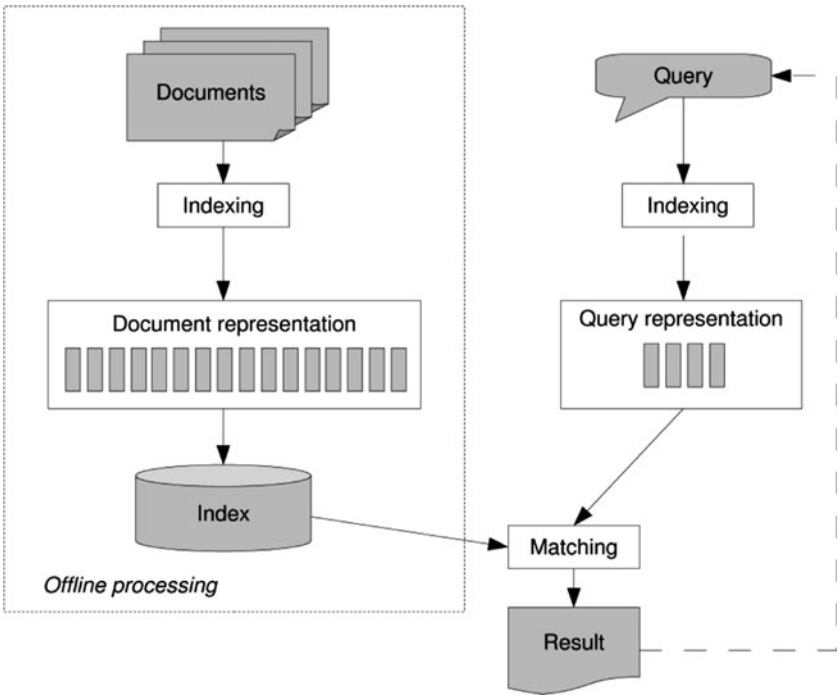


**Fig. 2.2** Basic conceptual architecture of an information retrieval system

[8] This is essentially an arbitrary data structure, built for efficient access to retrievable items. Details on the implementation of an index for an IR system are given in Section 2.5.2.

information stored in the index to look up features from the user's queries and calculate a score for each matching document (matching phase, see Section 2.5).

The mechanisms used in the indexing phase often need to be adapted to the characteristics of different languages. We have highlighted some of the most prominent issues in this regard; it is however clearly not practical to cover all possible implications for every language. Care needs to be taken on a case-by-case basis at system design time. The mechanisms described for the matching phase are by and large language-independent.

We will restrict the discussion to systems designed for text retrieval. Many of the mechanisms discussed in this section can be adapted for multimedia retrieval, which will be discussed in Chapter 6. When we talk about 'indexing' in the following, we are talking about building a *full-text index* of the documents, making documents accessible via each of their content-bearing words – this task should not be confused with cataloguing documents via the use of descriptors and then building an index of this catalogue.

## 2.4 Indexing Phase

During the indexing phase, a representation of the documents and queries suitable for storage in the system and for subsequent retrieval is built in successive steps through application of indexing components. Indexing of documents is carried out 'offline' – i.e., usually independently of query processing.

Apart from being a practical necessity to implement an *efficient* retrieval system, indexing is also a crucial aspect of *effective* information retrieval, since there is a large degree of freedom in how authors encode information in unstructured or semi-structured (textual) documents and in how searchers express their information needs. Information retrieval systems have the task of matching queries and documents on the basis of these different formulations of the same information.[9] Linguistic phenomena tied to this task include but, depending on the language, are not necessarily limited to:

- Morphology
- Synonyms
- Homonyms
- Paraphrases
- Metaphors
- Spelling variants and errors
- Transliteration

[9] This is in contrast to databases, where every effort is made to store data in a normalised form.

These linguistic problems are addressed by various *normalisation*, *reduction* and *expansion mechanisms* during the indexing process. It is theoretically possible to apply some of the mechanisms detailed below 'on-the-fly' during matching; however, in practice, this is almost universally avoided due to performance considerations.

The retrieval system can only retrieve items in the matching phase based on the representation of the items in the index; it is thus evident that information that is lost, or garbled, during the indexing phase can no longer contribute to the item's score during matching, and may effectively make it impossible to retrieve this item in response to a given query. Implementing a careful, effective indexing process is consequently of high importance.

Indexing is often implemented in the form of an 'indexing pipeline', where a series of sequential steps is executed to obtain 'features' that are then used to build the index structure proper. The most common indexing steps are:

1. Pre-processing;
2. Language identification;
3. Document formation;
4. Segmentation, tokenisation, parsing;
5. Feature normalisation;
6. Enrichment.

### *2.4.1 Pre-processing (Step 1)*

Information retrieval systems nowadays often fill a role analogous to 'data integration', by providing users with a single interface to access information from many sources, such as multiple different databases, intranet content, personal files and folders and other data collections ('integration at search time'). Pre-processing ensures that items from these sources are compatible with subsequent indexing steps. Most prominently, inconsistencies in the encoding systems used by different sources can arise, especially when multiple languages, potentially using non-Latin scripts, are involved. Language-specific character encodings such as KOI-8 (Russian), Shift JIS (Japanese) and GB2312 (Simplified Chinese), among others, remain widespread, and potentially need to be handled if the system is to cover these languages. Character set handling issues can be largely avoided in later steps if all documents are converted to a common character encoding that encompasses all languages of interest and their properties. Unicode seems the obvious choice to do this, and also harmonises well with XML.[10] If keeping exclusively within English

[10] Unicode is the current *de-facto* standard for character encoding of written text for applications that need to cover many diverse languages (www.unicode.org). XML is a text format intended for the creation of structured (text) documents that are machine-readable (http://www.w3.org/XML/).

and other Western European languages, the use of some of the ISO-8859 code sets, especially ISO-8859-1, may also be possible and more space efficient. The issues in converting between the different encodings fall outside the scope of this book.[11]

None of the methods or algorithms discussed in the following sections makes specific use of properties of particular file formats (such as the Portable Document Format (PDF) or the Microsoft Word document format). Typically, if used at all, only the most basic structural information from such formats (such as titles, headings, etc.) is processed. Structural information on this level can easily be encoded in XML. Note that this represents no significant restriction on the possibilities available for document presentation to the user; the use of an XML representation of the documents can be restricted to the processing by the retrieval system, presenting the documents that have been retrieved in their original format.

Additional pre-processing is possible at this stage. Most notably, such pre-processing can include:

- Removal of non-content bearing sections (headers, footers, navigational areas, copyright messages, etc.). These can lead to spurious matches during later retrieval. These sections can still be shown when the retrieved document is presented to the user.
- Duplicate or near-duplicate detection (fingerprinting). Duplicate detection typically uses a checksum mechanism[12] for easy comparison of documents. The checksum method however breaks down when documents undergo many very small revisions as is common with dynamic content on the World Wide Web, where, for example, documents may be supplemented with a time-stamp that is recalculated at each access. Detection of very similar documents ('near-duplicates') essentially requires an effort that increases quadratically with the size of the document collection.[13] Fingerprinting and sketching[14] methods can help to reduce this computational complexity (Broder 2000).

### 2.4.2 *Language Identification (Step 2)*

Some of the processing in later indexing steps, such as tokenisation, stopword elimination and stemming (see Steps 4 and 5) is usually language-dependent.

[11] For more details, see e.g., Erickson (1997).

[12] The checksum is typically a numerical value that acts as a 'summary' of the underlying data and that is sensitive to (slight) changes therein. While two random pieces of text may share the same checksum, it is highly improbable that a text would still generate the same checksum after modification when using an appropriate algorithm for checksum calculation. Exact duplicates of documents will share the same checksum.

[13] Conceptually, every document needs to be compared to all other documents in the collection.

[14] A short, usually fixed-length, representation of the document is used that, unlike checksums, allows for similarity comparisons.

In cases where there is no clear indication of the language of a given document or query (e.g., through metadata fields or user profile information), a *language identification component* needs to be employed.

Major approaches to language identification include:

- Detection based on stopword usage;
- Detection based on character *n*-gram frequency;
- Hybrid methods.

The concept of stopwords will be properly introduced in Step 4. For the purpose of language detection, stopwords are defined to be the most frequently used words (or word forms) as determined from representative samples of a language. These lists of frequent words will for most languages contain many determiners, conjunctions, prepositions and interjections, most of them comparably short words. The use of a frequency analysis based on these lists has been shown to be effective for language detection.

Alternatively, methods based on character *n*-grams can be used. The concept of character *n*-grams will be more fully detailed in Step 5. To form character *n*-grams, the text drawn from a representative sample for each language is partitioned into overlapping sequences of *n* characters. The frequency counts of the resulting *n*-grams form a characteristic 'fingerprint' of the language, which can be compared to a corresponding frequency analysis on the text for which the language is to be determined (Dunning 1994).

Hybrid approaches apply variants of both the stopword and character *n*-gram methods sequentially, and assign languages to text on the basis of both outcomes.

### 2.4.3 Document Formation (Step 3)

Documents stored in information retrieval systems can vary considerably in length. A single document collection indexed by a system can, for example, contain both lengthy newspaper articles and very short news headlines. The weighting schemes discussed in Section 2.5 can handle variations in document lengths to different degrees, although it is sometimes unclear how they have been optimised with respect to this issue. It can thus be prudent to choose a smaller or larger granularity than the entire document for retrieval, both to address retrieval effectiveness, but also to improve the presentation of results to the searcher. The choice of a different granularity can be especially interesting if single documents address multiple different topics. There are scenarios where systems need to operate on the sub-document (passage, paragraph or sentence) level and scenarios where the focus is on the super-document (set of documents, folder or linked documents) level. In some cases, such scenarios can be accommodated by splitting or merging documents prior to indexing. If this is not the case, the architecture of the system, and its processing of the documents needs to be adapted.

Ways to handle indexing at a sub-document level include (Kaszkiel and Zobel 1997):

- Fixed-length passages (documents are split in chunks containing a fixed number of tokens each, either with or without overlap).
- Paragraph boundary detection (potentially using structural information, such as XML mark-up).
- Sentence boundary detection.

Indexing at the super-document level may include:

- Collating text from several hyperlinked documents.
- Enriching the text of a document by including information from metadata, annotations or comments to the document, or from 'anchor text'[15] in linked documents.

### 2.4.4 Segmentation, Tokenisation, Parsing (Step 4)

Before retrieval, documents need to be segmented into shorter units, in order to allow matching with the query (which typically is either a short natural language description or a set of keywords). For most European languages, the obvious choice is segmentation into words (*word segmentation*). It should be noted, however, that while the term 'word' has a very specific definition in linguistics, in information retrieval the meaning of 'word' can be more ambiguous. As we will shortly discuss, it is not always clear what are the best 'words' from a retrieval perspective. A more neutral way to discuss these issues is to talk of 'terms' or 'features', when referring to the units that are ultimately output by the information retrieval system after Step 4. If a system is focused on retrieval of text, 'term' is often used to describe these units, while 'feature' is a good choice if non-textual content is also covered in the discussion. Both terms and features occur as a stream of tokens, an ordered list of units output by the segmentation or tokenisation component. Step 4 covers the methods to produce valid units for retrieval, where the terms or features are usable for subsequent retrieval. Most index structures in information retrieval systems do not allow matching on parts of terms or features – only full, exact matches of terms or features are possible. It is therefore crucial to produce not only a valid set, but the 'right' set of features that leads to the best possible matches and therefore to maximum effectiveness during retrieval. This aspect will be covered in more detail in Step 5.

The easiest option to produce a valid stream of tokens from text written in European languages is to segment using whitespace characters (space, newline, tabulator, etc.). All the characters between two sequences of whitespace characters are treated as a token. However, this is usually not a good option, as special

[15] That is, the text underlying a hyperlink that is clicked by the user to follow the link.

characters such as punctuation marks are retained after such processing. These characters can prevent later matches. An obvious solution to this problem is the restriction of tokens to sequences of alphanumeric characters. Care is needed to ensure that the definition of character classes used for this partition is appropriate for the languages to be handled (e.g., covering the necessary characters with diacritical marks).

There may be the need for more sophisticated processing if the system is to be robust especially with regard to named entities. By restricting tokens to alphanumeric characters, a number of issues arise. Potential terms such as 'O'Brien', 'F/A-18', 'Coca-Cola', 'Yahoo!Mail' and others are split into multiple tokens. If such splits are to be avoided, either a dictionary of named entities or more sophisticated linguistic processing is needed.

The situation is more complicated in some Eastern Asian languages, notably the most widely used languages in the region, Chinese and Japanese. In these languages, whitespaces are generally not inserted between 'words', with whole sentences written as continuous strings of characters. Analogous to the treatment of text in European languages, these sentences must be split into units suitable for retrieval. This is not a simple task, as for non-trivial sentences there are often multiple different possible splits into individual words, with the wrong choice making later matching impossible. Literature dealing with the problem usually recommends one of two basic alternatives: the use of word *n*-grams[16] or the use of a specialised segmentation component. In Chinese, the characters ('logograms') represent 'basic concepts', and each word in the Western sense can potentially be represented by a number of logograms. Often it is possible to infer the meaning of a word that is expressed by multiple logograms from the meaning of the individual underlying characters. The simplest strategy is therefore to use single Chinese characters as the unit for retrieval, but since there is important additional meaning encoded in the character combination, the effectiveness of such systems is usually not optimal. Another solution is the use of bigrams of Chinese characters, i.e., overlapping pairs of characters. The use of single characters for retrieval is consequently also called 'unigram' indexing. Unigram and bigram strategies can be combined, with the segmentation component outputting a stream of both unigrams and overlapping bigrams.

As an alternative to the use of word *n*-grams, word segmenters are available for Chinese. These attempt to find the most plausible splitting of a sentence into Chinese words of arbitrary length. Segmenters adapted for use in IR systems do not necessarily need to produce a linguistically correct segmentation for effective retrieval. It may well be that the more simplistic word *n*-gram method allows interesting conflations of related words that would otherwise be represented by longer character strings. Similar effects can be observed when using stemming and decompounding for European languages (see Step 5).

[16] While structurally the same as character *n*-grams for Latin script, as used in Step 5, they are named differently to account for the different role of Chinese characters.

When comparing the use of word *n*-grams to the use of a word segmenter for Chinese, the unigram + bigram combination can be competitive with full word-based segmentation (Abdou and Savoy 2006).

Japanese uses characters originally borrowed from Chinese ('Kanji') as well as two additional syllabaries, namely 'Hiragana' and 'Katakana' plus the Latin alphabet. As in Chinese, sentences are written as continuous strings of characters. Segmentation can take place in the form of unigrams, bigrams, combinations of unigrams and bigrams or can be word-based, again analogous to Chinese. There are also reports of competitive performances for the combination of unigrams and bigrams compared to full word-based segmentation (Savoy 2005).

Most information retrieval systems contain a component that removes *non-content bearing tokens* (also known as 'stopwords' or 'stop words') from the segmented document stream. Historically, the use of such a component stems from the observation that frequency counts of words are very uneven in document collections. Very few words are used extremely often, while most words are used only very rarely (a rule also referred to as Zipf's law). This distribution has been experimentally verified for many different languages.

When processing the tokens during indexing, a substantial part of the tokens can thus refer to very few different words. For example, when processing a full year of articles from the German newspaper *Frankfurter Rundschau*, roughly 800,000 unique features (word forms) were found. However, the ten most frequent features make up 16.6% of tokens, and the top 50 most frequent features cover 33% of tokens. In contrast, a clear majority of features (nearly 500,000) occur only once or twice in any of the articles. Most of the highly frequent features are determiners, conjunctions, prepositions, interjections and the like (see Table 2.1 for example lists of highly frequent words in German and English).

It can be argued that most of the meaning of a text is retained even when these words are deleted,[17] and thus historically, in the development of information retrieval systems, they were eliminated primarily in order to reduce the size of the resulting index. The system uses a list of these 'stopwords' for this purpose. Today, it can be argued that generally index size is no longer a pressing concern, and stopword elimination purely to reduce the index size should no longer be necessary.

Apart from reducing the number of tokens to be indexed, stopword elimination has also been shown to be beneficial for retrieval effectiveness as measured by recall and precision. This result should be interpreted carefully. Many academic experiments report average performance over a number of queries. A small increase in average performance may well hide a performance regression for a substantial number of individual queries. Furthermore, even though stopwords are often termed

[17] This argument is obviously not without problems as very frequent words such as 'not' can easily change the meaning of a sentence to the contrary. However, the weighting approaches discussed in Section 2.5 are based mainly on frequency counts, and do not employ deeper semantic analysis of text.

**Table 2.1** The ten most frequently used words for English and German.[a] These words would be very likely candidates for inclusion in stopword lists

| English | German |
|---|---|
| the | der ('the') |
| to | die ('the') |
| a | und ('and') |
| of | in ('in') |
| and | den ('the') |
| in | das ('the') |
| for | von ('by') |
| that | mit ('with') |
| is | im ('in the') |
| said | zu ('to') |

[a] As determined by the processing of 1 year's worth of newpaper stories for each language (published by the *Los Angeles Times* and *Frankfurter Rundschau*, respectively). For the German list, the corresponding most frequent translation to English is given in parentheses.

'non-content bearing tokens', clearly *any* elimination of tokens from a text leads to information loss, however small. Often very frequent words such as 'the' and 'who', which in many circumstances may not be essential for human comprehension of a sentence, can carry critical importance in special contexts, for example when looking for information on the rock band 'The Who'. The fact that certain weighting schemes benefit from stopword elimination should therefore be seen more as a deficiency of such weighting schemes (by not properly handling the stopwords) rather than as proof of the value of stopword elimination to boost retrieval effectiveness. Recent work investigating which weighting schemes are particularly robust with respect to lack of stopword elimination suggests that there are significant differences between different schemes. By choosing a robust weighting scheme stopword elimination can be avoided in many scenarios. Alternatively, the stopword list can be kept as short as possible (Dolamic and Savoy 2009).

### *2.4.5 Feature Normalisation (Step 5)*

Once valid candidates for retrievable features have been identified, it is possible to normalise them further, in the hope of enhancing their potential to enable matching between query and documents. As natural language provides many different ways of conveying the same information, it is unlikely that the formulation of their information needs by users will always match the terms used by the authors of the documents containing relevant information. The main reasons for mismatches are the use of synonyms or of alternative surface forms of the same lexeme, and different applications of writing conventions, e.g., in the use of diacritics, capitalisation, spelling, etc. Problems of synonymy are usually handled by

**Table 2.2** Example rules from the different steps of the Porter stemmer (cited from Porter 1980). The measure *m* counts the number of vowel-consonant sequences in a word. As can be seen from the examples given for step 4 and step 5, rules can produce strings that do not constitute valid English words

| Example rule | Example result of rule |
|---|---|
| Step 1: SSES → SS | caresses → caress |
| Step 2: (m > 0) ATOR → ATE | operator → operate |
| Step 3: (m > 0) NESS → | goodness → good |
| Step 4: (m > 1) IBLE → | defensible → defens |
| Step 5: (m > 1) E → | probate → probat |

employing some form of query (or even document) expansion, e.g., by employing a thesaurus. Such methods would typically be implemented in Step 6.

Much research has been conducted on addressing the issue of differing word surface forms as, depending on the language, the number of such forms for the same lexeme can potentially be very high, due to variations in grammatical gender, number, case, etc. There are two basic word formation processes in play: inflection and derivation.

*Stemming* is an attempt to minimise mismatches between queries and documents due to the use of differing word forms. A stemming component will remove common suffixes or prefixes from words, thus approximating the lemma (i.e., the 'base' or dictionary form) of that word. It is not usually a requirement for a stemmer to correctly identify the lemma, or indeed to produce even a valid word; instead the focus is on obtaining a conflation effect and thus optimising retrieval performance. For many languages, stemmers are built based on a set of rules for affix removal, while in some languages stemming only becomes practicable after including dictionary resources as well.

To illustrate the workings of a rule-based stemmer, in Table 2.2 we present a small sample of rules taken from the widely-used Porter stemmer for the English language[18] (Porter 1980). The Porter stemmer removes suffixes in five, iterative steps from arbitrary character strings assumed to be English words.[19] In total, the stemmer uses roughly 60 rules, with each rule operating on a combination of a set of preconditions, a suffix that is removed, and optionally a new suffix that is attached. Many rules require a minimal length of the character string for the rule to be applied, with a measure of the number of vowel-consonant sequences in the string used to this end.

While it seems intuitive that some normalisation of word surface forms to common representations is desirable given the likelihood of a mismatch between query and document, an analysis of the problem shows that care is needed. While

[18] Adaptations of the Porter stemmer to many different languages are today available.

[19] In particular, there is no attempt to detect names and foreign words, which are thus potentially stemmed as well should they match the rules.

the meaning of a word in many cases does not change between different inflected forms, there are situations where there is a shift, and where the difference is crucial. In some cases, the user would explicitly ask for something in the plural, knowing that any occurrence in the singular of the same concept is not likely to lead to relevant information. Normalisation of word forms originating from derivation can also be misleading. For example, the word 'formation' may be derived from 'form', but it is not immediately clear if documents containing the latter word would be helpful when the user is searching for information related to the former concept.

In academic IR literature, the terms *overstemming* (conflation of two word forms or words which is detrimental to retrieval effectiveness; too 'aggressive') and *understemming* (failure to conflate two word forms or words which is detrimental to retrieval effectiveness; too 'conservative') are often used. A good stemming component for information retrieval strikes the right balance in word form reduction to maximise retrieval effectiveness. Questions of acceptance of stemming by the user are not well researched. Stemming often leads to artificial, truncated representations of words, and these are not always readily recognisable by the user as desirable for retrieval (e.g., as demonstrated, the Porter stemmer for English suffers from this phenomenon. A further example is the conflation of both 'initial' and 'initiative' to 'initi', which is not an immediately recognisable character string for users. Due to overstemming in this example it is hard to relate back to the underlying unstemmed word form). While stemmed forms can often be hidden from the view of the user by using them exclusively for internal representation, there are cases where the user is prompted for interactive feedback on search terms during retrieval (such as in relevance feedback, see Section 2.5.6). In such cases stemmed representations may not be usable.

Stemming has been shown to be effective for a large number of languages, and in particular for those languages with a complex morphology, both inflectional and derivational. Braschler and Ripplinger (2004) give a short overview of work related to stemming and cite an incomplete list of languages where stemming has been shown to be beneficial: Slovene, Hebrew, Dutch, German, Italian and French. Even though there is extensive work available on stemming in English, results on the effectiveness of stemming for that language are inconclusive, most likely due to English having a comparatively simple morphology. At most small benefits have been reported (Harman 1991, Hull 1996). Depending on the application, it may thus be beneficial to allow expert users to toggle the stemming function on and off, to allow control for understemming and overstemming effects. Stemming can be detrimental to retrieval efficiency, which may be a concern in systems with massive processing loads.

For the Arabic language, 'stems' can be reduced further to 'roots'. While an Arabic stemmer will typically remove both prefixes and suffixes from a word, infixes are also removed to obtain the root. Early experiments in Arabic information retrieval seemed to indicate that this further reduction leads to improvements in retrieval effectiveness, but later experiments on larger test collections did not validate this result, instead showing a better performance using stems. Using roots may be an interesting option if recall has to be optimised (El-Khair 2007).

An issue closely related to stemming is that of *decompounding*. A number of languages, such as many Germanic languages (German, Swedish, Dutch), Finnish and Korean, among others, make extensive use of a word formation mechanism where new compound words are formed from multiple 'basic' words, and where the compound word is written with no whitespace separating the constituents. It is often possible to express the same concepts using a phrasal expression. This type of variation in formulation of the same concept, coupled with a potentially infinite number of such compound words, presents an important challenge to IR systems handling these languages: noun compounds in particular often carry important, unambiguous meaning, and a mismatch between compound and phrasal expressions containing the same constituents results in significant degradation of retrieval effectiveness. A decompounding component is used to split compound words into their constituents during indexing. Decompounding is usually applied both for queries and documents.

A similar compound formation mechanism exists in the English language, although less frequently (e.g., airplane from air + plane, software from soft + ware). Furthermore, such compound words in English are usually lexicalised, and often acquire a meaning that no longer directly derives from the sum of their constituents (such as 'software'). In general, the phenomenon is not of much concern for English language retrieval.

When compounding is an important feature of a language, such as in German, good handling of compounds becomes important for effective retrieval. Multiple authors have demonstrated substantial gains by splitting compounds (see e.g., Braschler and Ripplinger 2004). Decompounding can also be seen as a problem related to word segmentation. Unfortunately, there are few systematic studies into different decompounding algorithms.

As a language-independent alternative to stemming, *character n-gram techniques* are helpful[20] (McNamee 2009). For character *n*-gram retrieval, words are split into sub-units, i.e., a set of overlapping strings of characters, typically between four and six characters long. For example, the word 'airport' may be split into character *n*-grams of length 4 as follows: '_air', 'airp', 'irpo', 'rpor', 'port', 'ort_'. The technique usually yields a number of common character *n*-grams for pairs of words that should be conflated, and these *n*-grams tend to give an adequate representation of the stem of the word. Specifically, matches on only parts of words become possible. A drawback is that this technique inflates the size of the index, with the actual size depending on implementation and choice of length of *n*-grams. There is also an issue of acceptability to the user, as unrelated words in queries and documents may be matched on the basis of common character *n*-grams, making it hard for users to understand why certain irrelevant documents are returned by the system.

[20] Not to be confused with word *n*-grams used to solve the segmentation problem in Eastern Asian languages, see Step 4.

Additionally to the issues with differing word surface forms, problems with inconsistency or variations in usage of writing conventions can arise. Examples include the use of capitalisation in English, where words at the beginning of a sentence or in titles are capitalised even when otherwise written in lowercase, and use of diacritics in French where the diacritical marks are seldom used when a corresponding character is written in uppercase. Similar issues exist sometimes in typed text, where in languages such as German and French, users often do not write characters with diacritics if they are not easily available on a keyboard, and revert to using corresponding 'basic' characters or character combinations instead.

Nearly all information retrieval systems convert the entire text to lowercase. There seems to be a broad consensus that it is difficult to realise improvements in terms of the later matching stage by keeping upper/lower case distinctions, with any potential gain being offset by the difficulties of handling exceptions related to inconsistent use and special instances as mentioned above. However, some processing components that can be used during indexing employ deeper grammatical analysis that may depend on case information, which thus has to be preserved prior to the application of such components.

The picture is less clear with respect to the handling of diacritics. The issue of diacritics removal is not extensively researched. An exception to this is work by McNamee and Mayfield (2003), which reports on the changes to retrieval effectiveness when removing or retaining diacritical marks for eight languages.[21] They observe only small, statistically insignificant differences with respect to the handling of diacritics. It is probably advisable to remove diacritics in order to maximise the potential of matches between queries and documents, unless there is an issue with presenting the thus normalised features to the user in a subsequent interactive process.

### *2.4.6 Enrichment (Step 6)*

A number of additional processing steps can be undertaken before the document is finally inserted into the index. We only briefly mention some of the most common options here.

Phrase (multiword) detection addresses issues of polysemy and homonymy (i.e., with words that have multiple meanings). Often such ambiguous words occur as part of a larger, multiword expression that is unambiguous. By detecting these multiword expressions and joining them into single indexing units (thus effectively employing a process that is the reverse of decompounding discussed in Step 5), retrieval effectiveness, especially precision, may be enhanced, potentially at the

[21] Dutch, English, Finnish, French, German, Italian, Spanish, Swedish.

expense of recall. Overall, findings on the effectiveness of using phrases for retrieval have been mixed (see e.g., Mitra et al. 1997).

Thesaurus expansion can alleviate problems with synonymy, i.e., the use of different words to convey the same meaning. Usually, manually prepared resources are used acting as a form of 'look-up table', with synonyms being added to the stream of indexing units. It may be necessary to use domain-specific resources, both to enhance vocabulary coverage, and to avoid over-aggressive expansion by including many obscure terms. Alternatively, there are approaches that use resources generated by statistical approaches on suitable training data, e.g., similarity thesauri (Qiu and Frei 1993).

Named Entity Recognition (NER) is related to multiword detection: often entities have complex names composed of multiple constituents; by handling such complex names as single indexing units, ambiguity is decreased. Also, names potentially have different representations, differing in length, formality, etc., or using different spellings and transliteration. By adding an identification step (Named Entity Recognition and Identification (NERI)), these alternative forms can be replaced by a common representation. If translation is introduced into the retrieval system, NER/NERI can be essential in avoiding wrong translations of names (such as e.g., translating the name 'Bush', interpreting it as a reference to a shrub).

## 2.5 Matching Phase

Matching of queries and documents takes place on their respective indexed representations. The mechanisms described in this section are to a high degree language-independent, insofar as they operate on exact matches at feature level. However, a number of the steps outlined in Section 2.4, such as feature normalisation and enrichment, can lead to matches between queries and documents that contain none of the search terms in their original surface form as entered by the user. This is consistent with the task of an information retrieval system to return as much relevant information as possible (optimising recall), while keeping the number of false matches (irrelevant documents retrieved) minimal (optimising precision). It is not the retrieval of documents that match certain 'patterns' (the surface form of the search terms) that is the goal, but the retrieval of documents that are relevant to the information need by the user.

### *2.5.1 'Bag of Words' Paradigm*

The output of the indexing pipeline (succession of indexing steps) as described in Section 2.4 is a stream of tokens, representing the occurrence of indexing 'features' in the associated documents. In the following discussion of matching, we will

concentrate on features derived from textual information ('terms'), but more generally the features can potentially be arbitrary abstract units, such as features representing phonetic information, information contained in non-textual input data (e.g., edges, colour histograms, audio characteristics), or metadata information (category assignments, information about the document source, etc.). Most of the considerations in the following sections also apply to such types of features with little or no adaptation. For our immediate purposes, features are derived during indexing from words by successive application of indexing steps, which usually affect the character string representing the word. The dominant matching approaches treat the resulting streams of tokens as an unsorted 'bag' (multiset), i.e., they ignore the order in which the underlying words originally occurred in the text. This use of a so-called 'bag of words'[22] has implications when matching on phrasal expressions, represented by a string of multiple words.[23] It should also be noted that features introduced during enrichment (Section 2.4.6) potentially have no natural placement in the ordering of tokens, and thus their use is facilitated by adopting the 'bag of words' approach.

As an example for the representation of documents and queries in 'bag of words' form, consider the abstract of this chapter as a potential searchable document (see Table 2.3). Let us assume that the user inputs 'index structures in IR systems' as a query. This query shares no common words with the text sample given apart from the single stopword 'in', which can be assumed to occur in most English texts of non-trivial length. Also given in the table are the bag of words representations for both the query and the text sample. Both text and query were indexed as described in Section 2.4, with the use of tokenisation, conversion to lowercase, stopword elimination (Smart stopword list[24]) and stemming (using the widely used Porter stemmer). Obviously, after indexing there is a much better match on 'index' and 'system'. Some finer issues and limitations of stemming also become apparent: the conflation on 'structures' is probably an undesirable one: while the abstract talks of 'structuring the discussion', the user is looking for 'index structures'. Furthermore, no conflation between 'information retrieval' and 'IR' takes place – this form of matching would require additional resources such as a domain-specific thesaurus.[25]

[22] While 'bag of words' is the customary choice of term to describe the approach, it would be more helpful to use 'bag of tokens', to indicate the nature of the units contained in the indexed representation.

[23] These implications are tied to the usual assumption that all indexing features that have corresponding tokens in the bag are independent of each other – an assumption that is clearly violated in the case of phrasal expressions.

[24] The freely available and often used Smart stopword list is rather exhaustive, containing entries such as 'need', and thus nicely illustrating the issues with stopword elimination discussed in Section 2.4.4, by suppressing the final word of the abstract in the indexed representation.

[25] Whether the text sample is actually relevant to the user's query is a different question; one which is only answerable if the underlying information need, user preferences, etc. are known.

**Table 2.3** Original version and bag of words representation for a text sample and a possible associated query, respectively. Unlike this example, in real systems, a bag of words has no particular ordering

| |
|---|
| 'The information retrieval system stands at the core of many information acquisition cycles. Its task is the retrieval of relevant information from document collections in response to a coded query based on an information need. In its general form, when searching unstructured, natural language text produced by a large range of authors, this is a difficult task: in such text there are many different valid ways to convey the same information. Adding to the complexity of the task is an often incomplete understanding of the desired information by the user. In this chapter, we discuss the mechanisms employed for matching queries and (textual) documents within one language, covering some of the peculiarities of a number of widely spoken languages. Effective within-language retrieval is an essential prerequisite for effective multilingual information access. The discussion of within-language information retrieval or monolingual information retrieval can be structured into two main phases: the indexing phase, commonly implemented as a pipeline of indexing steps, producing a representation that is suitable for matching; and the matching phase, which operates on the indexed representations and produces a ranked list of documents that are most likely to satisfy the user's underlying information need' |
| access(1) acquisit(1) ad(1) author(1) base(1) chapter(1) code(1) collect(1) commonli(1) complex(1) convei(1) core(1) cover(1) cycl(1) desir(1) difficult(1) discuss(2) document(3) effect(2) emploi(1) essenti(1) form(1) gener(1) implement(1) incomplet(1) index(3) inform(10) languag(5) larg(1) list(1) main(1) match(3) mechan(1) monolingu(1) multilingu(1) natur(1) number(1) oper(1) peculiar(1) phase(3) pipelin(1) prerequisit(1) produc(3) queri(2) rang(1) rank(1) relev(1) represent(2) respons(1) retriev(5) satisfi(1) search(1) spoken(1) stand(1) step(1) structur(1) suitabl(1) system(1) task(3) text(2) textual(1) underli(1) understand(1) unstructur(1) user(2) valid(1) wai(1) wide(1) |
| 'index structures in IR systems' |
| index(1) ir(1) structur(1) system(1) |

### *2.5.2 Inverted Index*

Every information item (document, text, but also the query) in the retrieval system is represented by one 'bag of words', an unordered multiset of tokens obtained during the indexing process. The core task of the matching component is to compute the similarity between the bag representing the query and the bags for every document contained in the system, i.e., the system conceptually has to compute a very large number of similarity scores (e.g., in the case of web search services, in the order of billions). The formula guiding this computation is called the 'weighting scheme'. In order to efficiently compute the scores, two basic simplifications are used by nearly all retrieval systems:

1. Documents that do not contain any of the features contained in the indexed representation of the query are assigned a score of 0.
2. Only exact matches between the features in the indexed representation of the query and the indexed representation of the document are considered.

Both points should not be confused by fact that the system may match a document with a query based on words occurring in different surface forms, or even based on words not occurring in the original document or initial query

(expansion terms): it is the bag of words representation of the query and document that will contain exact matches on feature level.

The two assumptions above make the use of an 'inverted index' possible: the system builds and maintains a data structure (usually implemented in the form of a hash table[26]) that allows efficient look-up of a feature, returning a list of all documents containing the given feature. Essentially, to produce a similarity score for every document in the system in response to a query where the query has $n$ different associated features,[27] the system has to perform $n$ look-ups in the inverted index, collecting the lists of documents containing each feature. All documents not contained in the union of these $n$ lists receive a score of 0, and need not be considered further. For all other documents, scores are calculated (usually added up) according to the number of features they contain. Lastly, the system has to sort the documents by score and return the ranked list. To allow efficient summing up of similarity scores, an accumulator structure[28] is used, which allows storage of partial sums, and easy updating of these partial sums based on an identification of the associated document (a 'document identifier', 'docid'). Figure 2.3 shows how the inverted index associates each feature (term) in the collection with a table of (document identifier, frequency) pairs that denote how often this term occurs in a document. Thus the term 'inverted': while a 'bag of words' makes it possible to identify which features are associated with a given document, the inverted index makes it possible to identify which documents are associated with a given feature.

Please note that the use of hash tables for implementation of the inverted index carries an important implication which is the basis for the second simplification proposed above: while look-up of exact matches at feature level in hash tables is very efficient, it is not possible to implement partial matching at feature level with reasonable effort using the same data structures. Specifically, this means that no querying with wildcard characters[29] (*, ? etc.) is possible – inexact matches between query and documents are solely enabled through the normalisation during the indexing phase.

### *2.5.3 Basic Matching Algorithm*

The above considerations lead to the following basic algorithm for weighting:

[26] A data structure that allows location of items at a 'cost' that is typically independent of the size of the structure (Sedgewick and Wayne 2011).

[27] Derived from $m$ original search terms, where potentially $m <> n$.

[28] Also typically implemented in the form of a hash table.

[29] A wildcard character is a placeholder for one or more characters that are unknown or unspecified.

**Initialise** accumulator, set number of entries to 0
**For each** feature *f* in the set of features representing the query *i* **do**
    **Look up** feature *f* in inverted index
    **Fetch** list *l* of identifiers of documents containing the feature *f*
    **For each** identifier of a document *j* in the list *l* do
        **Calculate** a weight *w* for the match (feature *f*, document *j*)
        **Add** weight to accumulator value for document *j*
    **Next**
**Next**
[*If applicable*: **Normalise** the scores in the accumulator]
**Sort** documents according to their accumulator values
**Return** ranked list of identifiers of the top *n* documents

If multiple queries are evaluated in one 'batch', the use of a query index (see Figure 2.3), associating a 'query identifier' with a list of the features pertaining to the query, is helpful.

The different weighting schemes discussed below mainly differ in how they calculate the weight *w* and how they normalise the scores prior to compiling the ranked list. To address the twin problems of calculating the weights and normalising the scores, there are two main approaches:

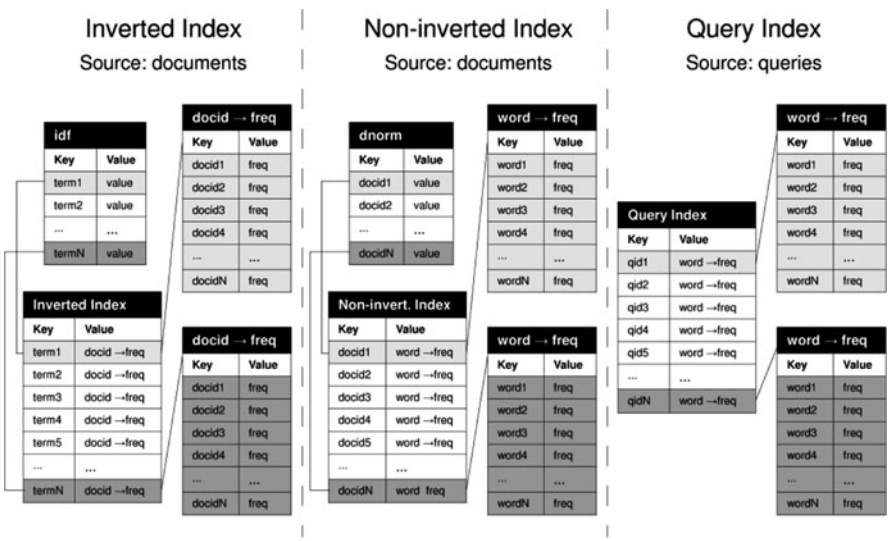


**Fig. 2.3** Indexes typically used for retrieval. The 'inverted index' is mandatory to efficiently compute most weighting schemes. Some weighting schemes also benefit from building an analogous 'non-inverted index' and 'query index'

1. Weighting schemes based on the 'vector space model', initially developed at Cornell University in the 1960s, that interpret query-document similarity as a distance measure in a high-dimensional space of document features (Salton et al. 1975).
2. Probabilistic weighting schemes, that aim to estimate the probability of relevance of a document with respect to a given query as accurately as possible (see e.g., Robertson et al. 1980).

Both approaches have been fine-tuned over decades, and modern implementations consistently provide very good retrieval effectiveness across many different languages when compared with alternative approaches. Other effective approaches include ranking using language models (Ponte and Croft 1998, Hiemstra and de Jong 1999) and 'divergence from randomness' (Amati and Rijsbergen 2002). We also cover the language modelling approach briefly, because it has an interesting extension to cross-language information retrieval (for this extension, see Section 3.3.2).

The discussion of models and weighting schemes uses the notation summarised in Table 2.4.

### 2.5.4 Vector Space Model

The *vector space model* is based on the notion of an $m$-dimensional vector space where the different dimensions represent the unique features contained in the document collection, i.e., if the collection contains 800,000 unique features, the corresponding vector space will have a dimensionality of 800,000.[30] Such a high dimensionality may sound prohibitive, and may be difficult to mentally visualise, but in practice, there is no problem in calculating the necessary operations efficiently (e.g., scalar product, sums and vector length).

To understand the underlying idea, a simple visualisation of the vector space model is helpful (see Figure 2.4).

Every document $d_j$ is mapped into the (orthogonal) vector space, and is represented by a single, $m$-dimensional vector $\overrightarrow{d_j}$ containing as many components as there are features in the collection, i.e.,

$$\overrightarrow{d_j} := \left(w(\varphi_0, d_j), .., w(\varphi_k, d_j), .., w(\varphi_{m-1}, d_j)\right)^T$$

[30] Complexities arise when one considers the possibility of queries containing features that do not occur in any documents of the collection. Theoretically, the vector space is extended by those features (i.e., the vector space represents an 'indexing vocabulary' that extends beyond only those features present in the collection), but this case can be ignored for the discussion of the basic workings of the model as presented here.

**Table 2.4** Notation used for discussion of models and weighting schemes

$D$: set of documents (the document collection)
$N$: number of documents in the collection ($N := |D|$)
$d_j$: single document ($d_j \in D$)
$l_j$: length of document $d_j$, as used for Lnu.ltn and BM.25 weighting schemes (varying definitions)
$\Delta$: average document length, as used for Lnu.ltn and BM.25 weighting schemes ($\Delta := \left(\sum_{d_j \in D} l_j\right)/N$)
$\lambda$: balancing factor, as used for Lnu.ltn weighting scheme and language modelling approach (different definitions)
$q$: single query
$\Phi$: indexing vocabulary
$\varphi_k$: single indexing feature ($\varphi_k \in \Phi$)
$\Phi(d_j)$: set of features representing document $d_j$ ($\Phi(d_j) \subset \Phi$)
$\Phi(q)$: set of features representing query $q$ ($\Phi(q) \subset \Phi$)
$w(\varphi_k, d_j)$: weight of feature $\varphi_k$ for document $d_j$
$w(\varphi_k, q)$: weight of feature $\varphi_k$ for query $q$
$\mathrm{ff}(\varphi_k, d_j)$: feature frequency of feature $\varphi_k$ for document $d_j$
$\mathrm{tf}(\varphi_k, d_j)$: term frequency of feature $\varphi_k$ for document $d_j$, synonymous for our purposes with $\mathrm{ff}(\varphi_k, d_j)$
$\mathrm{aff}(d_j)$: average feature frequency in document $d_j$ ($\mathrm{aff}(d_j) := \left(\sum_{\varphi_i \in \Phi(d_j)} \mathrm{ff}(\varphi_i, d_j)\right) / |\Phi(d_j)|$)
$\mathrm{df}(\varphi_k)$: document frequency of feature $\varphi_k$
$\mathrm{idf}(\varphi_k)$: inverse document frequency of feature $\varphi_k$ ($\mathrm{idf}(\varphi_k) := \log\left(\frac{1+N}{1+\mathrm{df}(\varphi_k)}\right)$)
$m$: dimensionality of the vector space used for matching
$\overrightarrow{d_j}$: vector representation of the bag of tokens associated with $d_j$ ($\overrightarrow{d_j} := \left(w(\varphi_0, d_j), .., w(\varphi_k, d_j), .., w(\varphi_{m-1}, d_j)\right)^T$)
$\vec{q}$: vector representation of the bag of tokens associated with $q$ ($\vec{q} := (w(\varphi_0, q), .., w(\varphi_k, q), .., w(\varphi_{m-1}, q))^T$)
$\mathrm{RSV}(q, d_j)$: retrieval status value of document $d_j$ with respect to query $q$, as determined by a given retrieval method; used for ranking
$D^{rel}(q)$: set of documents that are relevant with respect to query $q$ ($D^{rel}(q) \subset D$)
$D^{non}(q)$: set of documents that are non-relevant with respect to query $q$ ($D^{rel}(q) \cup D^{non}(q) = D$)
$\alpha, \beta, \gamma$: tuning parameters for Rocchio relevance feedback formula
$P(R|d_j, q)$: probability of relevance given document $d_j$ and query $q$
$P(\varphi_k|R, q)$: probability of occurrence of feature $\varphi_k$ in any given relevant document for query $q$
$P(\varphi_k|\overline{R}, q)$: probability of occurrence of feature $\varphi_k$ in any given non-relevant document for query $q$
$M_j$: unigram language model associated with document $d_j$
$P(q|M_j)$: probability for generating query $q$ based on model $M_j$ assigned to document $d_j$
$P(\varphi_k)$: probability of drawing feature $\varphi_k$ randomly from the entire collection
$P(\varphi_k|M_j)$: probability of drawing feature $\varphi_k$ randomly from document $d_j$ associated with model $M_j$

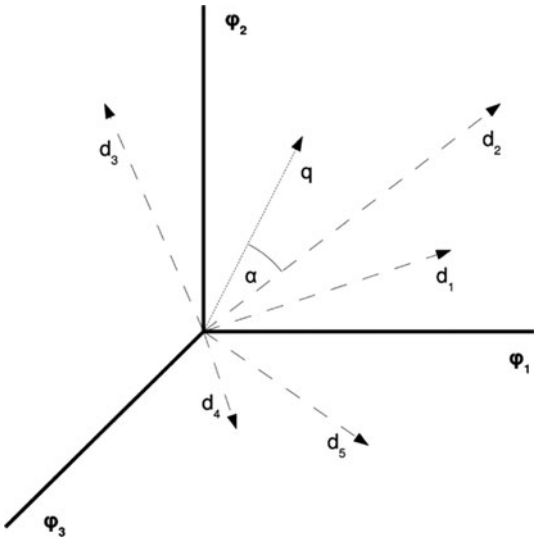


**Fig. 2.4** Representation of a query and a small document collection (containing five documents) in a three-dimensional vector space. A vector space of the order 3 implies that only three different features occur in the indexed representations of the documents

where $w(\varphi_k, d_j)$ is the weight of feature $\varphi_k$ for document $d_j$, and $m$ is the number of different features in the collection. Analogously, the query $q$ is mapped to a vector $\vec{q}$, where

$$\vec{q} := (w(\varphi_0, q), .., w(\varphi_k, q), .., w(\varphi_{m-1}, q))^T$$

Only those components that correspond to features represented in the 'bag of tokens' for the document or the query are assigned a non-zero (usually, but not necessarily, positive) value. All other components are set to 0. Please note an important consequence of the choice of an orthogonal vector space: as each indexing feature represents one of the (orthogonal) axes, an implicit assumption that all features (and thus their associated words or word forms) are mutually independent is made. This assumption clearly does not hold up to characteristics of natural languages, and is a limiting factor in the effectiveness of weighting schemes derived from this model.

### *2.5.5 The tf.idf-Cosine Weighting Scheme*

For those features actually represented in the bag for the document or query, the vector space model mandates no specific way to weigh their contribution to the vector; however, the most popular choices are all based on variations of 'tf.idf feature weighting'. Essentially, the weight of a feature in a given document vector is maximised if the feature is representing a token that is:

1. Frequent in the bag for the document that corresponds to the vector; and
2. Rare in the remainder of the document collection (as determined by the respective bags).

For rule 1, the 'feature frequency' (ff), also often called 'term frequency' (tf)[31] in the textual case, is used: $\mathrm{ff}(\varphi_k, d_j)$ denotes the number of occurrences of tokens corresponding to feature $\varphi_k$ in the bag representing document $d_j$ (one value per feature/document pair).

For rule 2, the 'document frequency' (df) is used: $\mathrm{df}(\varphi_k)$ denotes the number of documents in the collection that are associated with a bag that contains tokens corresponding to feature $\varphi_k$ (one value per feature in the collection). As the weight is maximised for features with low document frequency (see rule 2 above), the 'inverse document frequency', calculated as

$$\mathrm{idf}(\varphi_k) := \log\left(\frac{1+N}{1+\mathrm{df}(\varphi_k)}\right)$$

where $N$ is the total number of documents in the collection, is used. For standard tf.idf weighting, the two values for ff and idf are multiplied:

$$w(\varphi_k, d_j) := \mathrm{ff}(\varphi_k, d_j) * \mathrm{idf}(\varphi_k)$$

where $w(\varphi_k, d_j)$ is the weight of the component associated with feature $\varphi_k$ in the vector representing document $d_j$. The weights for the components of the vector representing the query are calculated analogously.

Overall similarity between the vector representations of a query $q$ and a document $d_j$ (the retrieval status value $\mathrm{RSV}(q, d_j)$) is determined by calculating the cosine of the angle $\alpha$ between the two vectors. This angle will be small if the document and query share many features.[32] To compute the cosine of the angle, the scalar product of the two vectors is used, which is then normalised by the vector lengths:

$$\mathrm{RSV}(q, d_j) := \frac{\vec{q} \cdot \overrightarrow{d_j}}{||\vec{q}||\ ||\overrightarrow{d_j}||} = \cos\alpha$$

The RSV will be high if the angle is small (with a maximum of 1 for an angle of 0 degrees), and will be low for larger angles (with a minimum of 0 in the case of two orthogonal vectors). Please note that the Euclidean distance between the vectors cannot alternatively be used to determine similarity, as the distance tends to be

[31] The use of 'term frequency' actually gives rise to the common name for this weighting scheme: tf.idf-Cosine, although ff.idf-Cosine would be equally justified, especially when considering non-textual information.

[32] The angle will be small independently of the length of the document. Since queries are typically much shorter than documents, this is a desirable feature.

dominated by the different lengths of queries and documents. The overall weighting scheme ('*tf.idf-Cosine weighting scheme*') is thus:

$$w(\varphi_k, d_j) := \mathrm{ff}(\varphi_k, d_j) * \mathrm{idf}(\varphi_k)$$

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) * \mathrm{idf}(\varphi_k)$$

$$\mathrm{RSV}(q, d_j) := \frac{\sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) * w(\varphi_k, q)}{\sqrt{\sum_{\varphi_k \in \Phi(d_j)} w(\varphi_k, d_j)^2} \sqrt{\sum_{\varphi_k \in \Phi(q)} w(\varphi_k, q)^2}}$$

The basic matching algorithm (Section 2.5.3) adapted for calculation of the '*tf.idf-Cosine* weighting scheme' reads as follows:

**Initialise** accumulator, set number of entries to 0
**Set** query norm *qn* to 0
**For each** feature $\varphi_k$ in the set of features representing the query *q* **do**
  **Look up** feature $\varphi_k$ in inverted index
  **Fetch** list *l* of identifiers of documents containing the feature $\varphi_k$
  **Calculate** $w(\varphi_k,q) = \mathrm{ff}(\varphi_k,q) * \mathrm{idf}(\varphi_k)$
  **Add** $w(\varphi_k,q) * w(\varphi_k,q)$ to query norm *qn*
  **For each** identifier of a document $d_j$ in the list *l* do
    **Calculate** $w(\varphi_k,d_j) = \mathrm{ff}(\varphi_k, d_j) * \mathrm{idf}(\varphi_k)$
    **Add** $w(\varphi_k,d_j) * w(\varphi_k,q)$ to accumulator value for $d_j$
  **Next**
**Next**
**Compute** square root of query norm *qn*
**For each** identifier of a document $d_j$ in the accumulator **do**
  **Normalise** the score by dividing by the query norm *qn*
  **Normalise** the score by dividing by the document norm $dn(d_j)$ of document $d_j$ (can be precomputed independently of queries)
**Next**
**Sort** documents according to their accumulator values
**Return** ranked list of identifiers of the top *n* documents

This weighting scheme is also sometimes referred to as the 'ntc.ntc' weighting scheme. The document norm values $dn(d_j)$ can be precomputed easily if a 'non-inverted index', which associates every document with its corresponding bag representation, is available (see also Figure 2.4). These values are independent of a specific query.

The vector space model allows easy interpretation of various further operations related to document retrieval: instead of calculating the similarity between the vector representations of the query and the documents, one can also calculate the similarity between the vector representations of:

- Two different documents (determining inter-document similarity, and potentially clustering a document collection according to these similarities).
- A document in the role of 'query' compared to a collection of user queries ('profiles') (implementing notifier 'push' systems that store queries representing long-term information needs and match a stream of documents against them).
- Two different queries (determining further queries related to a given query – e.g., for use in recommender systems that point the user to associated content of interest).

All these operations can be efficiently computed using this same architecture and approach.

A modern, effective weighting scheme based on the vector space model and the simpler tf.idf-Cosine approach is 'Lnu.ltn', using 'pivoted document length normalisation', defined as follows (Singhal et al. 1996):

$$w\left(\varphi_k, d_j\right) := \frac{\frac{1+\log\left(\mathrm{ff}\left(\varphi_k, d_j\right)\right)}{1+\log\left(\mathrm{aff}\left(d_j\right)\right)}}{\lambda l_j + (1-\lambda)\Delta}$$

$$w(\varphi_k, q) := (1 + \log(\mathrm{ff}(\varphi_k, q))) * \mathrm{idf}(\varphi_k)$$

$$\mathrm{RSV}\left(q, d_j\right) := \sum_{\varphi_k \in \Phi(q) \cap \Phi\left(d_j\right)} w\left(\varphi_k, d_j\right) * w(\varphi_k, q)$$

where $l_j$ is the length of document $d_j$ (different definitions of 'length', such as e.g., byte length or number of tokens are possible), $\Delta$ is the average document length in the collection and $\mathrm{aff}\left(d_j\right)$ is the average feature frequency in document $d_j$. The parameter $\lambda$ is derived from suitable training data and balances between the actual document length and the average document length.

### 2.5.6 *Relevance Feedback*

There is an interactive extension to the basic vector space weighting, which allows the user to judge an initial list of retrieved documents for their relevance. On the basis of this additional information the retrieval system can determine a new ranking ('relevance feedback').

If the user judges some or all the documents in the initial ranked list for relevance (assigning a binary score – relevant or irrelevant), the vector representing the query can be rotated away from the irrelevant documents towards the relevant documents (see Figure 2.5). A new vector representing a new, modified query is thus constructed – and can be interpreted in terms of the features that make up this vector. In a sense, through this rotation, the system 'identifies' new features that best represent the documents judged as relevant by the user. As a consequence of

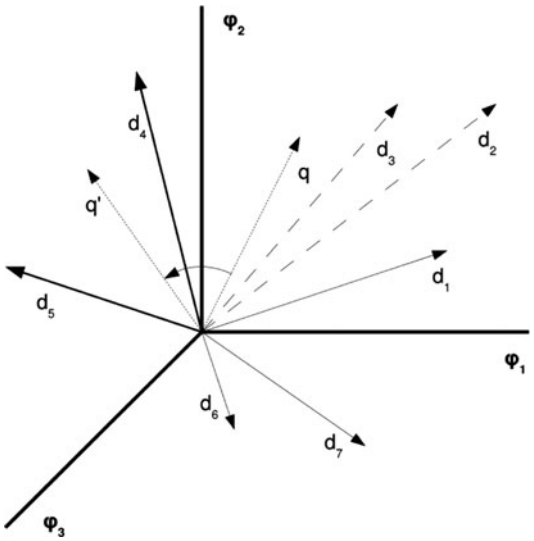


**Fig. 2.5** Relevance feedback. A vector space of order 3 is shown. The concept generalises to vector spaces of arbitrary order. Documents judged relevant by the user ($d_4$ and $d_5$) are represented by a bold vector, and documents judged non-relevant ($d_2$ and $d_3$) are represented by a dashed vector. All other documents are not judged by the user. Based on the judgments, the query vector $\vec{q}$ is rotated, and a new vector q' is obtained that represents a new selection of query terms

constructing the new, rotated query vector, the system can return an updated list of documents with vector representations most similar to that new vector, or it can present the features representing the new query vector to the user to assist in formulating a new, improved query. The new vector can be used directly for retrieval (thus effectively discarding original query features unless they are still represented in the vector), or to construct a new hybrid query composed of original query features and expansion features derived from the new vector (Salton and Buckley 1990).

Formally, the new vector $\overrightarrow{q'}$ is obtained as follows (known as the 'Rocchio method'):

$$\overrightarrow{q'} := \alpha \frac{\vec{q}}{||\vec{q}||} + \frac{\beta}{|D^{rel}|} \sum_{d_j \in D^{rel}} \frac{\overrightarrow{d_j}}{||\overrightarrow{d_j}||} - \frac{\gamma}{|D^{non}|} \sum_{d_k \in D^{non}} \frac{\overrightarrow{d_k}}{||\overrightarrow{d_k}||}$$

where $D^{rel}$ and $D^{non}$ are the sets of *known* (i.e., those judged by the user) relevant and irrelevant documents, respectively. $\alpha$, $\beta$ and $\gamma$ allow a tuning of the influence of the different components (original query, known relevant documents, known irrelevant documents). Typically, $\beta$ is chosen to be greater than $\gamma$, and $\gamma$ is often chosen to be 0 (i.e., relevant documents have a greater influence on the new vector than irrelevant documents).

In a variant of 'relevance feedback', the interactive selection of relevant documents by the user is omitted, and replaced by the assumption that the top $n$ documents obtained by the initial ranking are all relevant. This assumption will only yield useful results (i.e., improvements in retrieval effectiveness) if indeed all or nearly all of the documents used for feedback are relevant. If too many irrelevant

**Table 2.5** Two examples of queries expanded by the use of pseudo relevance feedback. Expanded queries (on the right) are lightly stemmed (e.g., 'Claes' → 'clae')[a]

| Original query | New query after pseudo relevance feedback |
|---|---|
| Who won the gold medal in the super G in Lillehammer at the Olympic Winter Games 1994? | olympic roffe-steinrotter event gold super steinrotter downhill won race time world medal slalom ski super-g skier roffe top us g |
| Why was the secretary general of NATO forced to resign in 1995? | clae agusta brussel mr foreign prime alliance party nato general year minister belgian belgium affair secretary scandal government resignation helicopter |

[a] These queries were taken from the CLEF evaluation campaign. Examples produced by Jacques Savoy, Université de Neuchâtel.

documents erroneously contribute to the computation of a new query vector, retrieval effectiveness will degrade. Given a highly effective retrieval method for initial ranking, this variant of relevance feedback, called 'pseudo relevance feedback' or 'blind relevance feedback' has been shown to improve average retrieval effectiveness in many retrieval scenarios. Coupled with the fact that no interactive participation by the user is needed, pseudo relevance feedback can be a very attractive tool for boosting recall. In any given operational setting, however, it is necessary to carefully consider whether degradation of retrieval effectiveness for some queries, which is in practice unavoidable, is acceptable.

Table 2.5 shows two examples of English queries that were expanded by pseudo relevance feedback. The original version of the query is shown on the left, and the expanded version is shown on the right. As the expansion terms are calculated using the indexed representation of the query, the expanded query is shown in this form, i.e., after case normalisation, stopword removal and (conservative) stemming. Expansion terms are determined by assuming the first 5 documents to be relevant, and producing a maximum of 20 terms. Analysing the expanded version of the first query, we can see that the system 'learned' the name of the winner of the sports event in question ('Diann Roffe-Steinrotter'), and her nationality ('US'), which made it possible to pull in more relevant content (increasing the recall). In the second query, which derives from a political topic, we again see the system picking up the name of the politician in question ('Claes'), but also a number of terms related to the reason for his resignation ('Agusta', 'helicopters', 'scandal'). For these examples, the original query features have been discarded, but can be included again if they are part of the set of 20 new expansion terms.

### *2.5.7 Probabilistic Weighting Schemes*

The class of weighting schemes discussed in this section is based on the idea of calculating the probability of relevance (i.e., of a 'relevant match') given a document $d_j$ and the query $q$ issued by the user. The ultimate goal is thus to calculate

$P(R|d_j, q)$ for all documents in the collection, and then rank according to the probabilities obtained. Since relevance is a subjective notion, it is clear that only estimates of the desired probabilities can be calculated, and that even these estimates draw on a number of assumptions that hinder the ability of the retrieval system to produce an accurate ranking. Specifically, the following compromises are made to derive a formula suitable for ranking documents:

1. Instead of calculating an estimate of $P(R|d_j, q)$, a score $\text{RSV}(q, d_j)$ is calculated, where $\text{RSV}(q, d_j) := f(P(R|d_j, q)) + g(q)$ and $f()$ is an order-preserving function. In practice, this leads to an identical ranking, but can be a serious drawback when real probabilities are needed.[33]
2. It is assumed that documents represented by the same 'bag of words' after the indexing phase are identical for retrieval purposes (this means, e.g., that the ordering of words is not used for ranking purposes).
3. It is assumed that all features (terms) in the collection are pair-wise independent.[34] This is clearly not the case in any non-trivial natural language utterance, and is a very problematic assumption. It is of critical importance for the derivation of the weighting scheme, however, as it makes it possible to base the score on estimates of the probability of relevance given the query and the occurrence of individual query terms in the document.
4. Only the presence or absence of a term in a document is used to calculate the estimate, not its frequency of occurrence.
5. It is assumed that all features (terms) that are not part of the query are equally distributed in relevant and non-relevant documents.
6. Based on the assumptions 1–5 above, scores $\text{RSV}(q, d_j)$ can be computed if the probabilities $P(\varphi_k|R, q)$ and $P(\varphi_k|\overline{R}, q)$, i.e., the probabilities that a specific feature occurs in any given relevant or non-relevant document, respectively, are known. These probabilities could be derived from training data; however, they are usually unavailable. Thus, it is assumed that the number of relevant documents is much smaller than the number of non-relevant documents, i.e., $D^{rel}(q) \ll D^{non}(q)$. In practice, $D^{rel}(q)$ can then be set to 0 for further calculation, even though this clearly would lead to a paradoxical situation if actually true.

The above set of compromises allows the derivation of the so-called 'Robertson-Spärck Jones weighting scheme' (using assumptions 1–5), and when additionally including assumption 6 leads to weighting based on idf-style weights of individual features that is very similar in nature to the tf.idf-weighting discussed in the context of the vector space model:

[33] Real probabilities would allow easy thresholding, which is important, e.g., for cutting off result lists at a certain probability level; or for filtering tasks, where documents would only be returned when exceeding the threshold.

[34] Compare this to the vector space model, where the orthogonality of the axes that represent the features implicitly leads to the same assumption.

$$\mathrm{RSV}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} \log\left(\frac{\frac{1}{2} + N - \mathrm{df}(\varphi_k)}{\frac{1}{2} + \mathrm{df}(\varphi_k)}\right)$$

with $N$ being the number of documents in the collection. Further details of the derivation of this weighting scheme can be found in Schäuble (1997).

When taking feature frequencies into account (i.e., relaxing the limitation in assumption 4), a significantly more effective weighting scheme can be derived. This new weighting scheme is widely known under the name 'BM.25'[35] (Walker et al. 1998). The BM.25 weighting scheme is defined as follows:

$$w(\varphi_k, d_j) := \frac{3 * \mathrm{ff}(\varphi_k, d_j)}{2\left(0.25 + 0.75\frac{l_j}{\Delta}\right) + \mathrm{ff}(\varphi_k, d_j)}$$

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) * \log\left(\frac{0.5 + N - \mathrm{df}(\varphi_k)}{0.5 + \mathrm{df}(\varphi_k)}\right)$$

$$\mathrm{RSV}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) * w(\varphi_k, q)$$

with $l_j$, the document length, being defined as the number of tokens occurring in a document, and $\Delta$ being the average document length. This scheme has proven to be very effective on a wide range of different test collections covering a substantial number of different languages.

### 2.5.8 Ranking Using Language Models

Instead of estimating the probability of relevance of a document given the query, as is the basis for the weighting schemes described in the previous Section 2.5.7, a new class of weighting schemes can be derived on the basis of estimating a 'query generation probability' given a document. For our purposes, it is assumed that a unigram language model $M_j$ is associated with each document $d_j$. This model assigns probabilities of occurrence $P(\varphi_k|M_j)$ to each feature $\varphi_k$, while making the assumption that all terms occur independently. We can then calculate the overall probability for generating the entire query based on the model assigned to each document, which is effectively the probability that the query is randomly sampled from a document, and then rank the documents according to this probability. Please

[35] Sometimes incorrectly referred to as 'Okapi weighting', after the name of the retrieval system that originally implemented the scheme.

note that this idea implicitly assumes that users choose 'reasonable' query terms, i.e., that they have some notion of which terms should occur in relevant documents.

The desired probability is calculated as follows:

$$P(q|M_j) = P(\varphi_1, \varphi_2, .., \varphi_n|M_j) = \prod_{i=1}^{n} \left((1-\lambda)P(\varphi_i) + \lambda P(\varphi_i|M_j)\right)$$

This formula combines global information about a feature ($P(\varphi_k)$, the probability of drawing feature $\varphi_k$ randomly from the entire collection) with local information about a feature ($P(\varphi_k|M_j)$, the probability of drawing feature $\varphi_k$ randomly from document $d_j$). This combination also avoids a zero overall score in cases where a document does not contain all query terms. The parameter $\lambda$ is determined from suitable training data. $P(\varphi_k|M_j)$ is estimated as $P(\varphi_k|M_j) = \frac{\mathrm{ff}(\varphi_k, d_j)}{\sum_{\varphi_i \in \Phi(d_j)} \mathrm{ff}(\varphi_i, d_j)}$ and $P(\varphi_k)$ is estimated as $P(\varphi_k) = \frac{\mathrm{df}(\varphi_k)}{\sum_{\varphi_i \in \Phi} \mathrm{df}(\varphi_i)}$. The formula can be rewritten to a vector product formula, which can be implemented in much the same way as the weighting schemes previously discussed (Hiemstra and de Jong 1999):

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q)$$

$$w(\varphi_k, d_j) := \log\left(1 + \frac{\mathrm{ff}(\varphi_k, d_j)}{\mathrm{df}(\varphi_k) \sum_{\varphi_i \in \Phi(d_j)} \mathrm{ff}(\varphi_i, d_j)} \cdot \frac{\lambda \sum_{\varphi_i \in \Phi} \mathrm{df}(\varphi_i)}{1-\lambda}\right)$$

$$\mathrm{RSV}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) * w(\varphi_k, q)$$

This weighting scheme allows an elegant extension for the cross-language case, see Section 3.3.2 for details.

### 2.5.9 *Off-Page Information: Page Rank*

We have limited the discussion of matching to weighting schemes that exclusively use 'on-page information', i.e., information that is readily and explicitly contained in the document proper (possibly transformed and enriched through the indexing process). If documents are embedded in a larger context, e.g., through hyperlinking or through associated resources, information from this context ('off-page information') can also be used for retrieval and ranking. The World Wide Web, with its extensive use of hyperlinks, is a good example. By combining a document with parts of the documents that link to it (e.g., with the anchor text, the portion of text

that describes a hyperlink) new terms that describe the content of the original text are potentially added. In a special case, a Spanish language link to an English document may actually provide some translated search terms that may help cross-language retrieval. Similarly, textual descriptions of links may assist in the retrieval of non-textual media, such as images.

Hyperlink structure can also be helpful for ranking: the famous PageRank algorithm first employed by the Google web search service uses the number of links that point to a page as an indication of its 'importance' (or 'popularity') for a query. PageRank does not treat all links as equal: links from highly 'important' pages are given higher weight. As this 'importance' is in turn derived from the PageRank of those pages, the process of calculation is an iterative one.

The original formula for the calculation of PageRank of a given page A is defined as follows (Brin and Page 1998):

$$\mathrm{PR}(A) := (1-d) + d \sum_{B_i \text{ links to } A} \frac{\mathrm{PR}'(B_i)}{\mathrm{C}(B_i)}$$

where $\mathrm{PR}'(X)$ is the PageRank of (web-)page X from the iteration immediately prior to the current one and C(X) is the count of the number of links pointing outwards from page X. $d$ is a dampening factor (usually set as $d = 0.85$). One way to view the resulting values is to liken them to the probabilities that a user that randomly clicks their way through the Web would end up on a particular page. The method works best if links are placed solely on the basis of the merit of the content being linked; unfortunately, as commercial interest in obtaining a good PageRank value for a web page has increased, this is often no longer the case on today's World Wide Web. The different approaches employed to manipulate PageRank and to combat such manipulation are outside the scope of this discussion.

## 2.6 Summary and Future Directions

Effective within-language (monolingual) information retrieval is an essential prerequisite for multilingual information retrieval. In this chapter we have presented the most common mechanisms for the implementation of such retrieval. The main building blocks are an indexing phase (Section 2.4), which prepares documents (offline) and queries (at query time) for matching, and a matching phase (Section 2.5), which calculates retrieval scores and produces a ranked result list. There is considerable freedom in how the same information can be expressed in natural language. Authors can freely choose between synonyms, phrasal expressions and in some languages can form (new) compound words. Metaphors, regional variations and other subtleties complicate matters further. It follows that information retrieval cannot be implemented by simply using pattern matching for all search terms entered in a query (Sections 2.1, 2.2). IR systems need to retrieve

documents based on partial matches, be it that only a portion of the search terms occur in the document, or that some search terms take related forms that differ from the ones used in the query. Probabilistic weighting schemes address these considerations directly, attempting to estimate the probability of relevance of a document given a query (Section 2.5.7), but the same provisions for partial matching are also implicitly contained in the older vector space model (Section 2.5.4). Retrieval takes place on an inverted index (Section 2.5.2) that, while very efficient for looking up exact matches, is not suited for inexact matching of features. The indexing phase thus plays a crucial role in segmenting (Section 2.4.4), normalising (Section 2.4.5) and enriching (Section 2.4.6) document content, and thus creating a bag of words representation (Section 2.5.1) of the documents which is suitable for matching. In some cases, relevant information may even share no common terms with a query. Mechanisms such as relevance feedback can help obtain a match in such situations, by automatically extracting new search terms from other documents found to be relevant, and then starting a new search with an expanded query (Section 2.5.6).

The mechanisms covered in this chapter have a long history in both research and implementation, and all the approaches covered are well introduced and have been generalised to many settings. Still, when focusing on the multilingual aspect, challenges remain. Difficulties related to the handling of different encodings and formats have lessened with the wider adoption of Unicode and XML, which are highly suitable for handling practically arbitrary textual content. However, while the language-dependent indexing mechanisms for language identification, stopword removal, stemming, decompounding and various enrichment options are well researched for the most common languages, work remains for less frequently spoken languages (and for special technical terminology).

## 2.7 Suggested Reading

There is no shortage of resources, both in print and online, that provide a background to the design and implementation of within-language information retrieval systems. For example, the book by Manning et al. (2008) entitled *Introduction to Information Retrieval* is a good starting point that includes topics such as index construction and compression, vector space and probabilistic information retrieval models, relevance feedback and web crawling.[36] The recently published second edition of *Modern Information Retrieval* by Baeza-Yates and Ribeiro-Neto (2011) is a very comprehensive treatment of all things related to information retrieval, including excellent coverage of many of the aspects discussed in this chapter. Croft et al's. (2009) book entitled *Search Engines: Information Retrieval in Practice*

[36] Also available online at http://nlp.stanford.edu/IR-book/information-retrieval-book.html

focuses more on web search but also covers the basics of information retrieval and includes an accompanying Java-based open source search engine called Galago and a set of programming exercises. From the perspective of library and information science, the book *Introduction to Modern Information Retrieval* by Chowdhury (2010) offers a broad introduction to information retrieval, but also includes chapters oriented towards the theory of classification and cataloguing, bibliographic formats and subject indexing. If an understanding of the historical origins of indexing methods and weighting schemes is desired, *Readings in Information Retrieval* by Spärck Jones and Willett (1997) is an invaluable resource. Lastly, the TREC campaigns have been a great driver in the development of effective systems for within-language IR. The proceedings of the different TREC conferences are available online.[37]

## References

Abdou S, Savoy J (2006) Statistical and comparative evaluation of various indexing and search models. In: Proc. AIRS2006. Springer-Verlag LNCS 4182: 362–373

Amati G, Rijsbergen CJV (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4): 357–389

Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval, 2nd edn. ACM Press, New York

Braschler M, Ripplinger B (2004) How effective is stemming and decompounding for German text retrieval? Inf. Retr. 7(3–4): 291–316

Braschler M, Gonzalo J (2009) Best practices in system and user-oriented multilingual information access. TrebleCLEF Project: http://www.trebleclef.eu/

Braschler M, Peters C (2004) Cross-language evaluation forum: objectives, results, achievements. Inf. Retr. 7(1–2): 7–31

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. 30(1–7): 107–117

Broder AZ (2000) Identifying and filtering near-duplicate documents. In: Proc. 11th Annual Symposium on Combinatorial Pattern Matching (COM '00). Springer-Verlag, London: 1–10

Chowdhury GG (2010) Introduction to modern information retrieval. Neal-Schuman Publishers

Croft B, Metzler D, Strohman T (2009) Search engines: information retrieval in practice. Addison Wesley

Dolamic L, Savoy J (2009) When stopword lists make the difference. J. of Am. Soc. for Inf. Sci. 61 (1): 200–203

Dunning T (1994) Statistical identification of language. Technical Report, CRL Technical Memo MCCS-94-273

El-Khair IA (2007) Arabic information retrieval. Annu. Rev. of Inf. Sci. and Technol. 41(1): 505–533

Erickson JC (1997) Options for the presentation of multilingual text: use of the Unicode standard. Library Hi Tech, 15(3/4): 172–188

Harman D (1991) How effective is suffixing?. J. of the Am. Soc. for Inf. Sci. 42(1): 7–15

[37] See http://trec.nist.gov/

Hiemstra D, de Jong F (1999) Disambiguation strategies for cross-language information retrieval. In: Proc. 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '99). Springer-Verlag, London: 274–293

Hull DA (1996) Stemming algorithms - A case study for detailed evaluation. J. of the Am. Soc. for Inf. Sci. 47(1): 70–84

Kaszkiel M, Zobel J (1997) Passage retrieval revisited. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '97). ACM, New York: 178–185

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press

McNamee P (2009) JHU ad hoc experiments at CLEF 2008. In: Proc. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). Springer-Verlag LNCS 5706: 170–177

McNamee P and Mayfield J (2003) JHU/APL experiments in tokenization and non-word translation. In: Proc. 4th Workshop of the Cross-Language Evaluation Forum (CLEF 2003). Springer-Verlag LNCS 3237: 85–97

Mitra M, Buckley C, Singhal A, Cardie C (1997) An analysis of statistical and syntactic phrases. In Proc. 5th International Conference in Computer-Assisted Information Retrieval (RIAO 1997): 200–214

Ponte J, Croft B (1998) A language modeling approach to information retrieval. In Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '98). ACM, New York: 275–281

Porter MF (1980) An algorithm for suffix stripping. Program, 14(3):130–137. Reprint in: Spärck Jones K and Willett P (eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco: 313–316

Qiu Y, Frei H-P (1993) Concept based query expansion. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '93). ACM, New York: 160–169

Robertson SE (1977) The probability ranking principle in IR. J. of Doc. 33(4): 294–304

Robertson SE, van Rijsbergen CJ, Porter MF (1980): Probabilistic models of indexing and searching. Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '80). ACM, New York: 35–56

Robertson SE, Maron ME, Cooper WS (1982) Probability of relevance: a unification of two competing models for document retrieval. Info. Tech: R and D. 1: 1–21

Salton G, Wong A, Yang C (1975) A vector space model for automatic indexing. Commun. of the ACM 18(11): 613–620

Salton G, Buckley C (1990) Improving retrieval performance by relevance feedback. J. of Am. Soc. for Inf. Sci. 41(4): 288–297

Savoy J (2005) Comparative study of monolingual and multilingual search models for use with Asian languages. ACM Trans. on Asian Lang. Inf. Proc. 42(2): 163–189

Schäuble P (1997) Multimedia information retrieval. Content-based information retrieval from large text and audio databases. Kluwer Academic Publishers

Sedgewick R, Wayne K (2011) Algorithms, Fourth Edition, Section 3.4 "Hash Tables". Addison-Wesley Professional

Singhal A, Buckley C, Mitra M (1996) Pivoted document length normalization. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '96). ACM, New York: 21–29

Spärck Jones K, Willett P (1997) Readings in information retrieval. Morgan Kaufmann

Tague-Sutcliffe J (ed.) (1996) Evaluation of information retrieval systems. J. Am. Soc. for Inf. Sci. 47(1)

Walker S, Robertson SE, Boughanem M, Jones GJF, Spärck Jones K (1998) Okapi at TREC-6, automatic ad hoc, VLC, routing, filtering and QSDR. In Voorhees EM, Harman DK, (eds.) The Sixth Text REtrieval Conference (TREC-6). NIST Special Publication 500–240: 125–136

Ubiquity (2002) Talking with Terry Winogard. Ubiquity Magazine 3(23)

# Chapter 3
# Cross-Language Information Retrieval

*"Mind the language gap!"* [1]

**Abstract** Cross-Language Information Retrieval (CLIR) systems extend classical information retrieval mechanisms to allow users to query *across* languages, i.e., to retrieve documents written in languages different from the language used for query formulation. Chapter 3 discusses a range of options to implement CLIR: systems can be distinguished both by what they translate (the queries, the documents or a combination of both), and how they translate (by using machine-readable dictionaries, applying statistical approaches, or incorporating machine translation systems). Providing systems for CLIR remains difficult, with translation ambiguity and insufficient lexical coverage being among the most prominent problems. The quality of available resources for different languages also varies greatly. Combination approaches may help by leveraging multiple types of translation resources. Finally, the chapter covers the issues encountered when extending a CLIR system to handle many languages.

## 3.1 Introduction

Chapter 2, on within-language retrieval, has shown how the implementation of information retrieval mechanisms is essentially separated into two phases: an indexing phase (where indexing is an offline process for documents) and a matching phase. While we have discussed adaptations for different languages, the mechanisms presented in that chapter deal with monolingual document collections, making the collection only available for access in the language the documents were written in.

[1] Not attributable to a single source. A play on the London Underground announcement 'mind the gap'.

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_3, © Springer-Verlag Berlin Heidelberg 2012

In Chapter 3, we now discuss how to extend the Information Retrieval (IR) system to meet the needs of diverse *multilingual contexts*:

1. The document collection is monolingual, but users can formulate queries in more than one language.
2. The document collection contains documents in multiple languages; users can query the *entire* collection in one or more languages.
3. The document collection contains documents with mixed-language content, and users can query the *entire* collection in one or more languages.

In all three cases, the system has to manage a language mismatch between the query and at least part of the document collection. Systems that involve retrieval crossing a language barrier are commonly called *Cross-Language Information Retrieval* (CLIR) systems. Alternatively, the terms 'cross-lingual' or 'translingual' can be found in the academic IR literature. Some form of CLIR is necessary to cover all three cases noted above. Systems that allow retrieval in any or all languages, i.e., covering all the above cases plus within-language (monolingual) retrieval, are typically called M*ultilingual Information Retrieval* (MLIR) systems. More broadly, it is also possible to speak of *Multilingual Information Access* (MLIA) systems.

## 3.2 Implementation of Cross-Language Information Retrieval

### *3.2.1 Query Translation and Document Translation*

For those MLIR systems that incorporate CLIR, some provision needs to be made for crossing the language gap between query and documents. The obvious choice is to use some form of translation, either:

- of the query (*Query Translation* (QT); translation into the language of the documents);
- of the documents (*Document Translation* (DT); translation into the language of the query); or
- of both query and documents (combination of QT and DT; translation into an intermediary language ('pivot language')).

There is, however, a fourth, less obvious choice:

- translation of neither query nor documents (i.e., no translation; direct matching across languages).

We will briefly discuss this 'no translation' option in Section 3.2.2.

The choice between query translation and document translation is a crucial one; especially when the number of languages to be covered by the system grows. There are inherent advantages and disadvantages to both options. *Document translation*

allows shifting the whole translation process to the offline portion of the CLIR system, with retrieval proper taking place on the translated versions of the documents. This avoids any speed penalty caused by translation at retrieval time. When translating entire documents, ample context is usually available to disambiguate terms with multiple meanings. If the translation output is humanly intelligible, it can be reused for document presentation in later stages of the IR cycle. On the down-side, translating whole document collections implies a multiplication of the storage footprint of the system; a problem that increases in severity as the number of languages increases. Translation of full documents is also a time-consuming process, a problem that can be prohibitive especially with large and dynamic collections. Finally, translation systems potentially improve over the years; in order to take advantage of these improvements, and to maintain consistency, the document collection has to be periodically re-translated.

*Query translation* avoids these problems; storage requirements remain essentially unchanged from the monolingual scenario. Often, users form their queries from a (short) sequence of nouns in their base form (lemmas); these types of queries lend themselves well to efficient dictionary look-ups. However, a (substantial) performance penalty at retrieval time is still incurred, as the query has to be translated 'on-the-fly'.[2] As stated in Chapter 2, user queries tend to be very short (often just two or three keywords) and thus offer little context in which to handle ambiguous terms. This problem is often handled by using expansion techniques, as described in Section 3.3.3. When the CLIR system is used to produce mixed-language result lists, i.e., result lists computed for multilingual document collections, query translation implies a *merging step*, where the output from a number of retrieval steps for each of the languages needs to be combined. Document translation avoids this important issue, which is detailed below in Section 3.4.2.

### *3.2.2 No Translation*

There are special cases when the use of an explicit translation mechanism can be entirely avoided. Linguistics uses the term 'mutual intelligibility' to describe the phenomenon that there are some language pairs where speakers of one language may understand the other language despite never having received formal training in that language. The key to this phenomenon is a similar vocabulary, i.e., the languages share many cognates (words of common etymological origin with similar or identical surface forms). Examples include the Nordic languages Danish, Swedish and Norwegian. By using *substring matching*, e.g., by the forming of *character n-grams* (see Section 2.4.5), matches between query and documents can be established without using proper translation. For 'very similar' languages, the number of such

[2] It is possible to use a caching mechanism for very frequent queries.

substring matches will be quite high and offset accidental matches (McNamee and Mayfield 2004). Alternatively, a modified strategy is possible when using whole words: here the 'proper' cognates are supplemented with pairs of 'near-cognates', determined by using a 'pseudo' spelling correction algorithm on one of the languages (e.g., assuming English to be a form of 'mis-spelled' French) (Buckley et al. 1998). Finally, Gey (2005) points out that in some cases Kanji characters in Japanese queries can be used directly after a simple codepage mapping to search Chinese text, and vice versa. While the above are commonly called 'no translation' approaches, it can be argued that they still constitute special cases of simulating a 'translation effect'. These approaches typically do not extend beyond pairs of closely related languages. Even then, problems with so-called 'false friends' may occur. Consider, e.g., the word 'gift', which in English most commonly denotes a present or a talent, but in German is a translation of the English word 'poison'. Exploring these special 'non translation' cases further falls outside the scope of this book.

### 3.2.3 *Different Types of Translation Resources*

Options for incorporating translation into a cross-language information retrieval system are basically threefold:

1. The use of *machine-readable, humanly crafted dictionaries* (and thesauri, word-lists and other similar resources) (see Section 3.3.1);
2. The use of translation resources generated by *statistical approaches* from suitable training data (see Section 3.3.2);
3. The use of a 'full' *machine translation* system (see Section 3.3.4).

Combination approaches, using any two or all three of the above listed types of translation options, are also possible (see Section 3.3.5).

Implementation paths for cross-language information retrieval are more 'open' than for within-language information retrieval. Where for the latter the combination of the methods outlined in Chapter 2 (indexing, usually including stopword removal and stemming, followed by the use of an effective weighting scheme for retrieval) will lead to solid performance in most circumstances, it is more difficult to generalise for the cross-language case. In addition to making basic choices between query translation and document translation and deciding which type of translation resources to use, there are also a number of mechanisms to consider, such as query expansion techniques, which may be very beneficial in some scenarios but unsuitable for others. This chapter focuses on the techniques that are employed most widely, rather than going into detail about techniques that only apply to specific scenarios. CLIR has now matured to the point where, despite the remaining difficulties, 'blueprints', i.e., configurations that have been shown to lead to reasonably robust CLIR performance over a variety of operational settings, start to emerge. Retrieval on collections containing many different languages remains the

most difficult case (Mandl et al. 2008). We will comment in the conclusions to this chapter on the expected range of CLIR performance of state-of-the-art approaches for different language configurations and retrieval contexts (Section 3.5). However, we urge readers to familiarise themselves with the experimental settings used to derive the mechanisms by following up the references listed at the end of the chapter, and validating them against their own retrieval environments.

A 'perfect' translation mechanism would effectively reduce the CLIR problem to a combination of using the translation mechanism to translate either queries or documents, followed by employing a monolingual, within-language information retrieval system for indexing and matching. It was this theoretical scenario that motivated the division of the description of multilingual information retrieval into two chapters, covering within-language retrieval first in Chapter 2. In practice, however, translation introduces new problems into IR systems, which make simple translation-plus-monolingual-retrieval combinations problematic.

It is beyond the scope of this book to list all the linguistic phenomena that affect translation quality. It is also important to note that metaphors, cultural differences and highly specialised terminology make it very difficult or effectively impossible even for skilled human translators to produce accurate translations of arbitrary text. This issue comes prominently into play when implementing a cross-language IR system. Translation can be approached from two different angles in CLIR: the system can either be engineered to produce a grammatically correct translation that is as accurate as possible, or it can produce translated representations that are specifically tailored to be robust towards the matching phase (thus optimising retrieval effectiveness), e.g., by producing multiple translation alternatives that are weighted by the further retrieval process. It is important to note that CLIR is not a translation exercise in the linguistic sense; the goal is not to produce perfect translations of queries or documents, but rather to render them in other languages in order to make effective retrieval possible (sometimes also referred to as 'pseudo-translation', to highlight this fundamental difference). In this sense, the translation portion of CLIR is a simpler problem than general Machine Translation (MT) that aims to provide an accurate translation of a text that can be understood by human readers.

### *3.2.4 Term Ambiguity*

As discussed in Chapter 2 in Section 2.5.1, a 'bag of words' paradigm underlies the most commonly used vector space and probabilistic retrieval methods. Information on document structure is thus lost; the retrieval scores are based solely on the occurrence of individual features in the documents of the collection. The conceptually simplest option to bridge the language gap is thus to translate such bags word-by-word. This strategy is, however, hindered by problems with translation ambiguities. Ambiguity tends to be high when considering isolated single words. The 'word-by-word' approach also breaks down when the same concept is expressed with a

varying number of terms in different languages. Consider for example the German word 'Fussballweltmeisterschaft', which translates to 'football world cup' in English.[3] Obviously, the reverse phenomenon occurs when we exchange source and destination languages – a sequence of words such as 'football world cup' would then best be translated by a single German word. When a query addresses multiple concepts, this *many-to-many (m:n) mapping* may lead to a shift in the ratio of terms representing each concept in the source and target languages. To complicate matters further, a similar effect also occurs when multiple translations for ambiguous terms are used. As a consequence, the length of a piece of text in the source language, and the length of its representation in the target language may differ widely, both overall, and looking at individual concepts contained in the text. This effect is problematic for retrieval (see e.g., (Hiemstra and de Jong 1999) for a discussion).

## 3.3 Translation Approaches for Cross-Language Information Retrieval

### *3.3.1 Machine-Readable Dictionaries*

In spite of the above considerations, in the simplest implementations of CLIR, text is 'translated' (or 'pseudo-translated') between two languages by replacing each term in the bag of words representing the text by *all* possible translations according to a suitable machine-readable dictionary.

A main limitation of the dictionary approach is that often dictionary-type translation resources contain the base form (lemma) only for their entries. For query translation, the ensuing problem may be less pronounced, as users frequently form their queries from a (short) sequence of nouns or adjectives in their base forms ('keyword-type queries'). However, for lengthier queries and especially for document translation, provision needs to be made to allow matching between different word forms and dictionary entries. Stemming (see Section 2.4.5) is one possibility to enable such matches, but may add noise terms (improper conflation) or increase ambiguity. As a consequence, dictionary-based translation mechanisms are most often found in CLIR systems using query translation.

Even after successful look-up of terms in a dictionary, two basic problems remain: (1) the ambiguity of individual terms, and (2) the mapping between (multiword) expressions, as outlined in Section 3.2.4. These issues explain the poor performance that is observed if such a simplistic word-by-word translation strategy is employed. By simply choosing all possible translations, the number of terms in

[3] Alternatively, it can be translated to 'soccer world cup' in American English, illustrating the regional differences that can come into play even within the same language.

the destination language that is substituted for each term in the source language can vary widely. As an example, consider the Spanish query 'Contrabando de Material Radioactivo'.[4] Using the Merriam-Webster Spanish Online dictionary[5] to replace each word with all given translation alternatives (ignoring the stopword 'de'), we obtain 'smuggling, contraband; material, physical, real, equipment, gear; radioactive'; i.e., the term 'Material' was substituted by five different English terms, and thus dominates the resulting English query. If no provision during retrieval is made for this difference (such as down-weighting the different translation alternatives), highly ambiguous terms, or terms that have many possible translation equivalents in the target language, are given a significantly higher weight than others. Even worse, in the case of a query with multiple search terms, where one source language search term has many possible translations in the target language, documents that contain all the translation equivalents for this one query term but no translations for other search terms may receive higher retrieval scores than documents that contain a single translation of each search term, due to the independence assumption usually inherent in retrieval methods using the 'bag of words' paradigm. Finally, for highly ambiguous terms, unwanted matches are frequently introduced into the retrieval process, potentially degrading retrieval effectiveness.

The ambiguity/synonymy problem can be addressed in two different ways: either by introducing some kind of disambiguation component into the translation mechanism or by selecting and weighting all or multiple translation options proposed for a single search term.

There are simple implementations that always choose the most frequent translation as pre-determined using training data, without applying any real disambiguation. Such approaches tend to improve results compared to simple 'pick-all-translations' variants that include all translation equivalents with equal weight. However, if the same approach is extended to retain (all, or a subset of) the different translations, instead using the training data to assign a form of 'translation probability' that can be included in the retrieval method, even better results can often be obtained. This is a viable approach especially if query translation is used (Hiemstra and de Jong 1999).

Introducing a proper word sense disambiguation component requires substantial architectural changes to the indexing components, as disambiguation is typically based on analysis of the context of the ambiguous term, which is not available if only bag of words representations are used for retrieval.[6] An alternative can be the use of co-occurrence data. Consider the Italian word 'calcio', which stands for both 'football' and 'calcium' in English. Different words will frequently co-occur with this term depending on its underlying actual meaning in a given context. This co-occurrence analysis can be carried out offline during the indexing of the documents.

[4] This is the title of topic 264 from the CLEF ad-hoc track. The (official, human) translation to English is 'Smuggling of Radioactive Material'.

[5] http://www.merriam-webster.com/spanish/

[6] And very limited in the case of query translation.

**Table 3.1** An example of the use of query structuring to group translation alternatives (cited from Pirkola 1998)

| | |
|---|---|
| English original topic formulation | What research is ongoing to reduce the effects of osteoporosis in existing patients as well as prevent the disease occurring in those unafflicted at this time? |
| English keyword query | osteoporosis prevent reduce research |
| Humanly translated Finnish query | osteoporoosi ehkäistä lieventää tutkimus |
| English query translated from Finnish by machine-readable dictionary; structured | #sum(osteoporosis #syn(prevent avert obviate obstruct hinder) #syn(alleviate mitigate reduce weaken abate relieve ease lighten) #syn (examination exploration inquest investigation report research scrutinity study)) |

However, with proper word sense disambiguation being a complex linguistic problem, the 'pick-all-translations' variants coupled with corpus frequency analysis, which work reasonably well, are often chosen as an alternative.

A second option for addressing the ambiguity problem is the use of *query structuring*: the weighting scheme is adapted to weigh the different translation alternatives for a source language word as a 'set', effectively assigning all alternatives one combined weight. Some retrieval systems offer a 'synonym operator' to this effect. The use of such 'structuring' has been shown to be consistently beneficial for a range of European languages (Hedlund et al. 2004). Table 3.1 shows an example of a Finnish query translated by a machine-readable dictionary into English, using query structuring. The origin of the query is a topic description taken from the TREC evaluation campaign (topic 216 'Osteoporosis'), which, in preparation of the experiment, was transformed manually into an English keyword query. This query was then, again manually, translated into the Finnish source query. The CLIR system produces the translated English representation shown at the bottom of the Table, where translation alternatives for individual keywords are grouped by using the 'synonym operator' ('#syn') (Pirkola 1998). A discussion of possible implementation options for this operator in a CLIR setting can be found in Darwish and Oard (2003).

Dictionary-based approaches suffer from the practical impossibility of compiling dictionaries that cover all the theoretically viable (or even practically occurring) words of a given language. Reasons for this impossibility include the constant evolution of languages, the wide variability of proper names, and the richness of the compound formation process in some languages. In general, these approaches benefit from a large and regularly updated dictionary, thus limiting the number of *Out-Of-Vocabulary* (OOV) terms. Out-of-vocabulary terms are typically copied as-is into the translated representation, in the hope that a direct match will be obtained (see also Section 3.2.2). A somewhat paradoxical situation arises when the size of the dictionary is aggressively expanded to include additional very rare terms or very rare translation alternatives for a dictionary entry. Such expansion may well lead to degraded retrieval effectiveness, in the former case due to accidental matches with proper names, and in the latter case if the disambiguation or weighting mechanisms

do not work well, with many of the additional translation alternatives not capturing the correct meaning of the source term and thus introducing more noise during the retrieval phase (see e.g., (Hedlund et al. 2004)).

### 3.3.2 Statistical Approaches

An interesting approach to solving both the ambiguity problem and the multiword problem, as well as the OOV issue to some extent, is the use of statistical approaches. This includes both the generation of machine-readable dictionary-type resources that are enriched with statistical translation probabilities, as well as approaches that can translate whole queries based on their representation in the vector space (by essentially mapping the vector in the source language to a corresponding vector space in the target language). This second approach allows the translation of queries 'as a whole' and thus addresses some of the issues of the mapping of multiword expressions quite elegantly.

The main approaches, such as *Latent Semantic Indexing* (LSI) (Littman et al. 1998) and *Similarity Thesauri* (Qiu and Frei 1993), basically work by exploiting term co-occurrence information in bilingual text corpora that acts as training data. All these approaches can exploit training corpora containing parallel texts, i.e., corpora containing documents with high-quality, manual translations into several languages. A list of available *parallel corpora* can be found in Moreau (2009). Many of these corpora comprise texts from international organisations or official bodies of multilingual countries, such as the proceedings of the Canadian parliament,[7] legislative texts of the European Union[8] or texts collected by the United Nations.[9] In general, available parallel corpora often contain domain-specific terminology[10] and are limited to a core group of widely spoken languages. It is often possible to align the contents of parallel corpora on the sentence level.

In order to cover domains and languages for which few or no parallel corpora are available, some approaches have been extended to also work on *comparable corpora*. Such corpora contain monolingual texts covering the same topic areas in more than one language, with links between related texts (constituting a 'document-level alignment'). As an example, both the TREC and CLEF campaigns have distributed collections of news texts by the Swiss News Agency (SDA/ATS) from the time period of 1988–1990 and 1994/1995, respectively. These collections contain documents for three official languages of Switzerland (German, French and Italian). The texts in these languages form three separate monolingual corpora,

[7] Hansard English/French corpus, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20.

[8] Europarl corpus, 11 languages, http://www.statmt.org/europarl/.

[9] UN Parallel Text English/French/Spanish corpus, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T4A.

[10] In addition to legislative texts other examples include scientific abstracts and texts from the tourism domain.

**Table 3.2** Three examples of German and French document pairs derived from document-level alignment of the TREC SDA/ATS comparable corpora. The document titles are shown, with English translations in parentheses. While the documents are not direct translations, they clearly cover the same events[a]

| Title of German document | Title of French document |
|---|---|
| 'New York Times': Erster 'Sternenkrieg'-Laser getestet<br>('*New York Times': First 'Star Wars' laser tested*) | Etats-Unis: premier essai grandeur réelle d'un puissant laser de l'IDS<br>(*United States: first full-scale test of a powerful SDI laser*) |
| Goria zu Besuch in Malaysia<br>(*Goria on visit in Malaysia*) | Le premier ministre italien Goria en visite en Malaisie<br>(*Italian Prime Minister Goria on visit in Malaysia*) |
| Condor-Maschine bei Izmir abgestürzt: Mutmasslich 16 Tote<br>(*Condor plane crashes near Izmir: presumably 16 deaths*) | Un avion ouest-allemand s'écrase près d'Izmir: 16 morts<br>(*A Western German plane crashes near Izmir: 16 deaths*) |

[a] As quoted from Braschler and Schäuble (1998), as well as from unpublished work by the author.

with few direct translations, but cover many of the same local, national and world events (often to a different extent). By performing document-level alignment, subsets of the collections can be used to form comparable corpora. Unfortunately, no such inter-language links are provided by the publisher of the texts.

Parallel and comparable corpora, however, can also be acquired (semi-) automatically. Comparable corpora can be built by starting with collections of documents covering the same domain and (if applicable) the same time-frame. Alignment is then carried out by using indicators of relatedness such as a similar publishing date, occurrence of proper names, cognates and numbers, as well as any potentially available classifiers (Braschler and Schäuble 1998). Some examples of pairs of German and French documents that can be derived automatically in such a way from the TREC SDA/ATS collections are given in Table 3.2. A rich source for collecting pairs of translated text is the World Wide Web. With many offerings on the Web today reaching a global audience, increasing numbers of web pages are translated into multiple languages. These translations can be identified by using special characteristics, such as hyperlinks with special descriptions ('click here for the French version') or filenames ('home-fr.html') (Nie et al. 1999). More recently, there has been work on how to build comparable corpora based on Wikipedia[11] texts. Many articles in Wikipedia contain links to articles about the same topic in other languages. By pairing such linked documents for the languages with the most extensive coverage, comparable corpora with more than 100,000 document pairs that cover a wide range of topics can be constructed (see e.g., (Yu and Tsujii 2009)). Corpora built using these methods contain a varying number of erroneous alignments. The robustness of the mechanisms that exploit these 'noisy' comparable corpora is not well documented.

[11] Wikipedia is a multilingual, online, community-based encyclopedia. See http://www.wikipedia.org/.

**Table 3.3** Two sample entries from bilingual similarity thesauri[a]

| English → French: offer | French → German: incendie (*fire*) |
|---|---|
| 4.7494 offre (*offer*) | 0.6321 brand (*fire*) |
| 4.5737 offert (*offered*) | 0.5130 feuer (*fire*) |
| 4.2255 comptant (*in cash*) | 0.4223 feuerwehr (*fire department*) |
| 4.0475 pret (*loan*) | 0.4160 brandstiftung (*arson*) |
| 3.8256 opa (*tender offer*) | 0.3865 flammen (*flames*) |
| 3.7980 faite (*done*) | 0.3703 brandursache (*cause of fire*) |
| 3.7215 prevoit (*calculates*) | 0.3686 sachschaden (*material damage*) |
| 3.5656 echec (*failure*) | 0.3623 braende (*fires*) |
| 3.5602 intention (*intention*) | 0.2779 brannte (*burnt*) |
| 3.5171 engage (*hire*) | 0.2581 ausgebrochen (*broke out*) |

[a] Quoted from unpublished work by the author.

Nie et al. (1999) report mixed results where, for some queries, they observed no degradation in retrieval effectiveness when using a quasi-parallel corpus automatically acquired from the Web versus a manually built one; but for other queries they reported performance losses which were directly attributable to errors in the web-based corpus. A superficial analysis of the alignment quality of the corpus automatically derived from SDA texts used for the examples that are given in Tables 3.2 and 3.3 indicates that at least 10% of the alignments were erroneous.

The data structures derived from training on parallel and comparable corpora are often comparatively large, covering several hundred thousand pairs of terms. When building translation resources on such training data, the translation probabilities will reflect the properties of the underlying documents, with e.g., a term such as 'strike' translated differently depending on whether documents covering sport or economic news are used. For an example of translations derived from bilingual similarity thesauri, see Table 3.3.

Both Latent Semantic Indexing and Similarity Thesauri can also be used for within-language retrieval (and were indeed at first exclusively used for this purpose). In this case, a *monolingual data structure* is built which is an alternative resource for query expansion or term conflation. In the monolingual case, it is possible to directly train on the same collection that is used for searching, avoiding issues of domain shift. Compared to relevance feedback, the expansion terms derived from this resource are query-independent, reflecting the (training) corpus as a whole; whereas the selection of expansion terms during relevance feedback is guided by the use of candidate terms in the top-most retrieved, human judged documents (and thus is very much query-dependent).

Latent Semantic Indexing works by loosening the assumption that all indexing features are mutually independent, instead mapping features that co-occur in similar contexts (i.e., the same documents) to locations that are close in the vector space. This is achieved by a dimensionality reduction of the vector space through Singular Value Decomposition (SVD). In the multilingual case, this training takes place on concatenated documents that contain sections of text in each of the languages to be covered (Littman et al. 1998).

The Similarity Thesaurus is built by essentially 'inverting' the retrieval process. Instead of documents, terms are retrieved. While information retrieval usually returns a list of documents estimated to be the most relevant for each query, for similarity thesaurus construction, a list of terms estimated to be most 'similar' for each term is returned instead. Common weighting schemes can be adapted for this inverted scenario, such as tf.idf-Cosine (see Section 2.5.5). For the construction of multilingual similarity thesauri, multilingual training documents are constructed in much the same way as for LSI (Sheridan et al. 1997).

Table 3.3 shows two sample entries from bilingual similarity thesauri. On the left, an example of an entry in a English-French similarity thesaurus built on approximately seven months of texts from the Associated Press news agency and the Swiss news agency SDA (and thus reflective of terminology used by a newswire) is shown. Some of the texts in English and French contain comparable content (and possibly quotations), but the texts are generally written by two independent organisations. Roughly 8,000 pairs of automatically aligned documents were used to build this thesaurus (noisy comparable corpus). On the right, an example of an entry in a French-German similarity thesaurus built on approximately 3 years of texts from the Swiss news agency SDA (and thus again reflective of terminology used by a newswire) is shown. The texts in French and German contain comparable content, but there are few direct translations. For this thesaurus, roughly 80,000 pairs of automatically aligned documents were processed (again forming a noisy comparable corpus). Scores of the two samples are given to illustrate the nature of the data structures produced, but are not comparable due to different training data.

An elegant way to incorporate translation probabilities from resources built by statistical approaches are weighting schemes based on language models, as discussed in Section 2.5.8. Please refer to that section for the basics of the monolingual form of these schemes. The schemes can be extended to cross-language information retrieval by directly replacing the probability $P(\varphi_i|M_j)$, with $\varphi_i$ being a term in the language of the query, by a weighted sum of the probabilities of its possible translations in the document language (see Table 2.4 in Chapter 2 for notation):

$$\begin{aligned} P(q|M_j) &= P(\varphi_1, \varphi_2, .., \varphi_n|M_j) \\ &= \prod_{i=1}^{n} \left( \sum_{j=0}^{k-1} P(\varphi_i|T_{i,j}) \left( (1-\lambda) P(T_{i,j}) + \lambda P(T_{i,j}|M_j) \right) \right) \end{aligned}$$

where each query term $\varphi_i$ has $k_i$ possible translations $T_{i,j}$ $(0 \leqslant j \leqslant k_i - 1)$ in the document language (Hiemstra and de Jong 1999).

Finally, the building of an intermediate data structure, such as a reduced dimension vector space (LSI) or a similarity thesaurus, can also be entirely avoided, and the aligned corpora instead exploited by evaluating the original language query on the aligned parallel or (noisy) comparable corpus first. This first retrieval step leads to a result list consisting of document pairs in the bilingual case, from which a new query in the target language can be constructed by using blind relevance

feedback (see Section 2.5.6). The query constructed from these target language expansion terms can be used as a 'pseudo translation' of the original query in a second retrieval step (Braschler and Schäuble 1998).

### 3.3.3 Pre-translation and Post-translation Query Expansion

Approaches using humanly crafted, machine-readable dictionaries can be combined with a statistical approach by using a relevance feedback mechanism. As explained earlier, 'out-of-vocabulary' terms present a serious problem, as these terms will go untranslated. To underscore this point, remember the discussion of the frequency of occurrence of words in textual corpora in Section 2.4.4: many word forms are used only very rarely, but collectively make up the significant majority of text. The fact that dictionaries usually only cover base forms is the lesser of two associated problems: good word normalisation components can help to map some of these word forms to the appropriate dictionary entries. However, these normalisation components would have to generate linguistically correct base forms, which is not the case when using simple rule-based stemming mechanisms during indexing for the generation of features (compare the Porter stemmer, Section 2.4.5). The deeper issue is that even if such a mapping is successfully implemented, inclusion of all possible base forms in a dictionary is not feasible. The ever evolving nature of languages, and the prevalence of proper names, some of which should be translated (e.g., 'Venezia' in Italian translates to 'Venice' in English, 'Venise' in French, 'Venedig' in German, etc.), while others should not (e.g., 'Waldheim' – the surname of the former UN secretary-general and Austrian president, which literally translates to 'Forest home' in English) make the compilation of an exhaustive dictionary for general purposes impossible.

The OOV problem is especially prominent when using query translation: an untranslated search term often implies the loss of entire concepts in the translated query. Failure to translate a word in the query can thus easily lead to failure to retrieve any useful information. In order to somewhat counter this OOV problem, the use of *query expansion* is often proposed. The rationale is that by expanding the query to include additional related terms, the probability of loss of a query concept is decreased. Query expansion is possible both prior to the translation (pre-translation) and after the translation (post-translation). *Pre-translation query expansion* directly strives to alleviate the OOV problem: a blind relevance feedback mechanism (see Section 2.5.6) is used to replace the query with an expanded query that contains the original query terms (this is optional) plus a number of expansion terms (usually in the range of 20 to 50 new terms) that best statistically add to the initial query as a whole.[12] By introducing these expansion terms, which are often synonyms or related terms, the probability that whole concepts of the query go untranslated decreases.

[12] The relevance feedback mechanism will effectively select terms that best differentiate between the most highly ranked documents after initial retrieval (before the feedback loop) and the rest of the collection.

However, the use of this technique requires the availability of a textual corpus in the query language to be used for the expansion that is suitably similar to the documents that are to be accessed in the target language. This requirement may be difficult to meet when handling less frequently-spoken languages, and even when such a corpus is available implies additional storage and handling. *Post-translation query expansion* tries to address problems caused by the translation component picking the wrong translation alternatives. By again using blind relevance feedback to expand on the translated query as a whole, additional synonyms are included and bad translation choices are down-weighted based on their dissimilarity to the overall query concept. Both pre-translation and post-translation query expansion have been shown to be useful tools to increase retrieval effectiveness (Ballesteros and Croft 1997).

Table 3.4 shows an example of pre-translation and post-translation expansion, as given in (Ballesteros and Croft 1997). They used a query taken from the TREC evaluation campaign (from the 'Spanish track'), which was translated using a machine-readable dictionary. A maximum number of 20 terms was used for the expansion. Note that, in this example, original query terms were not necessarily retained if their weight after expansion fell below the 20 term threshold (e.g., the term 'Mexico' is no longer present in the final query representation). The system that was used supports phrase matching (phrases are given in round brackets) and outputs multiple translation alternatives where applicable (in square brackets).

A study by McNamee and Mayfield (2002) links the effectiveness of pre- and post-translation query expansion to the quality of translation resources, by artificially degrading the quality of dictionaries and statistical translation resources. They show

**Table 3.4** Example for pre-translation and post-translation expansion. Quoted from Ballesteros and Croft (1997). The translations in italic have been added for the reader's reference and are not included in the original paper

| Query in English | the economic and (commercial relations) between Mexico and Canada |
|---|---|
| After pre-translation expansion | economic (commercial relations) mexico canada mexico free-trade canada trade mexican salinas cuba pact economies barriers |
| After pre-translation expansion and translation to Spanish | [económico equitativo][comercio negocio tráfico industria] [narración relato relación][Méjico México] Canadá [Méjico México][convenio comercial][comercio negocio tráfico industria] zona cuba salinas<br>*[equal economic][trade business traffic industry][narrative story relationship][Mexico Mexico] Canada [Mexico Mexico][trade agreement] [trade business traffic industry] area cuba salinas* |
| After pre-translation expansion, translation to Spanish and post-translation expansion, stemmed | canada (liber comerci) trat ottaw dosm (acured paralel) norteamer (est un)(tres pais) import eu (vit econom) comerci (centr econom) (barrer comerc)(increment subit) superpot rel acuerd negoci<br>*Canada (free trade) treaty Ottawa 2000 (side agreement) north American (united states) (three countries) import eu (vital economy) trade (central economy)(trade barrier)(sudden increase) superpower relationship agreement business* |

that the use of these expansion techniques can counter loss in retrieval effectiveness incurred through an (artificial) shrinking of lexical coverage. Indeed, they demonstrate very large performance gains that in some cases more than compensate for the larger coverage of the original, non-degraded translation resources. They consequently urge the use of such expansion techniques, especially when coverage of translation resources is poor.

### *3.3.4 Machine Translation*

The third major option to integrate translation in a retrieval system is the use of a *Machine Translation* (MT) component. While this would seem an obvious idea, there are a few major issues to consider when integrating MT, especially for query translation. First, queries sent to an information retrieval system are rarely full, grammatically correct sentences, which is what MT systems have been tailored to handle; rather, they are usually a sequence of search terms. Also, queries are typically short, and thus provide little context for translation. This difference can have a major impact on the effectiveness of the word sense disambiguation components integrated in MT systems. Second, much like approaches using machine-readable dictionaries, MT systems also suffer from the OOV problem. Furthermore, pre-translation and post-translation expansion strategies do not naturally combine with an MT approach, as they produce only long, non-grammatical sequences of additional search terms. It is also hard to tightly integrate MT systems, as they typically provide for end-to-end translation only: the input in the form of the text in the original language is replaced with the output in the target language, without giving any additional information on translation probabilities or weights. Because of this last issue, MT systems are often used to produce CLIR systems where the translation is used before the indexing phase, reducing the subsequent indexing and retrieval to monolingual problems (loose coupling).

For example, the internet search service Yahoo! separates the CLIR problem into a distinct translation and search phase (see Figure 3.1): first the user translates the query, and then they can start a web search.

### *3.3.5 Combination Approaches*

Experiments in the context of the CLEF evaluation campaigns have shown great promise for approaches that combine different types of translation resources. Most commonly, resources generated from statistical approaches are combined with either machine translation or machine readable dictionaries, but arbitrary combinations of these types are possible. *Combination approaches* have the potential to:

- Increase lexical coverage, e.g., by adding domain-specific dictionaries or statistically generated translation resources such as similarity thesauri which,

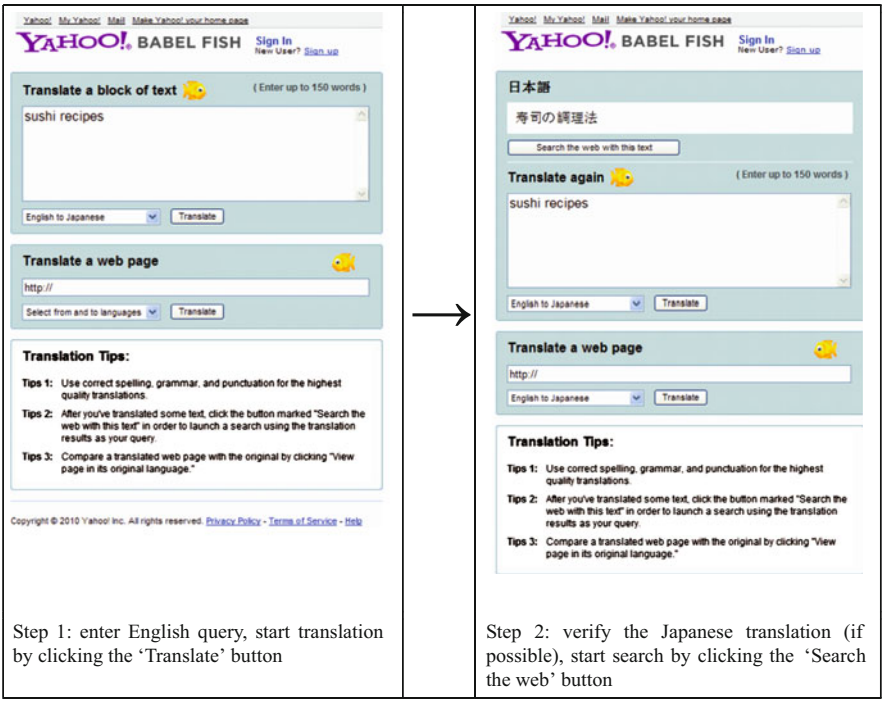


**Fig. 3.1** Example of a cross-language search service as offered by Yahoo! This screenshot is copyright of Yahoo!

depending on the training, can cover terms not available in other translation resources, such as proper names and newly emerging terminology.

- Improve robustness with respect to negative outliers, by decreasing the probability of negative outliers being highly ranked due to a defect in a single translation mechanism (e.g., the choice of the wrong translation alternative).
- Broaden the applicability of the system, by extending the number of domains or languages that can be covered (otherwise restricted to whatever is available for the type of translation resource chosen).

When using different types of resources, it is not normally possible to merge the resources directly, due to their different properties (e.g., a similarity thesaurus cannot be merged into a humanly generated machine readable dictionary, as the entries in the former are assignments of statistically similar terms, qualified with a similarity score, while the latter contains linguistically correct translation pairs). Instead, combination systems implement multiple retrieval processes, each using one type of translation resource, and ultimately merge the outputs of these steps into a single result list.

The problem of *merging* multiple result lists calculated on the same collection of documents is also known as 'data fusion', and is related to, but different from, the problem of merging multiple result lists calculated on arbitrary (and in particular

often disjoint) document collections (also called 'collection fusion'). For the latter problem, which often comes into play when many languages are handled simultaneously, see Section 3.4.2 below. The main issue in merging results obtained using different types of translation resources is the comparison of the retrieval scores (Retrieval Status Value(s) (RSVs)) established by the different retrieval steps. While the weighting schemes, such as tf.idf-Cosine and BM.25 (see Sections 2.5.5 and 2.5.7, respectively) establish a ranking in decreasing order of estimated probability of the relevance of the documents with respect to a query, the scores that are returned do not correspond to actual probabilities of relevance. Specifically, the rankings are merely order-preserving with respect to a theoretical ranking using 'real' probabilities. The absolute values are influenced by many factors that are specific to the weighting schemes (such as which collection statistics are used for the calculation). However, considering that the result lists cover the same underlying document collection, the lists will be at least in part a reordering of each other (with some extra items that do not appear in all the lists). This characteristic of the result lists lends itself to a merging strategy based on document ranks. There is nothing inherently 'multilingual' about this form of the merging problem: instead of merging multiple result lists stemming from different types of translation resources, merging of lists from retrieval with different weighting schemes (e.g., BM.25 and Lnu.ltn) or with different query formulations (e.g., stemmed and character *n*-gram), is equally possible. For this latter problem, see e.g., (Belkin et al. 1994), who propose the use of the median rank a document receives in the different result lists as the new 'score' for the combined list. As an alternative to using the ranks, Fox and Shaw (1993) give various ways to sum up the retrieval scores of the different lists. Since collection fusion, i.e., merging of result lists stemming from arbitrary collections, is generally a harder problem than data fusion, and as the latter problem can be viewed as a special case of the former, all the methods presented for the merging across languages problem in Section 3.4.2 can be adapted for data fusion purposes as well.

## 3.4 Handling Many Languages

### *3.4.1 CLIR Flows*

A special problem arises when the document collection contains documents in (possibly many) different languages. As mentioned earlier, it is possible to translate the queries, the documents, or both. Generally speaking, it is possible to use all types of translation resources (machine-readable dictionaries, resources generated from statistical approaches and machine translation) for both query translation and document translation, although some of the methods covered so far in this chapter lend themselves more naturally to one or other of the alternatives

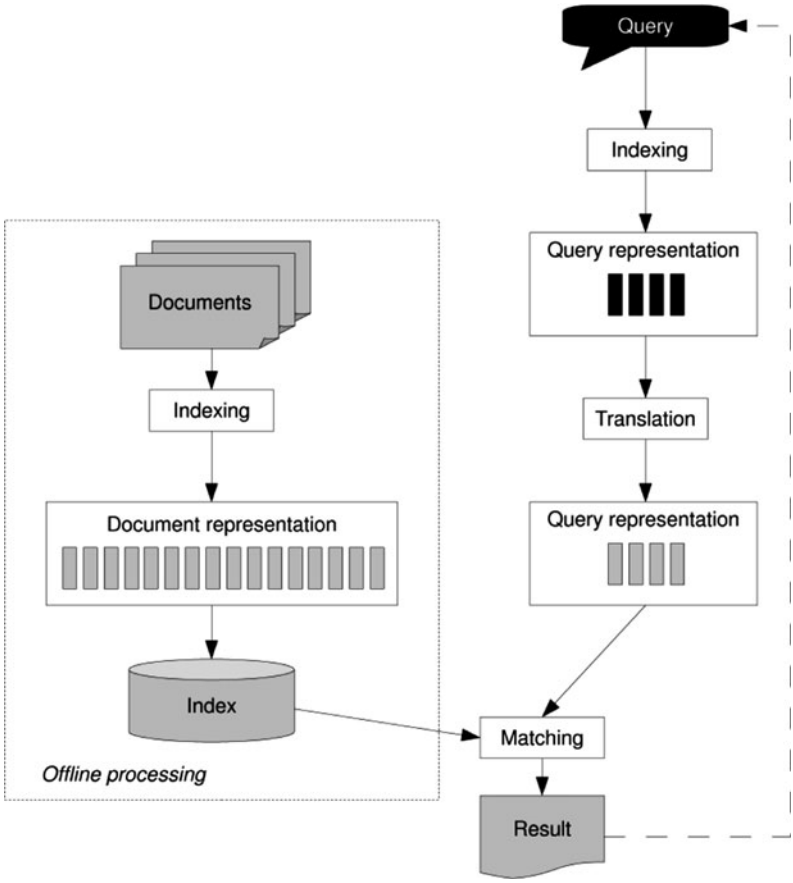


**Fig. 3.2** A basic system for bilingual CLIR. The flow is very similar to the monolingual case, with the addition of query translation. In some systems, the order of the indexing and translation steps are reversed

(e.g., pre-translation query expansion for query translation approaches). The decision on what to translate will influence the combination of the different indexing, matching and translation components. Figures 3.2–3.4 show three different possibilities.

Figure 3.2 shows a bilingual system based on query translation. This is possibly the simplest widely-implemented extension of an existing within-language system.[13] The system translates the query, and subsequently operates in the same way

[13] Ignoring the possibility of completely separating the translation from the IR system by pre-translating all documents and then using an unmodified monolingual IR system – an option which quickly becomes impractical if many languages need to be handled.

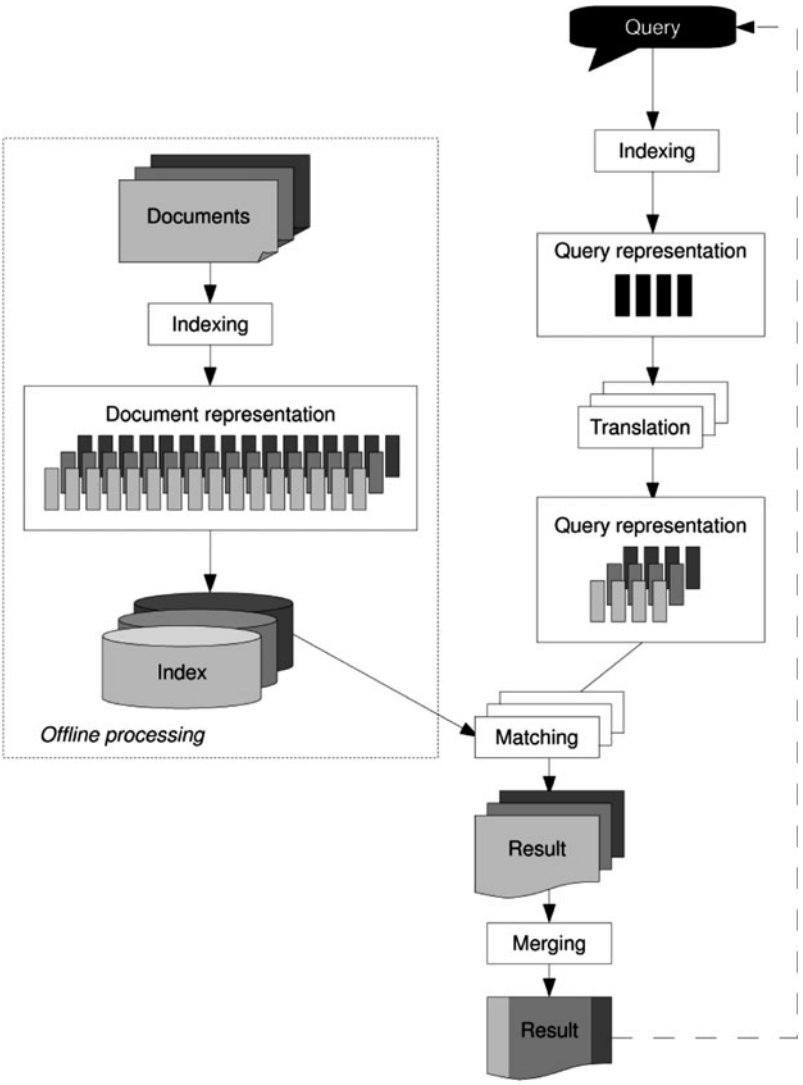


**Fig. 3.3** A system handling many languages via query translation. Systems can either use a set of monolingual indexes (pictured) or a single, unified index. Retrieval is in the form of a series of bilingual matching steps. A merging step is necessary to unify the result lists

as the basic within-language retrieval system shown in Figure 2.2 in Chapter 2. Note that the 'indexing' and 'translation' steps of the query can also be reversed in some system architectures (e.g., when incorporating machine translation).

Figure 3.3 shows a more advanced set-up, catering for a document collection containing many different languages. Documents are indexed in all of these

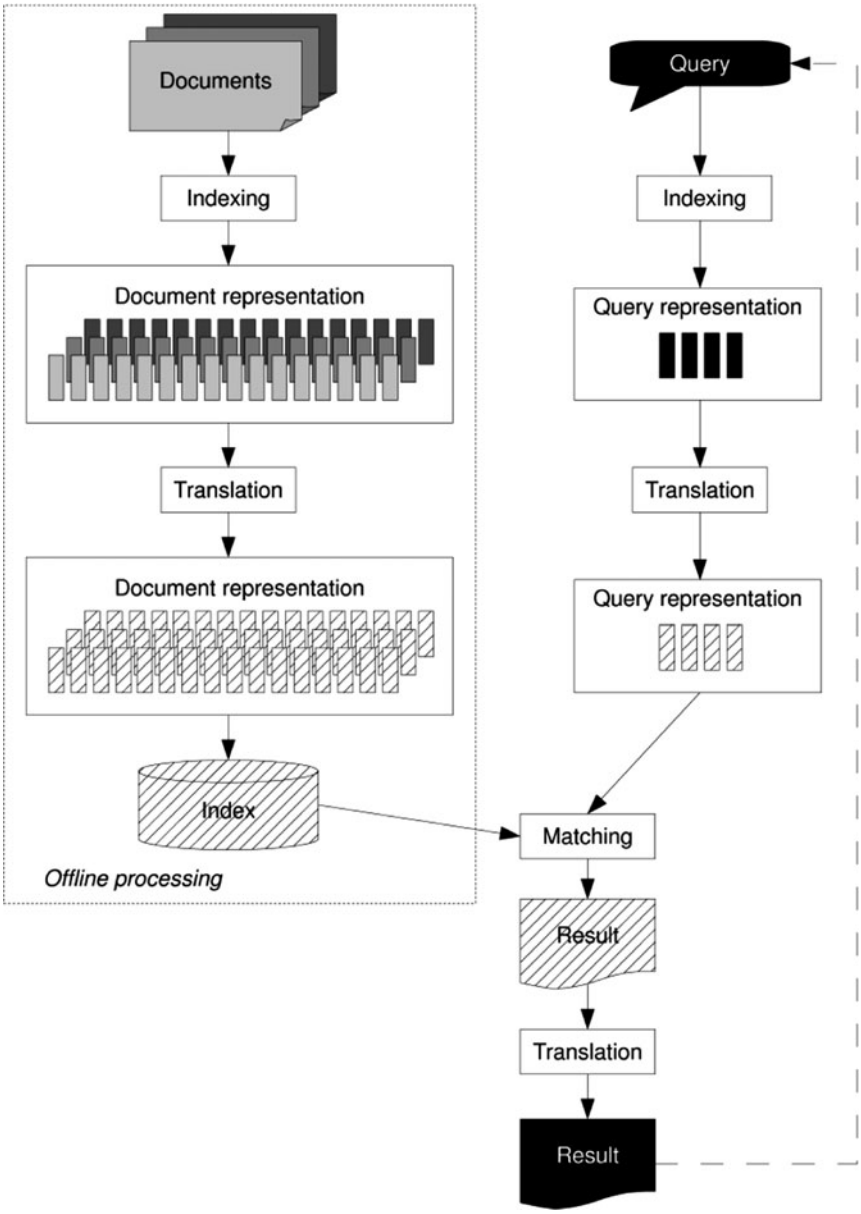


**Fig. 3.4** The flow of a system that uses a pivot language for matching. The pivot language need not necessarily be one of the document or query languages. Both documents and queries are translated to the common pivot language

languages, potentially using language-dependent indexing components (see Section 2.4 in Chapter 2). As a result, a set of different index structures (inverted indices), one for each language, is built. Alternatively, a unified index containing documents

from all languages can be built, but the system then needs to be able to compute its internal statistics taking this multilinguality into account. For retrieval, the query needs to be translated into all the languages, and a number of matching steps needs to be carried out. The result, a set of ranked lists, is finally merged in order to present the user with a single, unified result. Result list merging from disjoint collections is discussed below in Section 3.4.2.

Figure 3.4 shows an approach using the translation of documents instead of queries. In this approach, all documents are translated into a single language ('pivot language'), which may or may not be included among the set of languages used in the documents. The query is translated only once, into the pivot language (or, in bilingual CLIR cases, when no pivot language is necessary, not translated at all). The match is between the document and query representations in the pivot language and, after back translation (if necessary), a ranked result list in the user's preferred language is returned.

There are a number of other possible ways to combine indexing, translation and matching, but the three flows given as examples should offer a good understanding of the basic interaction between these components.

### *3.4.2 Merging Across Languages*

Figure 3.3 introduces a new problem into the multilingual information retrieval process that is not present in the bilingual CLIR case: how to merge the output from a number of different retrieval steps on disjoint document collections? The system initially produces one result list for each language and its underlying collection. This problem, also called *collection fusion* is related to the problem of data fusion for CLIR approaches that use a combination of different types of translation resources, as described in Section 3.3.5. As previously discussed in that section, the absolute scores obtained from the different result lists are influenced by the statistics of the underlying collection (e.g., through the use of the idf values) and thus are not comparable across different document collections or even across different queries. The fact that the scores from the multiple matching processes outlined in Figure 3.3 cannot be easily compared complicates this 'merging' or 'fusion' step still further.

The 'merging' or 'fusion' problem can be addressed in essentially two ways: (1) by using approaches that largely ignore the scores returned by the matching phase, and only focus on the ordering of the documents in the multiple lists to be merged, and (2) by using approaches that try to 'map' the scores, making them (more) comparable.

With the first option, 'interleaving' is the simplest choice: the multiple single-language lists are replaced by a multilingual, unified list where one document from each list is taken in turn, according to their rank in the original list. The scores are ignored beyond their role in establishing the ranking in the monolingual lists. This option suffers from the implicit assumption that all lists should contribute equally to the unified result lists. If the distribution of relevant items is skewed across the

different languages, significant performance loss will occur. There are variations of this strategy that draw from the individual lists relative to the proportion of documents of the corresponding language in the overall multilingual collection. These bring only small improvements, as the problem of skewed distribution of relevant items is a query-dependent one. There are also approaches that ignore the problem of incompatible scores, and merge the lists based on the 'raw scores' ('raw score merging'), e.g., by 'summing up' the scores (Fox and Shaw 1993). Again, this strategy has been shown to be too simple to solve the merging problem, and the performance is correspondingly poor (Savoy 2004). To verify the effectiveness of these simple strategies, it is possible to retroactively calculate an optimal merging result for training queries based on human assessments of relevance. On this basis it can be shown that all simple approaches to merging, such as 'raw score merging', 'interleaving', and their normalised counterparts, substantially underperform the (theoretical) optimal merging strategy, attaining a maximum effectiveness of roughly 66% in terms of average precision compared to the theoretical, optimal ranking (Braschler 2004a).

When trying to map the scores, a number of new approaches are possible. The most direct approach is to normalise the scores, to have them on the same scale. A comparison of some normalisation options can be found in Savoy (2004). The most successful of these was found to be 'combRSVnorm', defined as:

$$\text{combRSVnorm}(d_i) := \sum_{j=1}^{n} \left( \frac{\text{RSV}_\text{j}(d_i) - \text{MINRSV}_\text{j}}{\text{MAXRSV}_\text{j} - \text{MINRSV}_\text{j}} \right)$$

where the sum is calculated over the different original scores ($\text{RSV}_\text{j}$) that a document $d_i$ was assigned in the $n$ result lists to be merged, with $\text{MAXRSV}_\text{j}$ and $\text{MINRSV}_\text{j}$ being the highest and lowest score respectively of any document included in that list (in practice, for $\text{MINRSV}_\text{j}$ the score of a document at some cut-off point is used, e.g., the score of the document at rank = 1,000). A more elaborate approach calculates new scores by machine translating a sample of the top-ranked documents for each language to a common language, and then building a small, temporary, monolingual index of this document sample. On this new index, the scores are then recalculated, giving comparable scores. Clearly, this solution incurs a substantial computational penalty. Alternatively, instead of translating the top-ranked documents at retrieval time, a sample of documents from each language can be translated prior to retrieval, and a small index of these translated documents is then used to determine normalising factors for each language, either by using a set of training queries (obtaining a query-independent factor for each language), or by running the individual queries on the smaller index in addition to retrieval on the multilingual index (obtaining a query-dependent factor for each query and each language) (Si and Callan 2005). If it is possible to train on a set of queries by producing reasonably complete relevance assessments for them, logistic regression can be used to map the retrieval scores to real estimated probabilities, which then

are directly comparable. Savoy (2004) provides details of such an approach, and suggests summing the resulting estimated probabilities to obtain a score for the merged list. This approach appears able to outperform simpler methods based on interleaving or raw score merging, but its applicability is clearly limited due to its reliance on a set of training queries.[14]

### 3.4.3 Document Translation

Optimal merging remains an unsolved problem and the substantial performance degradation incurred can be an incentive to instead implement document translation should the computational and storage cost be deemed acceptable. This option is outlined in Figure 3.4. Using document translation implies the translation of the full document collection into a common pivot language, thus substantially increasing storage requirements. Typically, machine translation systems are used for this translation step.

### 3.4.4 Indirect Translation

The use of a pivot language can also be interesting in the case when no translation resources are available for direct translation between two languages. This situation most likely occurs when a pair of less-frequently spoken languages needs to be handled. In such cases, the only solution may be a two-step translation process that transforms the original input into a pivot language rendering, and only then translates the intermediary version into a representation in the destination language (indirect translation). In general, this option should only be chosen when no suitable direct translation resources are available, as it greatly increases the problems introduced by translation, with the second translation being based on a potentially faulty first translation. However, somewhat counter-intuitively, the use of a pivot language may increase retrieval effectiveness when the direct translation resources available are of much lower quality than the translation resources to the pivot language. This may be the case especially when using English as the pivot language, where a translation process L1 → English → L2 may benefit from better translation resources between L1/English and English/L2 than a direct translation process for L1 → L2 (Savoy and Dolamic 2009). Another interesting option is the simultaneous use of two different pivot languages in a process termed lexical triangulation. For example, by translating L1 → English → L2 and L1 → French → L2, two different representations of the original query in the destination

[14] The experiments cited in Savoy (2004) are based on the CLEF 2002 test collection. For training, the CLEF 2001 test collection was used.

language L2 can be obtained. These two representations can then be used to detect some of the issues with translation ambiguities between the languages, implicitly assuming that these ambiguities would occur in different contexts for the two different pivot languages (Gollins and Sanderson 2001). A major drawback of this idea is the added complexity of obtaining translation resources for two pivot languages, as well as the performance degradation incurred by using multiple translation steps.

## 3.5 Summary and Future Directions

This chapter has described the incorporation of mechanisms into an information retrieval system to allow querying across languages (Cross-Language Information Retrieval (CLIR)). While there are special cases where CLIR can be implemented without an explicit translation step (Section 3.2.2), usually some form of translation, either of the queries, the documents, or both, is incorporated into the system architecture (see Sections 3.2.1, 3.2.3, and 3.4.1).

While there is a relatively clear consensus on how to implement within-language retrieval for monolingual document collections, as described in Chapter 2, and while the indexing steps and weighting schemes described in that chapter are broadly applicable, there is a wider range of approaches to consider for implementing CLIR. The choice of which type(s) of translation resources to use, machine-readable dictionaries (Section 3.3.1), resources generated by statistical approaches (Section 3.3.2) or machine translation (Section 3.3.4), will also depend on the availability of such resources for the language pairs to be covered by the system. Combination approaches have shown great promise and may allow leveraging many different translation resources for the different languages (Section 3.3.5).

Cross-language information retrieval has not reached the same level of maturity as within-language (monolingual) information retrieval. While much progress has been made, as measured in evaluation campaigns such as CLEF and TREC, several issues remain. In the early years of CLIR research, *circa* 1997, the effectiveness of bilingual IR experiments was typically measured at 50–60% of that of comparable monolingual experiments. By 2003, this had improved to consistently better than 80% for the most commonly used language pairs (Braschler 2004b). Use of machine-readable dictionaries was most prevalent at that time, with combination approaches yielding the best results and shown to be reasonably robust, but incurring substantial additional cost by exploiting multiple translation resources. A study by the European Union FP7 Coordination Action TrebleCLEF attempted to generalise from the academic experiments reported on (mainly) in the CLEF evaluation campaigns, and also recommended the use of approaches that combine multiple types of translation resources, not least to alleviate problems of vocabulary coverage (Braschler and Gonzalo 2009). More recently, an analysis of CLEF 2009

results[15] has shown bilingual experiments obtaining an effectiveness of up to 99% of monolingual baselines, mostly through using the Google Translate MT service, and for combinations of widely-spoken Western European languages. Ferro and Peters (2010) ask in their discussion of these results: *"can we take this as meaning that Google is going to solve the cross-language translation resource quandary?"* Of course, the wide disparity in the quality and availability of translation resources for different language pairs and retrieval contexts, and the very different requirements that specialised vocabularies introduce to the CLIR problem, means that we are still far from 'one-approach-fits-all'. The excellent results from CLEF 2009 indicate that for the most frequently-spoken languages, when available, well-tuned MT systems can be a great fit. Combination approaches become more attractive when resources get more scarce, or vocabularies get more specialised, as they help by leveraging as much from the resources as possible. They may also increase the robustness of the system with respect to negative outliers. When very few or only poor resources can be obtained, 'no translation' approaches form a last resort; these have been shown to peak at approximately 50–60% of comparable monolingual performance, if queries are suitably long and contain named entities.

Looking at the implementation of mechanisms using these individual types of translation resources, machine translation allows a very loose coupling of translation with the rest of the retrieval mechanism, while machine-readable dictionaries are potentially available for specific domains of texts to be covered, and can be combined with query expansion techniques to address problems of vocabulary coverage (see Section 3.3.3). Finally, resources generated through statistical approaches reflect the vocabulary of their underlying training sets and therefore, if matched well with the document collection used for querying, can again help with out-of-vocabulary issues.

When using multiple types of translation resources to create intermediate retrieval results, issues of data fusion arise (see Section 3.3.5). The weighting schemes covered in Chapter 2 rank documents in the order of their estimated probability of relevance, but the absolute scores are not comparable across different retrieval methods or even different queries. We have listed a number of possibilities that allow the combination of multiple result lists that were produced on the same document collection. Related to this is the problem of collection fusion. Systems covering more than two languages via query translation are usually implemented by combining the output from multiple bilingual retrieval processes. Again, the scores of the intermediate bilingual retrieval runs are not directly comparable. The 'merging' of these intermediate results is an even harder problem than data fusion, as the scores are based on disjoint document collections (one collection per language). We discuss the merging problem in Section 3.4.2. Document translation avoids the merging issue, but implies (multiple) replication of large portions of the document

[15] Results obtained in the CLEF 2009 'Ad-hoc track'consisting of experiments on retrieval on the library catalogue records ('TEL collection').

collection and is a computationally costly option (see Section 3.4.3). Finally, the handling of many different languages remains a challenging problem. As the number of languages to be handled grows, it quickly becomes unfeasible to acquire and incorporate direct translation resources for every potential language combination. The use of an intermediary language (pivot language) can be a solution, but carries the risk of amplifying the rate of translation error, due to the use of 'double translation'. Perhaps surprisingly, for some language combinations indirect translation may be beneficial, specifically if no high-quality direct translation resources are available (see Section 3.4.4).

Looking to the future, difficulties remain in locating translation resources of sufficient coverage and quality for many less-frequently spoken languages. Systems that can incorporate many different types of translation resources and deal with uneven quality will therefore be important if more and more languages are to be handled. With the increasing number of resources and languages, questions of merging intermediate results, both calculated on the same collection (data fusion) or disjoint collections (collection fusion), become increasingly pressing. Here, state-of-the-art approaches are still significantly less effective than theoretically optimal benchmarks.

## 3.6 Suggested Reading

The first book focused entirely on MLIR/CLIR issues consisted of a collection of papers, most of which were revised versions of presentations at the Workshop on Cross-Linguistic Information Retrieval held during the SIGIR'96 Conference (Grefenstette 1998). It is interesting to look back at this volume today as it provides a good coverage of some of the earliest work in this area, which included experiments with machine translation, bilingual dictionaries, multilingual thesauri, and latent semantic indexing. A fairly comprehensive coverage of (mainly academic) R&D describing MLIR/CLIR experiments on European languages can be found in the CLEF Workshop Proceedings, published each year since 2000.[16] Other important sources of research literature in this area for languages outside of Europe are the TREC,[17] NTCIR[18] and FIRE[19] publications available online.

A good overview of many of the aspects covered in this chapter can be found in the book by Nie (2010) entitled *Cross-Language Information Retrieval*, which also provides a broader insight in the workings of machine translation systems.

[16] From 2000 to 2009 CLEF Proceedings were published in two forms: as online reports of experiments, see http://www.clef-campaign.org/, and in a more extended form including additional analyses of results in the Springer LNCS series, see http://www.springer.com/computer/lncs/.

[17] See http://trec.nist.gov/.

[18] See http://research.nii.ac.jp/ntcir/.

[19] See http://www.isical.ac.in/~clia/index.html.

# References

Ballesteros LA, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR 1997), ACM Press: 84–91

Belkin NJ, Kantor P, Cool C, Quatrain R (1994) Combining evidence for information retrieval. In: Proc. of the 2nd Text REtrieval Conference (TREC-2). NIST Special Publication 500-215: 35–43

Braschler M (2004a) Combination approaches for multilingual text retrieval. J. Inf. Retr. 7(1/2): 183-204

Braschler M (2004b) Robust multilingual information retrieval. Doctoral Thesis, Institut interfacultaire d'informatique, Université de Neuchâtel

Braschler M, Gonzalo J (2009) Best practices in system and user-oriented multilingual information access. TrebleCLEF Project: http://www.trebleclef.eu/

Braschler M, Schäuble P (1998) Multilingual information retrieval based on document alignment techniques. In: Proc. 2nd European Conference on Research and Advanced Technology for Digital Libraries, Second European Conference (ECDL 1998), Springer-Verlag: 183–197

Buckley C, Mitra M, Walz J, Cardie C (1998) Using clustering and super concepts within SMART: TREC-6. In: Proc. of the 6th Text REtrieval Conference (TREC-6), NIST Special Publication 500-240: 107–124

Darwish K, Oard DW (2003) Probabilistic structured query methods. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR 2003). ACM Press: 338–344

Ferro N, Peters C (2010) CLEF 2009 ad hoc track overview: TEL and Persian tasks. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments, Springer LNCS 6241: 13–35

Fox EA, Shaw JA (1993) Combination of multiple searches. In: Proc. of the 2nd Text REtrieval Conference (TREC-2), NIST Special Publication 500-215: 243–252

Gey FC (2005) How similar are Chinese and Japanese for cross-language information retrieval? In: Proc. of NTCIR-5 Workshop Meeting, December 6–9, 2005, Tokyo, Japan

Gollins T, Sanderson M (2001) Improving cross language retrieval with triangulated translation. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR 2001). ACM Press: 90–95

Grefenstette G (1998) (ed.): Cross-language information retrieval. Kluwer Academic Publishers

Hedlund T, Airio E, Kekustalo H, Lehtokangas R, Pirkola A, Järvelin K (2004) Dictionary-based cross-language information retrieval: Learning experience from CLEF. J. Inf. Retr. 7: 97–117

Hiemstra D, de Jong F (1999) Disambiguation strategies for cross-language information retrieval. In: Proc. of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL 1999). Springer-Verlag: 274–293

Littman ML, Dumais ST, Landauer TK (1998) Automatic cross-language information retrieval using latent semantic indexing. Chapter 5. In: Grefenstette (ed.) Cross-Language Information Retrieval. Kluwer Academic Publishers: 51–62

Mandl T, Womser-Hacker C, Di Nunzio GM, Ferro N (2008) How robust are multilingual information retrieval systems? In: Proc. ACM Symposium on Applied Computing (SAC 2008): 1132–1136

McNamee P, Mayfield J (2002) Comparing cross-language query expansion techniques by degrading translation resources. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR 2002). ACM Press: 159–166

McNamee P, Mayfield J (2004) Character n-gram tokenization for European language text retrieval. J. Inf. Retr. 7: 73–97

Moreau N (2009) Best Practices in language resources for multilingual information access. TrebleCLEF project: http://www.trebleclef.eu/

Nie J-Y (2010) Cross-language information retrieval. Morgan & Claypool

Nie J-Y, Simard M, Isabelle P, Durand R (1999) Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR '99). ACM Press: 74–81

Pirkola A (1998) The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proc. ACM SIGIR conference on research and development in information retrieval (SIGIR 1998). ACM Press: 55–63

Qiu Y, Frei H-P (1993) Concept based query expansion. In: Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993). ACM Press: 160–169

Savoy J (2004) Combining multiple strategies for effective monolingual and cross-language retrieval. J. Inf. Retr. 7(1/2): 119–146

Savoy J, Dolamic L (2009) How effective is Google's translation service in search? Commun. of the ACM 52(10): 139–143

Sheridan P, Braschler M, Schäuble P (1997) Cross-language information retrieval in a multilingual legal domain. In: 1st European Conference on Research and Advanced Technology for Digital Libraries (ECDL 1997), LNCS 1324: 253–268

Si L, Callan J (2005) Multilingual retrieval by combining multiple ranked lists. In: Peters C (ed.) Proc. of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005). Available at http://www.clef-campaign.org

Yu K, Tsujii J (2009) Bilingual dictionary extraction from Wikipedia. In Proc. of Machine Translation Summit XII (MTS-2009). Available at http://www.mt-archive.info/MTS-2009-Yu.pdf

# Chapter 4
# Interaction and User Interfaces

*"Multilingual interfaces as well as multimedia interfaces will be particularly important as the information society continues to become more global."*

Marchionini 1992

**Abstract** Increasingly people are required to interact or communicate with Information Retrieval (IR) applications in order to find useful information. This interaction commonly takes place through the user interface, which should help users to formulate their queries, refine their searches and understand and examine search results. In multilingual information access, the design of an effective search interface and the provision of functionality to assist the user with searching are vital as users cross the language boundary and interact with materials written in foreign, and potentially unknown, languages. Designing the user interface may also involve the localisation of existing material and services, in addition to providing cross-language search functionalities. This chapter focuses on the users of MLIR/CLIR systems and highlights the issues involved in developing user interfaces, including what form of multilingual search assistance can help users and how to design an interface to support multilingual information access.

## 4.1 Information Seeking and User Interaction

Real life information retrieval, whether monolingual, multilingual or cross-lingual, is interactive and dynamic and involves iterative cycles of interaction between users, content, information systems and the search context/environment. The information needs and tasks performed by the users of IR systems are often diverse and evolve over time, resulting in varied and complex information seeking behaviours.

The traditional view of information retrieval has typically been 'system-oriented' and focused on optimising the match between representations of the user's query and the documents being searched. However, more recently the focus has shifted to consider more 'user-oriented' aspects of search: human–computer interaction, the user's information-seeking behaviour, the contexts in which users search for

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_4, 

information and their cognitive abilities. User-oriented approaches to IR have also been motivated by the recognition of information not just as a 'thing' (i.e., a physical object), but also as something that, through interpretation and reasoning, can lead to changes in a person's state of knowledge.[1] Considering the user and their interactions with an information retrieval system is the focus of *Interactive Information Retrieval*. A user-oriented view has also been adopted in cross-language research within *interactive CLIR*, which is viewed as "*a process in which searcher and system collaborate to find documents that satisfy an information need regardless of the language in which those documents are written*" (Oard et al. 2008, p. 181).

Users will interact with an IR application through a user interface. This provides a communication channel that enables a dialogue between the user and the IR application. Typically, this dialogue will consist of a series of inputs and outputs where the user inputs a query; the application matches documents to the query and returns a list of search results; the user selects one of the search results; the application displays a full version of the corresponding document, and so on. The design of the user interface influences this communication through the provision of searching aids and by presenting the application in a way that helps users carry out their searching tasks. However, interactive IR is more than simply designing a good user interface; understanding the user's individual characteristics and their searching context is just as important. Interaction will occur at various levels, ranging from physically typing commands and selecting interface objects, through interpreting and reasoning at a cognitive level as people digest and use information to change their mental state.

The use of a particular IR application occurs within the wider context of *information seeking*[2]: resolving problems and accomplishing (work or leisure) tasks, going from conceptualising the information need to using the relevant information that has been discovered. Marchionini (1995) defines information seeking as "*a special case of problem solving. It includes recognizing and interpreting the information problem, establishing a plan of search, conducting the search, evaluating the results, and if necessary, iterating through the process again*." As well as problem solving, information seeking has also been considered within a broader framework of decision-making and cognitive models of learning.

Figure 4.1 provides a simplified illustration of the information seeking process in which recognising the need for information (discussed in Section 4.2) will give rise to information seeking activities to find documents that will help the user fulfil their information needs. This is similar to Figure 2.1, the information acquisition cycle presented in Chapter 2. However, it can be seen as more generic than Figure 2.1 as it captures not only the retrieval (or search) process, but also browsing activities which are often provided by IR applications and provide an alternative way of finding information.

[1] This is exemplified, for example, by Ingwersen's cognitive perspective on IR (Ingwersen 1996).

[2] Information seeking is also itself a part of the wider field of human information behaviour: "*the totality of human behaviour in relation to sources and channels of information, including both active and passive information seeking, and information use*" (Wilson 2000, p. 49).

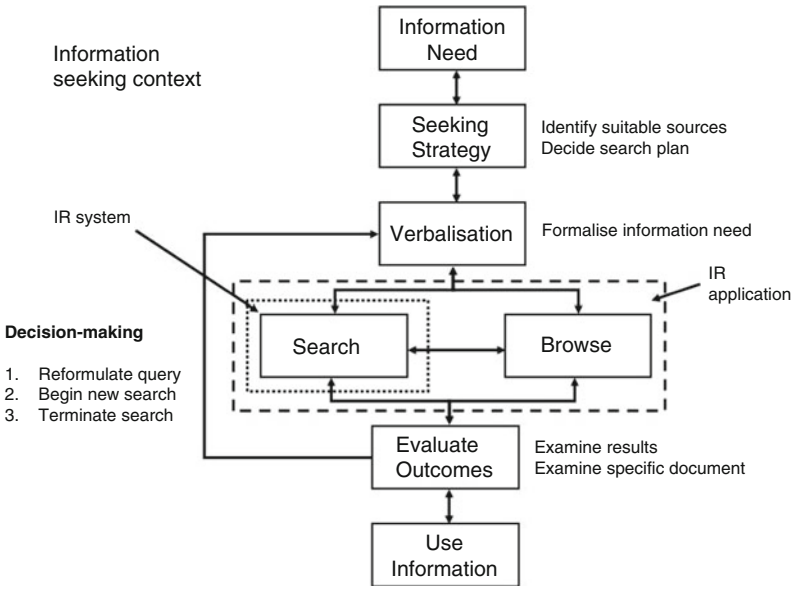


**Fig. 4.1** Simplified representation of the information seeking process (compare to Chapter 2)

Commonly, people will develop strategies to find the information they need and devise an approach/plan to locate relevant information (the 'seeking strategies' of Figure 4.1). This might include identifying particular search tools to use, particular sources of information to consult, and particular information seeking approaches to adopt, e.g., issuing queries known to work successfully in the past or starting with a generic search topic and then narrowing down to a more specific and precise definition of the information need. The user's information need must be formalised ('verbalisation' in Figure 4.1) and expressed to match the vocabulary of the IR application (the 'formulation/coding' step in Figure 2.1). Search and/or browse activities are then carried out to find potential relevant documents.[3]

The users' individual characteristics and their social or cultural situations will form a part of the *context* in which the search activities are conducted. Context can refer to personal characteristics of the user (e.g., age, gender, education, job experience, and social/economic status), the role a person plays in work or society (e.g., the job a person does and their position, experience, competencies and level of responsibility), or the environment in which someone searches (e.g., cultural norms, legislation and economic and political situation). Many retrieval tasks conducted by users are affected by factors, such as time, place, history of interaction, physical

[3] Note that for some tasks users may be unable to easily formalise their information needs as they may not start out with a specific goal or purpose (e.g., in the case of information exploration or serendipitous information seeking behaviours). Assisting users with articulating their information needs is one of the goals of an IR application.

environment and the task in hand and the search context will influence a person's information needs and their information seeking behaviours. For example, in academia the need for content in multiple languages and functionality to support cross-language searching is more prominent in fields such as literature, history and architecture than in mathematics and pure sciences (Clough and Eleta 2010). Context is also found in web search, for example, where results are served differently depending on the user's location (e.g., UK vs. US) and access device (e.g., desktop vs. mobile).

Common approaches that people use to locate information include *searching* for specific items or information about a topic in a structured way, *browsing* for information in a more random and unstructured manner or a combination of using both approaches. Searching with an IR application typically involves formulating queries (the 'formulation/coding' step in Figure 2.1): recalling potential words or phrases that reflect, in some way, an underlying information need. Queries can be expressed in natural language or involve specific operators, such as Boolean expressions. Searching can be highly effective, especially if the user knows what they are looking for or the query is specific, e.g., a known-item search.

Browsing, on the other hand, allows users to discover information in a more random and unstructured way and is often used when people do not have specific search goals or want to navigate through collections of information in a more informal and exploratory way (Marchionini 2006). There are many types of browsing including *systematic*, *exploratory* and *casual* that reflect what a user knows about their underlying information need. Systematic browsing is carried out when the user clearly knows what they are looking for (i.e., they have a specific 'query' in mind), such as locating a specific term in an online glossary or catalogue, or finding a known web page. Exploratory browsing, however, reflects the situation in which the user has less clear objectives about their information need. In casual browsing the user has no particular information need, e.g., scanning a daily news website hoping to find something interesting to read.

The search results, and possibly individual documents, are examined and compared to the original information need as the user evaluates the outcomes of their information seeking activities. This evaluation involves judging the relevance of retrieved items using criteria such as relevance of topic, authority, accessibility, quality, language and timeliness.[4] Depending on whether they have found the information they require, users will reformulate their query and begin a new search, or will stop. This, often iterative, interactive cycle of inputs and outputs between user and system is often termed a search *session*. Finally, relevant information is then used within the context of the underlying work or leisure task. This involves

[4] Often relevance is considered as whether a document is on the same topic as the user's query, however there are many other, more subjective elements to relevance that determine whether a user considers a result to be relevant or not (see, e.g., Barry and Schamber 1998, Steinerová 2008).

both physical and mental acts to incorporate the retrieved information into a person's existing body of knowledge.

This basic information seeking process will be similar for both monolingual and cross-language search tasks; differences may arise due to personal, or individual, differences such as linguistic ability. For example, users fluent in more than one language often switch between languages when carrying out search tasks (Aula and Kellar 2009); users may take longer to judge the relevance of results returned in a foreign language than in their native language (Hansen and Karlgren 2005); users may require different levels of support, e.g., for query formulation, depending on their reading and writing language skills (see Section 4.3). In addition, if we consider the wider information seeking context (as shown in Figure 4.1) then to make use of the documents returned by the MLIR/CLIR system, users may also have to translate them (either manually or using MT). This would not be required when searching for documents written in languages known to the user.

Building IR applications that effectively fulfil users' information needs is often problematic for a variety of reasons. For example, people can find it hard to articulate their needs and translate them into a representation appropriate for a specific searching system, might have difficulties with finding appropriate query terms, may be overwhelmed with too many search results, may not retrieve enough results or any results at all (zero hits), might have difficulty interpreting disorganised lists of results or with using specialised query syntax. This can be particularly challenging for users of cross-language systems when attempting to query collections in languages unfamiliar to them.

In addition, users' information needs can, and frequently do, evolve during the search process and relevance assessments change during the session. The result is that searching behaviour is often dynamic and iterative, commonly involving multiple searches and multiple systems. Understanding the different tasks that users perform and their search goals enables appropriate support mechanisms to be included in the search interface. The design and implementation of various search aids, together with good interface design, can help make search easier for the end users and support them during their information seeking activities, whether searching or browsing (see Section 4.4). IR applications that support other forms of searching activity, such as exploratory search, are also receiving current attention (Marchionini 2006, Wilson et al. 2010).

## 4.2 Users' Information Needs and Search Tasks

When a person uses an information retrieval application, e.g., a web search engine, we assume they come with an *information need*,[5] or some problem to solve, and by using the IR application they find information that helps fulfil that need. Users will

[5] The notion of an information need highlights that a gap in knowledge exists between what someone knows and what they need to know to fulfil some kind of underlying goal, problem or

come with different information needs, such as requiring a specific piece of information, getting an answer to a question, seeking advice or exploring a general topic. In the context of cross-language searching, users do not have a cross-language information need; rather they have a need that cannot be fully satisfied without finding information in languages other than their native one.

People's information needs are commonly driven by the goals they have, i.e., their purpose for information seeking, and the daily work (or leisure) tasks they perform (Vakkari 2003). For example, search tasks which may involve cross-language support include: searching a monolingual collection in a language that the user cannot read; retrieving information from a multilingual collection using a query in a single language and selecting images from a collection indexed with free text captions in an unfamiliar language.

Some of the tasks that people perform may be information searching tasks and could include the retrieval of a specific item (known-item searching), finding documents that contain material relevant to a theme/topic (subject searching[6]), or more generic browsing/exploration. These activities are similar for monolingual, multilingual or cross-language searching. In the context of web search a user's search tasks are often categorised into *navigational* – to find specific information, e.g., a website; *informational* – to find information about a specific topic; and *transactional* – to find a service to initiate further interaction (Broder 2002).

When developing IR applications, particularly when following a user-centred approach (see Section 4.5.1), users' information needs and tasks are often captured during a *requirements*[7] *elicitation* phase. Methods commonly used to establish the information needs people have and the information seeking tasks they perform include conducting questionnaires, interviews and surveys, observing users as they go about their daily activities, analysing the tasks that people carry out on a daily basis, or using implicit sources of evidence, such as transaction logs. Current practice suggests that the results from using different data collection methods should be combined to create a more complete set of requirements, an approach known as *triangulation* (Ingwersen and Järvelin 2005, p. 93).

To help illustrate the use of different approaches for establishing users' information needs we describe two user studies. In the first, Marlow et al. (2007) conducted an online user survey with realistic end-users of Tate Online, a website

work task. Taylor (1968) describes how information needs arise using a series of four stages: from the unexpressed (visceral) need – a conscious (or even unconscious) need for information, through to a conscious need – an ambiguous and rambling statement, to a more qualified and rational statement (formalised need), to a compromised need – the question expressed in a language understandable by the search system, e.g., as queries.

[6] A helpful analogy to highlight the diversity of subject search tasks is that of finding a needle in a haystack (Koll 2000). In this analogy the document(s) or information to be found are needles and the IR system or database(s) used are the haystacks. The point is that subject searching is like finding a needle in a haystack, but not all searches are the same.

[7] A requirement is "*a statement about an intended product that specifies what it should do or how it should perform*" (Preece et al. 2002, p. 204).

for Britain's national art gallery, to ascertain the user needs and requirements regarding implementing multilingual information access. The results from the survey were used to derive a set of requirements and recommendations for providing enhanced multilingual access to content on Tate Online. Areas of particular interest for investigation were:

- User characteristics – where are international visitors located and what languages do they prefer to use when surfing the Internet?
- Task analysis – why do these visitors currently use Tate Online, and what do they do there?
- Requirements –what type of increased multilingual functionalities do the site users need or want?

The second example is the MultiMatch[8] (Multilingual/Multimedia Access to Cultural Heritage) project. User studies involving interviews, questionnaires and analysis of transaction logs were conducted to ascertain the information needs of professional users in the cultural heritage domain (Minelli et al. 2007, Marlow et al. 2008b). As part of the MultiMatch project 15 expert cultural heritage users were observed at work and questioned about their typical multimedia and multilingual search tasks. Characteristics about individuals' characteristics (e.g., language skills and abilities) and their searching behaviours were gathered from representatives of different user groups (education, tourism and cultural heritage). Table 4.1 summarises the types of searches elicited from users which tended to fall into one or more categories relating to proper names, places, time, titles, and general subjects. Example tasks carried out by users included finding material to illustrate Power Point presentations for lectures (academics) and fact-checking details in the archive

**Table 4.1** Example searches by users in the cultural heritage domain

| Category | Example searches |
|---|---|
| Names | Edgard Varèse, Leonardo da Vinci, Terragni |
| Places | Novgorod, Nijmegen, Etruscan tombs |
| Dates (usually combined with another category) | Fourteenth century, 1945 |
| Titles | 'The Marriage at Cana', 'The Man Midwife', 'Nuit et brouillard' |
| General subjects | Gambling, people judging things, coffee, illustrated punch bowls |
| Category Combinations | |
| General | Youths from the 1960s smoking cigarettes, street shots from Paris, Moscow landscapes |
| Specific | Henri Matisse's murals and stained glass of the Rosary Chapel at Vence 1949–1951, Architecture of Dubai and Beijing, Land mine clearing in Holland after WWII (1945–1948), fourteenth century Humberware drinking jugs |

[8] http://www.multimatch.org/

databases (practitioners). Eleven of the 15 interviewees were bilingual or polyglots, and needed to use other languages at least occasionally in their work. Examples of multilingual searches included an Italian speaker looking for images of individuals in a German catalogue and a Dutch speaker looking for information from a Slovakian web page.

## 4.3 Users' Language Skills and Cultural Differences

The users of search systems will exhibit *individual differences* and these may well impact or influence their search behaviour, the design and personalisation of the interface and the requirements for specific search assistance that the IR system should provide. These differences may include the users' personal characteristics, such as age, gender, physical abilities, cognition, education, cultural background, training, motivation, goals, personality and language skills/abilities. These individual attributes, together with other contextual factors, such as the users' level of domain expertise (e.g., novice or expert), familiarity with a task or subject area and knowledge about the system interfaces being used to find information (collectively known as their *situation*[9]) may well affect their searching behaviour, shape their expectations from the search system and affect their success at finding relevant information. For example, users from a cultural heritage background searching for 'Madonna' will be expecting different results from users interested in pop music. In the case of cross-language searching two contextual factors of particular interest are the user's language skills and cultural differences.

The user of an MLIR/CLIR application will formulate queries in a source language that must be matched against documents written in one or many target languages. Individuals can have a range of both *passive* (listening and reading) and *active* (writing and speaking) abilities based on their mother tongue and other languages studied for any length of time. Users with multi-language skills (*polyglots*) may be able to formulate searches and judge the relevance of results in multiple languages, with varying degrees of success, but want the convenience of a single query, rather than issuing searches for each target language. These users may also require assistance with searching information in their non-native languages. On the other hand, some users have language skills in only one language (*monoglots*) and require assistance with the entire information seeking process. Users with poor foreign language skills may be able to read documents returned by a cross-language search system, perhaps with the help of a machine translation system, but find it difficult to formulate search queries effectively.[10] In fact many

[9] The notion of 'situation' describes a specific search scenario: a particular person carrying out a specific task in a given context at a particular time.

[10] Ogden and Davis (2000) summarise two types of users for CLIR: bilingual users who formulate their queries in their native language to retrieve documents in their second languages and monolingual users who are interested in finding information in other languages.

users, particularly on the Web, may have language skills in their mother tongue *and* English. In many cases users will turn to English if information on the Web in their native language is not available (Rieh and Rieh 2005).

Language ability is an important variable to consider when designing effective cross-language search tools as it can affect the user's search experience, their ability to interact with the searching system, and the search assistance they may require to find relevant items. The need for search assistance is substantially higher than in monolingual information retrieval: normally, the user can quickly adapt to characteristics of the system, but not to an unknown target language (Oard and Gonzalo 2002). To investigate the relationship between language skills and cross-language search functionalities used and appreciated by web users, Marlow et al. (2008a) conducted a task-based evaluation of web searching using Google Translate[11]: 12 participants were asked to search for web pages relating to 12 pre-defined topics with 4 topics in their native language, 4 written in a language for which they had a passive knowledge and 4 in an unknown language. Table 4.2 shows the frequency with which certain functionalities of Google Translate were used. The reliance upon query translation functionalities increased with language unfamiliarity: users were more likely to look at the translated versions of pages for unknown languages, and the original versions for passive languages. Query editing occurred only three times out of all 144 topics, and these were exclusively in the passive language condition. Based on the tools available (which offered limited editing assistance for translated queries), users were much more likely to reformulate or edit the query in the source language than to deal with the machine translation version.

The user's cultural background, which includes language, may also impact how users locate information and their satisfaction with search results. Culture is part of a user's environmental context and can affect various aspects of information seeking behaviour (Zoe and DiMartino 2000, Kralisch and Berendt 2004). For example, users from the UK searching for 'pants' or 'football' are likely to expect and want different results from a search engine than users from the US. Culture may also affect users' perceptions and choice of target language. For instance, Rieh and Rieh (2005) found that Korean students commonly selected material in English because of a perception that material of higher quality was only available in English.

**Table 4.2** Frequency of use of Google Translate functionalities for each topic, by language (Marlow et al. 2008a)

| | Query translation | Translated query editing | Original links viewed | Translated links viewed | Both links viewed |
|---|---|---|---|---|---|
| Native | 13 (27.1%) | 0 | 4 (8.3%) | 1 (2.1%) | 1 (2.1%) |
| Passive | 37 (77.1%) | 3 (6.3%) | 26 (54.2%) | 2 (4.2%) | 4 (4.2%) |
| Unknown | 46 (96.0%) | 0 | 14 (29.2%) | 19 (39.6%) | 9 (18.8%) |

[11] http://translate.google.com/translate_s

Cultural differences will also affect the design of user interfaces and are often considered during the localisation of information products and services, for example adapting websites to meet the linguistic and cultural needs of the local communities they target. The different versions are known as localised websites, or services, and often require specific design considerations, such as identifying which languages a website should be translated into, an awareness of cultural issues (e.g., the use of specific terminology or offensive references), the availability of resources (e.g., manpower, translation tools), technical and maintenance issues, how to measure success and issues surrounding design. Further consideration on localisation is given in Section 4.6.

## 4.4 Supporting Multilingual User Interaction

Users will come to an IR application with some kind of information need, which may be well-defined or vague. After selecting a system and collections to work on the user will then verbalise and express their information need as a query to an IR application, a process known as 'query formulation'. The query will be sent to the IR application, a black box to the user, and its response will be a set of results, which can be arranged or presented in various ways, such as a ranked list, a cluster of documents or some kind of visualisation. The user will scan, evaluate and interpret the results (i.e., evaluate the outcomes) and possibly view selected files when identifying relevant items. Many search sessions consist of successive searches rather than a single iteration and the user may decide to refine or reformulate the query and iterate the search cycle again, begin a new search or finish.

In this cycle of activities there are a number of points where users can interact with the search process: during query formulation (see Section 4.4.1), when exploring the search results and examining individual documents (see Section 4.4.2), and during query reformulation and refinement (see Section 4.4.3). However, users may also carry out browsing-based activities during their search for relevant documents and many IR applications will also provide functionality to support browsing, for example of the results set, and this may cause users to deviate from the standard retrieval process and intended navigational path (see Section 4.4.4). This combination of search and browse functionality enables support for a wide range of information seeking behaviours and goes beyond the traditional paradigms of keyword search to support more exploratory information seeking behaviours. This is becoming increasingly possible through the enrichment of content with metadata and explicit semantics and using technologies developed as part of the Semantic Web[12] (see, e.g., Wilson et al. 2010). In examples such as digital library systems, web portals and e-commerce applications, it is common to find search functionality embedded

[12] Tim Berners-Lee, who coined the term 'Semantic Web' in 2001, defines it as "*a web of data that can be processed directly and indirectly by machines*" (Berners-Lee et al. 2001).

within a larger application aimed at supporting a wider set of activities beyond keyword search, for example during decision-making.

Providing effective access to multilingual document collections undoubtedly involves further challenges for the designers of interactive retrieval systems. Users will have varying language skills and abilities and require additional assistance to overcome the language gap between the documents and their queries (Section 4.4.1). In many situations users can adapt to the particular characteristics of an IR application, however the need for search assistance is substantially higher in cross-language search where users may find it difficult, or even impossible, to adapt to unknown target languages.

In interactive CLIR users can be given additional support when formulating queries by assisting them with query translation: translating their query from a source to target language(s). Users can also be provided with support for viewing the results list to help them recognise potentially relevant items and with translating documents selected for use and extracting information to fulfil their overall information needs (Section 4.4.2). Table 4.3 provides a summary of functionalities to assist users that are applicable, both in monolingual and cross-language search scenarios (the cross-language functionality is highlighted in bold). Further examples of aids for supporting users' searching more generally can be found in (Hearst 2009). In addition to providing searching aids a MLIR/CLIR system must also provide support for a multilingual user interface by correctly displaying multiple fonts and characters and allowing users to select the interface language (Section 4.5.2).

### *4.4.1 Query Formulation and Translation*

In the first stage of user interaction, query formulation, the system must allow users to specify their queries, which can be expressed in various ways, such as keyword queries, natural language queries (including queries as questions), and paragraphs of text, queries containing Boolean operators and queries with command-based syntax.[13] Most IR applications will provide a search box where users can enter a textual representation of their information needs, for example a list of keywords. Queries specified by users searching the Web are typically short, varying from one to four words in length. For multilingual searching, query lengths tend to vary according to language and the scripts used, but otherwise follow general patterns of queries in English (Lazarinis et al. 2009). When querying visual media, e.g., images, the IR application may also allow users to provide a visual exemplar (Query-By-Visual-Example or QBVE), sketch shapes within the desired image or specify attributes such as colour (e.g., by picking colours from a palette).

[13] Although less common in general web searching command-line syntax is still widely used in interfaces to professional indexes and databases, such as those in the medical, scientific or legal domains.

**Table 4.3** Example support functionality offered by IR systems adapted from (Sutcliffe and Ennis 1998). Cross-language functionality is highlighted in bold

| Support type | Functionalities |
|---|---|
| Query formulation support (Section 4.4.1) | Specialised query syntax (e.g., Boolean operators) |
| | Natural language query input |
| | Advanced search editors |
| | Dynamic term suggestion/Auto-completion |
| | Query-by-example |
| | Query-by-pointing (e.g., on concept maps or images) |
| | Pre-formed queries (e.g., from existing thesauri or manually selected queries) |
| | Re-useable queries (e.g., from past searches) |
| | **Query translation (e.g., language selection, select/deselect translated terms, back translation of query terms)** |
| Evaluation support – document selection and examination (Section 4.4.2) | Provide summary of results, e.g., generate document surrogates, display number of hits |
| | View content of selected document |
| | View and mark/re-use results (relevance feedback) |
| | **Provide summary of results (e.g., present results written in different languages, generate and translate document surrogates)** |
| | **Translate selected document** |
| Query reformulation support (Section 4.4.3) | Automated term suggestion (e.g., related queries) |
| | Spelling error correction |
| | 'More like this' functions |
| | Search within results |
| | **Edit query translation (e.g., query expansion and translation refinement)** |
| Browsing support – collection and results (Section 4.4.4) | Controlled vocabularies and classification schemes (e.g., Dewey codes or faceted classification) |
| | Concept maps |
| | Navigational aids (e.g., hypertext) |
| | 2-D/3-D visualisations |
| | **Multilingual controlled vocabularies and classification schemes** |

The use of advanced search functionalities can help users be more specific about their search and enable results to be filtered according to certain criteria, e.g., specify that exact words or phrases must appear in the search results. More proficient users may utilise specific query syntax, such as Boolean operators, to create more complex queries. Many IR applications also implement advanced search functionalities to constrain the results, e.g., restricting search results to

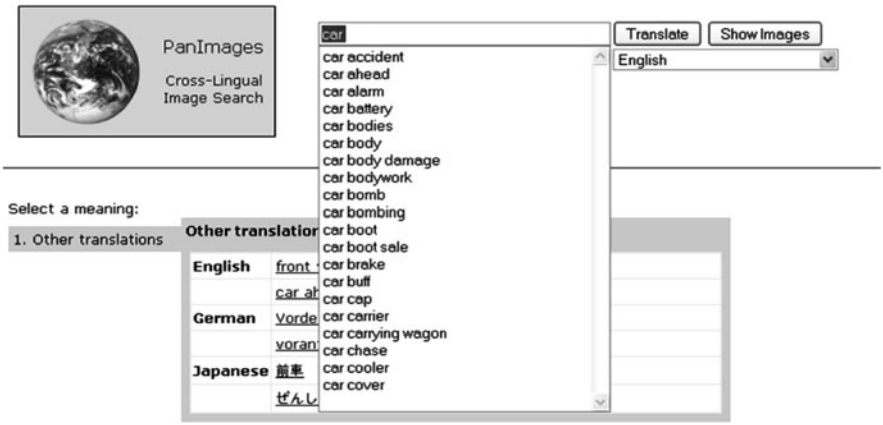


**Fig. 4.2** An example of auto-complete for a cross-language image search engine (Colowick 2008)

contain only documents written in specific languages, to contain only certain file types (audio files or web pages) or limited to a specific number of results.

Techniques such as *auto-completion* (or dynamic term suggestion) can also be used to assist users express their information needs. Auto-completion is often implemented as a drop-down box that updates dynamically as users type their query and can be selected at any time to initiate the search. Figure 4.2 shows an example of auto-completion from PanImages, a cross-language image search engine (Colowick 2008). The user submits a query in their preferred language (e.g., 'car' in English) and as the user types, a drop-down box shows further queries which the user can select. The entries in the drop-down box come from entries which can be found in the PanImages dictionary.

In the case of interactive cross-language IR, one of the predominant forms of search assistance at this stage is during *query translation*: mapping the query from one language into a different language(s). The amount of user involvement, or interactivity, in this process will depend on the user's language skills/abilities, the translation resources available for query translation and the design of the user interface and interaction model. Chapter 3 describes approaches for query translation including MT and those based on dictionary look-up and term statistics. From a user perspective, the use of MT leaves little room for user interaction in the translation process, especially in correcting or refining the translated query. The use of a bilingual dictionary, on the other hand, has the advantage that users can interact with the translation process, for example in the selection of alternative word senses. Previous research has shown that allowing the user to monitor and interact with query translation increases the performance of CLIR systems (He et al. 2003, Petrelli et al. 2002).

A further consideration for query translation is whether 'fully automatic' or 'user-assisted' query translation is used (Petrelli et al. 2004, Oard et al. 2008). In fully automated query translation the common mode of operation is for the user to enter the query, the system automatically translates the query with no user involvement and returns a list of search results. This is commonly implemented with MT

and the user can only modify (or refine) the query in the source language. Although this provides little user interaction, it may work well for users who have limited language ability in the target language(s). The alternative approach is to allow users to edit the query translations manually (e.g., selecting and deselecting translations) either prior to the search as an additional step in the retrieval process, or post-search. The advantage of the latter approach is that it provides immediate user feedback. If the results are not as expected the user can go back and post-edit the translated query terms and repeat the search. Table 4.4 summarises a number of aspects to consider in developing functionality to assist query formulation and translation.

To demonstrate the variety in functionality that can be offered to users we discuss a number of past and current MLIR/CLIR systems, both academic and commercial. Figure 4.3 shows an example of query formulation and translation implemented in the MULINEX system (Capstick et al. 2000). The user enters a query ('euro introduction'), selects the source and (possibly multiple) target languages (English, French and/or German) and then clicks 'search' (Step 1 in Figure 4.3). The system can automatically create a translation of the query, thereby hiding the query translation step. However, if the user wants to edit the query, a query assistant button guides the user to a page which allows the user to select/deselect up to three translations in the target language(s) – Step 2 in Figure 4.3. However, no back translation is provided for the different term translations. This may make it difficult for users with no knowledge of French or German to formulate queries. MULINEX is multi-language (German, English, and French) and a separate column of translations is provided for each language. It also suggests a list of additional terms the user might decide to include in the query.

Figure 4.4 shows the user interface developed for a cross-language multimedia information retrieval system in the Clarity project (Petrelli et al. 2004). The end users of Clarity were journalists working for BBC Monitoring (UK) and Alma Media (Finland). These users were polyglots. Clarity is an interesting cross-language system as the design of the interface followed a user-centred approach and involved real users carrying out realistic search tasks. Similar to MULINEX, users specify their source language and their target language(s). Because the system supports Baltic languages, such as Latvian and Lithuanian, the interface includes support for inputting characters with appropriate diacritics (e.g., ä, å and õ). The users enter their query in the source language and the system automatically selects term translations in the target languages and performs a search. Clarity has two interfaces: one which allows the users to modify the translation (the 'supervised mode') and the other which does not (the 'delegated mode'). Using the delegated mode, the user simply enters the query, clicks the 'search' button and the results are displayed. There is no user intervention during the query translation process. To modify the query, the user must re-enter it in the search box. In the supervised mode, the user has the option to edit the translated query terms and update the search. The users are also given *back translations* (shown in parentheses in Figure 4.4) of the translated query terms to help with selecting/deselecting translation terms.

**Table 4.4** Aspects to consider when developing functionality to support query formulation and translation

| Aspect to consider | Description |
|---|---|
| Automated or manual source language selection | Automatically detect the user's query (source) language or require that users specify their source language. Repeatedly selecting a source language may be a burden for the user, so may want to allow users to save this setting (along with other settings, such as the commonly searched source languages) |
| Making source language equivalent to the interface language | If the interface is localised to a particular language the users' source language could be assumed to be the same as the interface language. However, there may be cases when the user wants to select a source language that differs from the interface language |
| Automated vs. user-assisted query translation | Either automatically translate the query without user involvement (automated) or allow users to edit the query translation (user-assisted) |
| Allowing users to select one or multiple target languages | The simplest approach is to allow only the specification of a single source-target language pair (common for many library catalogue or digital library interfaces). However, this can be restrictive if users speak multiple languages fluently. Supporting multiple target languages may also be desirable when users do not know in which language potential relevant items may be written in. Thought needs to be given as to how users are able to add/remove target languages and how to scale up to greater numbers of languages |
| Showing non-translatable terms to the user | In some cases the users' query terms may be non-translatable and a decision should be taken whether to highlight these to the user and include them in the query |
| Allowing users to indicate non-translatable terms | Query syntax may be provided that will allow users to manually indicate terms that should not be considered for translation. For example, a user might indicate that a proper name such as 'George Bush' should not be translated (Bush would probably be incorrectly translated if performed in a word-by-word manner) |
| Automatically detecting phrases or provide suitable query syntax to indicate phrases | To prevent incorrect translation, phrases, such as 'still life', could be detected automatically or the appropriate query syntax provided to allow users to indicate phrases that should be treated as atomic units for translation |
| Named entity detection | In some cases named entities (e.g., people, places and organisations) could be manually or automatically identified from user's queries to reduce potential translation errors |

(continued)

**Table 4.4** (continued)

| Aspect to consider | Description |
|---|---|
| Providing back translations of translated query terms | In situations where users are required to select suitable query term translations (e.g., senses from a dictionary) but cannot engage with the target language(s) then back translations may be used to indicate the meaning of translated query terms |
| Supporting user-created dictionaries | It is possible that users may be able to offer additional or better translations for query terms than currently stored in the language resources used for translation. Functionality, such as user-generated dictionaries, could be provided for enabling this |
| Determining the number of senses to show users for ambiguous term translations | There may be many possible translations for some terms due to lexical ambiguity and a decision must be made whether to include all of these in the display and on how to order the senses |
| Providing support for inputting characters with diacritics and non-European languages | Users may have access to specialist hardware, such as a language-specific keyboard, but for universal access the interface should provide appropriate functionality for inputting characters with diacritics and queries in non-European language scripts, e.g., a virtual keyboard |

In the PanImages system (shown in Figure 4.2), the user submits a query in their source language (e.g., 'car'), possibly using the auto-complete function, and specifying their source language. The user does not specify a target language; rather the application searches its dictionary and returns a list of possible translations that match the query in various languages. The user can select one of the translations (and senses) and this executes a search for images that match the query in the associated image metadata. Note that this approach requires the user to select a translation before having immediate feedback in the form of search results. The application is only able to translate simple words or phrases as found in its dictionary, not multi-term queries. For example, the query 'still life painting' cannot be translated although the queries 'still life' and 'painting' can. A useful feature of this application is that users can edit the translation and this is stored in a user-created (or personal) dictionary, which is only accessible by registered users. Users can also add translations in new languages which are stored in the PanImages dictionary for further use and can also be shared with other users of the system.

As a final example, consider the interface for the 'Translated Search' feature in Google Translate[14] (shown in Figure 4.5). No assumption is made about the intended

[14] http://translate.google.com/

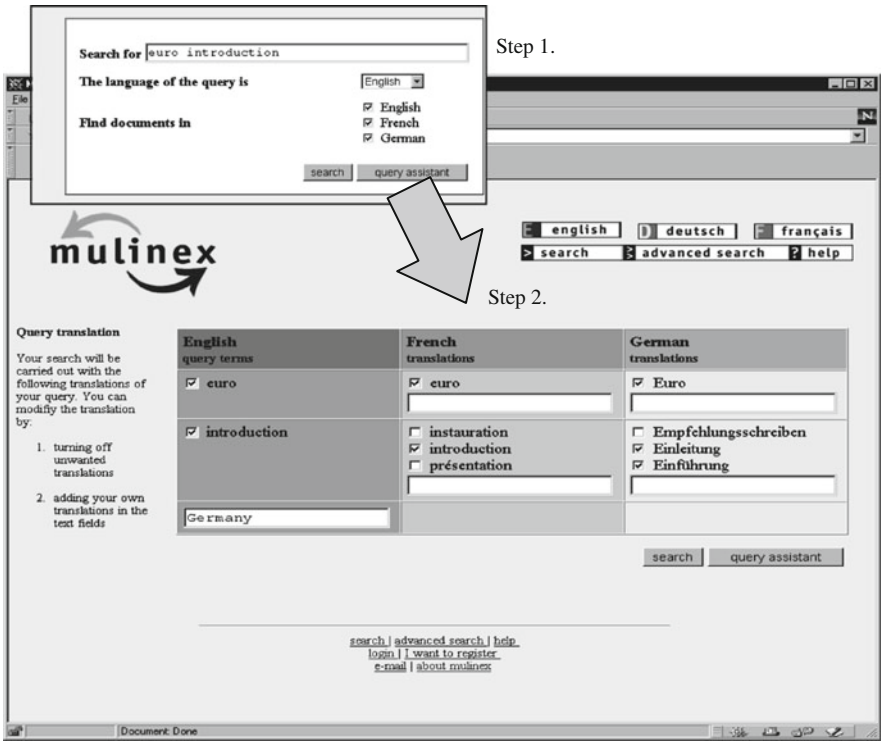


**Fig. 4.3** Example of query translation with the MULINEX system (Capstick et al. 2000)

users: they could exhibit varying language skills and be using Google Translate for a wide range of tasks. Google Translate supports general web searching and provides query translation for multi-term queries for a large number of language pairs. The user submits a query in the source language (e.g., 'still life paintings by van gogh'), selects a single target language and finally clicks the Translate and Search button. The application implements a fully automatic translation approach and presents search results immediately. Notice that there is no 'search all languages' option; just bilingual search between a specific language pair.

Translating the query into Dutch, for example, the interface provides the user with the estimated number of results and displays the translated query in the target language ('stilleven schilderijen van Van Gogh'). If the user is not satisfied with the search results or wants to refine the query they are able to modify the query in the target language, although this requires language-specific knowledge and overall there is limited interaction with the query translation stage.

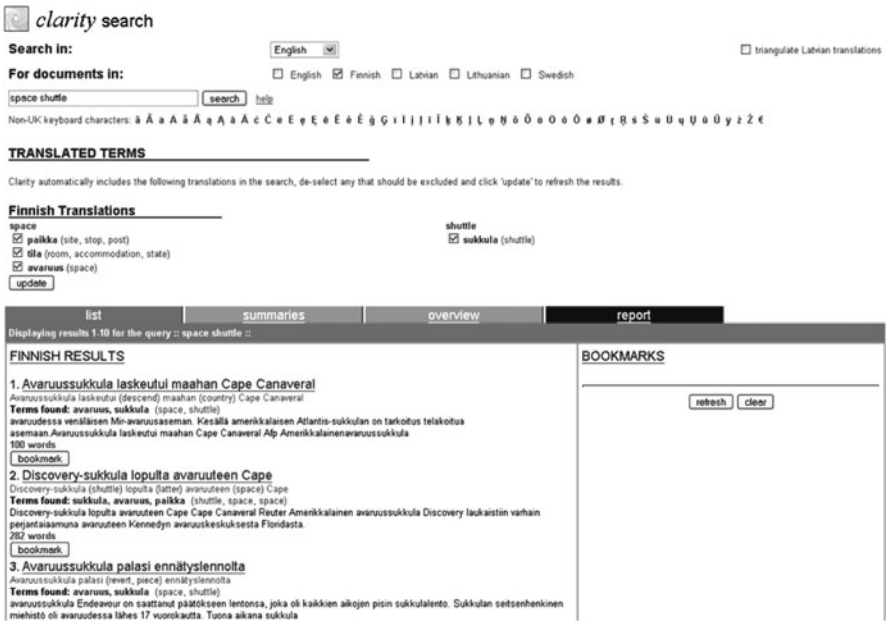


**Fig. 4.4** Example of query formulation and translation in Clarity (Petrelli et al. 2004)


**Fig. 4.5** Example of query translation with the web search function of Google Translate (translate.google.com) as provided at the time of writing. This screenshot is copyright of Google

### *4.4.2 Document Selection and Examination*

After users have issued a query they are presented with search results, which are examined and judged for relevance (i.e., evaluated). Users may require assistance as they scan, evaluate and interpret the results. The IR application can assist users in this stage by creating *document surrogates* that summarise the content of the documents being referenced by the search results. Common types of document surrogate include 'snippets' for plain text documents, bibliographic fields for catalogue records, thumbnails for images, audio segments for audio files and key frames for videos.

Generating useful document surrogates is an important area of research as users will often decide whether referenced documents are relevant or not from the surrogates rather than navigating to the document itself (Lorigo et al. 2006). Past work has demonstrated that query-biased summaries – summaries focused around the user's query terms rather than summarising the document as a whole – are preferable (Tombros and Sanderson 1998). Surrogates for text documents commonly also include highlighted terms in the snippets that match query terms the user has searched on (e.g., the bold text highlighted in Figure 4.6).

For interactive cross-language IR, translation may be necessary for document selection and examination depending on the user's language skills and intended use of the results. Of course, the question may arise: 'why might users want to be presented with documents they, presumably, cannot read?' Firstly, translation tools, such as MT systems, can be used to translate an entire document or result surrogates into the user's native language; secondly, it is possible to select a subset of terms, such as proper names, and translate only these; thirdly, some objects, e.g., such as images, are language-independent and can be used without translation; finally, users may simply want to know that relevant results *may* exist in different languages rather than use the information, e.g., patent lawyers checking for possible infringements retrieve candidate documents (which can then be checked more carefully). Table 4.5 provides a number of aspects to consider when developing search assistance aids to support users in the stage of document selection and examination.

To assist users with document selection, surrogates of the results must be translated into a language the user can understand and use. There are several approaches for doing this. For example, in the case of a ranked list of results the existing surrogates (e.g., summaries and document titles) could be translated individually before being presented to the user. Alternatively, the entire results page, which could contain surrogates, could be run through an MT system. A final approach could be to extract selected terms from documents referenced in the results list (e.g., nouns and noun phrases) and translate these into the language of the users' query. This has the advantage that it is often possible to translate specific phrases more accurately than entire texts.

Oard et al. (2004, pp. 8–9) discuss further studies for document selection and examination and show that word-by-word translations (e.g., noun phrases from the retrieved documents) are sufficient for users to judge the relevance of documents; although people may be less confident in their judgements than judging documents in their native language. For document examination it is common for applications to provide a link to an MT version of the original document. Often this version is generated dynamically when the user requests to view a translation of the document as it may be more efficient than translating all documents in a collection into all possible languages. In some cases users may be willing, and able, to accept imperfect translations of texts (Resnik 1997, Marlow et al. 2007).

The layout of search results can also impact the ability of users in reviewing and using them and aspects of design such as usability must also be considered. Commonly results are presented as ranked lists with around ten hits per page, although this is changing as search systems interleave different kinds of media in

**Table 4.5** Aspects to consider when developing functionality to support document selection and examination

| Aspect to consider | Description |
|---|---|
| Highlighting query terms | Often highlighting the query terms in some manner (e.g., font style or colour highlights) is beneficial for monolingual and cross-language search interfaces |
| Creating document surrogates | This will vary by media type. For textual documents, one could select a subset of words to translate into the user's source language (e.g., proper names or the title of a web page). Surrogates can be created in advance or dynamically and on-the-fly (e.g., to generate query-biased summaries for textual documents). Translation of the surrogates may be done in advance or dynamically and on-the-fly |
| Displaying results of more than one target language | If searching many target languages thought must be given on how to present the results. For example, they could be grouped by language (within a ranked list or on different layers using tabs), or documents of different languages interleaved in a single results list. It may also be necessary to deal with duplicate documents (in different languages) |
| Displaying results for different media types (e.g., web pages, images, etc.) | The presentation of different media types will vary. For example, documents may be presented as ranked lists; images are more commonly presented in a grid format. The former will typically be represented by a textual surrogate (e.g., title, URL); the latter by a thumbnail image |
| Providing different sorting mechanisms and filter options | In some instances it is preferable to provide users with the functionality to sort results, e.g., sorting library catalogue entries by author name or year. Also, allowing users to filter results based on parameters, such as size of item, media type, language, etc. enables refinement of results. Thought may also have to be given to sorting text written using different alphabets |
| Providing translation in English of documents in the target language(s) | Many users, particularly on the Web, may have language skills in their mother tongue and English. In many cases users will turn to English if information on the Web in their native language is not available. Therefore providing an English translation may provide a wider range of users with some level of translation support |
| Providing translations of documents in the target language(s) | If the user wants to examine a document it may be necessary to translate it. This translation can be done using MT as the user requests to view a document or generated in advance (but this must be done for all possible/likely source languages). The user should be able to get access to the document in the original target language as well as the translated version |


**Fig. 4.6** Ranked results and document surrogates (snippets) from Google Translate (translate.google.com). This screenshot is copyright of Google

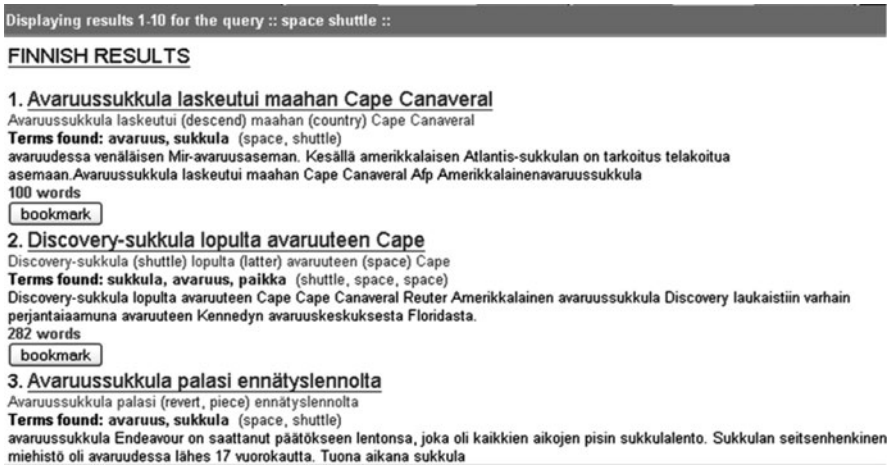


**Fig. 4.7** Document selection in the Clarity system (Petrelli et al. 2004)

the results (e.g., images or news). Results are commonly ranked according to some underlying principle, often depending on the type of media. For example, web pages are often ranked according to a notion of authority (i.e., how many other web pages link to them); results from searching news stories are commonly ranked in chronological order (i.e., most recent first).

Figure 4.6 shows a typical Search Engine Results Page (SERP) from the 'Translated Search' function of Google Translate. The results consist of document surrogates (snippets) that include a title, query-biased summary and URL (e.g., http://pintura.aut.org/). The right hand side shows results in the original language (Spanish) with translated versions (English) on the left. Query terms and translations are highlighted in bold text. Figure 4.7 shows the results presented as a ranked list for an English search ('space shuttle') to return Finnish documents in the Clarity system. Query-biased summaries are generated for the results, terms from the titles have English translations in parentheses, query terms found in the document are displayed, and when searching multiple languages the results are

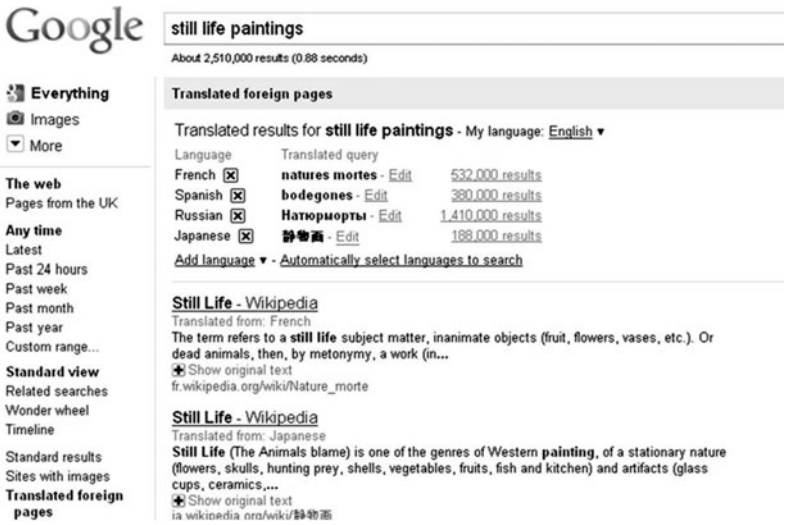


**Fig. 4.8** Example results presentation for Google's web search interface (Google.com) after selecting the 'Translated foreign pages' search option. This screenshot is copyright of Google

grouped into their target languages (e.g., if the target languages are specified as Finnish and Latvian, the Finnish results are presented first followed by the Latvian).

Figure 4.8 shows an example of results presentation and user interaction with the standard Google web search interface[15] after the user selects to view 'Translated foreign pages' for the query 'still life paintings'. This invokes a query translation step showing the user the number of results for each translated query. If they have the relevant language skills they can edit the query and also add further languages for translation or remove them. The resulting document surrogates (snippets) are translated into the source language (English) automatically and the target language of the result indicated to the user under the heading 'Translated from'. When the user selects a result link to view, the underlying document pointed to by the link is translated into the source language, which creates a monolingual search experience in the users' source language.

The presentation of results and search functionalities offered to users may also differ by media type. For example, it is common to display image search results as a two-dimensional grid of image thumbnails, possibly with additional metadata such as the size of image and the URL of the document containing the image. Since it is increasingly difficult to display all information in the limited space of one screen,

[15] This cross-language search functionality is different from Google Translate and is available as a part of the main web search interface. Google Translate, on the other hand, is a stand-alone translation service that can be used for various purposes, such as translating a given URL or a piece of text, but also allows users to perform a web search starting from Google Translate.


**Fig. 4.9** Results presentation for FlickrArabic (Clough et al. 2007)

there is often a balance that must be struck between showing a small amount of detailed information and providing a large amount of more abstract information.

Figure 4.9 shows FlickrArabic (Clough et al. 2007), a system for searching images from Flickr.com, a large-scale, web-based image database based on a large social network of online users. The application is used to manage and share personal (and increasingly more commercial) photographs. The FlickrArabic application translates a users' query from Arabic into English, which is used as an interlingua, or pivot language, to translate into French, Spanish, German, Italian or Dutch. This is necessary because many translation resources only offer Arabic to English. The English translation is shown to users who can modify the query if they have sufficient language skills. In the case of polysemous Arabic queries, translations for all senses are displayed to the user and they are able to accept all or just to select the correct translations. Translation between Arabic and English is performed using a bilingual dictionary. Translation between the English and other languages is performed using an online MT system and therefore users have little control over translation. The results in different target languages are shown in the tabs, which upon selection translate the English query into the selected language. This is an interesting cross-language application as results often reflect cultural differences in an obvious manner (e.g., when searching for cars in German many pictures of German manufactured cars are shown; in English cars made by American manufacturers are shown).

### 4.4.3 Query Reformulation

Query reformulation is one of the key stages in the search process. After conducting an initial search and depending on the results, users may wish to refine their search to improve the results or to reflect changes in their information need. This can often involve the user broadening or narrowing a search through the deletion or addition of query terms (query expansion). Various studies have shown that over half of users reformulate their queries (Jansen et al. 2005, Wildemuth 2006) and supporting this activity is an important part of designing interactive search interfaces. Many approaches have been suggested for assisting users with query reformulation. For example, displaying search terms in such a way that users can edit them, providing term suggestions, and allowing users to search within results.

Studies have shown that query reformulations often fall into general categories. For example, Rieh and Xie (2006) investigated query reformulations by analysing the query logs from a major search engine and grouped reformulations into the following major categories:

- *Specified reformulation*. In this case users will typically add terms to the query to narrow the subject of their search, e.g., 'dancing' to 'ballroom dancing'. Users may interact in this way to deal with the search engine returning too many results. This type of reformulation was found to account for 29.1% of all reformulations.
- *Generalised reformulation*. In this case users will typically delete terms to broaden the subject of their search, e.g., 'Freudian theory' to 'Freud'. Users may do this if the search engine returns too few results. This type of reformulation was found to account for 15.8% of all reformulations.
- *Parallel reformulation*. In this case the users will use a mix of adding, deleting or replacing terms (e.g., with synonyms), for example to search on different aspects of the same topic, e.g., 'photos Sheffield' to 'maps Sheffield' to 'hotels Sheffield'. This was found to be the most frequent type of reformulation accounting for 51.4% of all reformulations.

Further patterns of query reformulation may include: changes in word ordering, URL stripping (e.g., 'http://yahoo.co.uk/' to 'yahoo'), expansion of acronyms, addition or removal of whitespace characters (e.g., 'wal mart tomatoprices' to 'walmart tomato prices'), spelling correction and use of morphological variants (e.g., 'running over bridges' to 'run over bridge').

One way that search engines support query reformulation is through the use of spelling error detection and correction and using *automated term suggestion* to suggest alternative query terms. These can be used to expand a user's query and may appear as the user types (dynamic term suggestion). For example, a Google search for 'harvrd referencing' initiates a search for '*harvard* referencing' instead and offers the following related searches: 'harvard referencing examples', 'harvard referencing internet', 'harvard referencing websites', 'harvard referencing software', 'harvard referencing bibliography', "harvard referencing footnotes', 'harvard

referencing magazine' and 'apa referencing'. In a MLIR/CLIR system the query expansion terms or suggested alternatives will be offered in the language used for querying (the source language). For example, searching Google for 'voitue citroen' (where the French word for 'car' – 'voiture' – has been misspelled) generates a search for '*voiture* citreon' and the following user feedback[16]: 'Showing results for voiture citroen'.

Query reformulation suggestions are often based on past user interactions recorded in transaction logs or based on thesaurus look-up. EuroWordNet for example, is a multilingual database constructed from language-specific WordNets joined through an Inter-Lingual Index to provide translation equivalent relations. Individual WordNets are built around the concept of synsets which are groups of semantically equivalent words or phrases, and joined via semantic relations such as hypernymy, hyponymy and antonymy. For example, for the word 'car', the synset [motor vehicle; automotive vehicle] is a hyponym (kind-of), the synset [cab; taxi; hack; taxicab] a hypernym (is-a), [car door] a meronym (part-of). These relations provide a structure which can also be used to support query translation and expansion over the different languages included in EuroWordNet (see, e.g., De Luca and Nürnberger 2006).

Suggestions for query reformulation can also be based on using *relevance feedback* where users can indicate documents which are relevant to their search and from which suggested terms are derived. Automated approaches that do not involve user interaction (*pseudo* or *blind relevance feedback*) can also be used where the top $n$ documents are assumed to be relevant and used to derive $k$ expansion terms (see Chapter 2). Relevance feedback can also be used to show related articles and provide 'find similar' or 'more like this' functionalities.

### 4.4.4 Browsing and Visualisation

In addition to supporting querying, the IR application may also want to support the wider information seeking behaviour of users, for example that people may also browse for relevant information: following links from one piece of information to another. Approaches may be used to cluster (or group) related items, such as search results on specific topics, that may also exploit existing knowledge sources. Most searchable information resources provide a combination of free-text search and browsing functions. In Wikipedia, for example, users can begin their search by issuing a query to find a page of interest, and then navigate through browsable links to other pages within Wikipedia, including pages on the same topic in other languages, or pages from external sites.

[16] If French was selected as the interface language the phrase 'Showing results for' is changed to the equivalent 'Résultats pour'.

Functionalities that aim to assist users with browsing help users to navigate an information space, e.g., a set of search results, a collection of documents or an interconnected network of nodes, such as a website. These may include organising and presenting content in a form that is navigatable (e.g., through hyperlinks), providing aids to help users navigate and locate themselves in the information space (e.g., links and history lists), cross-referencing and linking between content (e.g., using glossaries, controlled vocabularies or bibliographic information) and providing overviews of the information space using visualisation techniques or logical structures (e.g., a table of contents).

Advances in areas of content classification and enrichment have generated content with more explicit semantic information. Techniques such as high-level overviews, the grouping and connection of related objects to create rich conceptual graphs, rapid previews of objects (e.g., 'dynamic query' interfaces) and visualisation techniques ranging from ranked lists, clustered result displays, tag clouds, cluster maps, and data-specific designs such as timelines can all help users to understand data structures and infer relationships, thereby facilitating information access (see, e.g., Wilson et al. (2010) for further details). To provide browsing functionalities the items in a collection may be connected through an ontology[17] or taxonomy to provide a common structure and vocabulary with which to navigate items (e.g., navigating to related items).

Collier et al. (2006) discuss the creation of a multilingual taxonomy (Chinese, English, Korean, Japanese, Thai and Vietnamese) to support access to information about rapidly spreading infectious diseases and to provide global health monitoring. The resulting multilingual taxonomy allows users to navigate health information through the conceptual structure provided by the taxonomy. In addition, the taxonomy can be used to refine a user's queries, e.g., as a source of broader and narrower terms. For example, the English concept 'bird flu' is mapped to DISEASE_373 ('avian influenza') which has a range of equivalent expressions in various languages (e.g., โรคไข้หวัดนก in Thai and 家禽ペスト in Japanese). The taxonomy, and associated linked documents, can be browsed within an online application called Biocaster[18] (see Figure 4.10). Items in the underlying taxonomy are hierarchically ordered and browseable. Selecting a concept from the taxonomy, e.g., 'chills', displays details about the concept including a description and multilanguage versions of the concept.

An alternative approach to organising information using a hierarchical classification scheme is to use a *faceted* classification scheme. This is often more suitable for collections with multiple dimensions and allows users to explore the content by filtering information. A faceted classification scheme allows items to be assigned to

[17] An ontology is an explicit or formal specification of concepts/terms in a particular domain and the relationships between them. The concepts can be arranged in any manner; a taxonomy, on the other hand, is a more restrictive form of ontology that requires that the concepts are arranged in a hierarchy.

[18] http://born.nii.ac.jp/

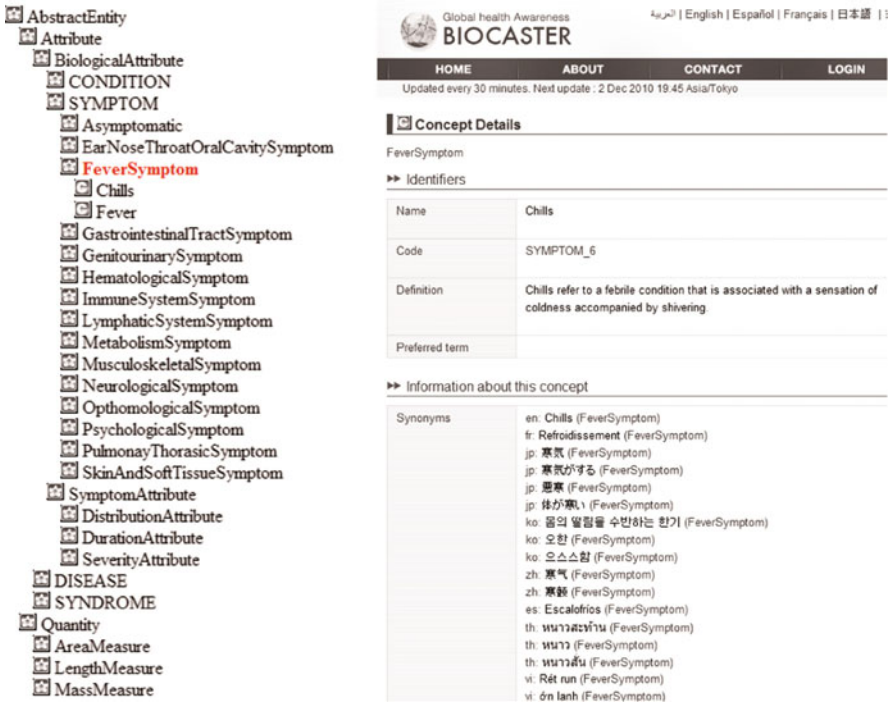


**Fig. 4.10** An example of browsing a hierarchically ordered taxonomy (from Biocaster)

multiple classes and then ordered in multiple ways rather than a single pre-defined ordering as encapsulated in a taxonomy. Each facet has an associated set of possible values and can be used for faceted searching and browsing. Hearst (2006) compares clustering versus faceted categories for information exploration and highlights the benefits of using a faceted scheme in offering a more flexible data model for interactive search.

Figure 4.11 shows an example of faceted browsing and timeline visualisation for artworks and artists from Tate Online, Britain's national art gallery, which was developed as part of the MultiMatch project (Clough et al. 2008). The underlying information is mapped to a faceted classification scheme that can then be navigated and visualised in various ways. The facet 'nationality', for example, shows all possible values, such as 'American', 'Australian', 'British', etc. The number of items associated with the value is shown as a numeric value for each value in the facet. Selecting multiple values filters the collection and the timeline presents filtered results (artists in this example) chronologically.

The previous examples assume that similar items in a collection are already connected, however there may be situations when this is not the case and similar items are therefore grouped on-the-fly. For example, Figure 4.12 shows a form of visualisation based on 'concept hierarchies' used for a text-based English-Italian cross-language image retrieval application (Petrelli and Clough 2006). An online MT

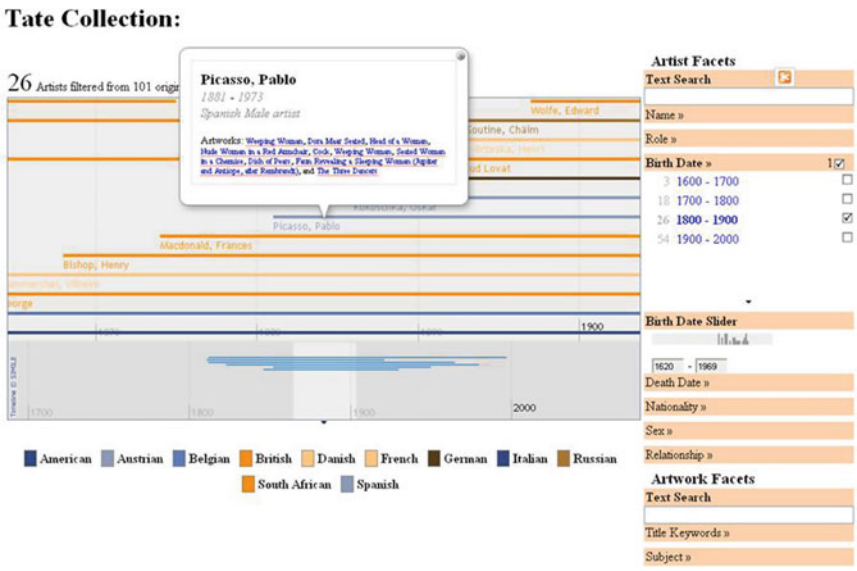


**Fig. 4.11** Example timeline view of artists from Tate Online (Clough et al. 2008)

system was used to translate the user's search request (in Italian) into English, the language of the document collection. The retrieved images are organised into a hierarchical menu based on the overlap of concepts automatically extracted from the image metadata and translated from English into Italian. The extracted terms are organised into a hierarchy of terms (e.g., vehicle $>$ car) where travelling lower down the hierarchy represents a narrowing or subset of the results set. The entire interface (including results) was then translated into Italian using a customised wrapper for the Babelfish online translation service,[19] and finally displayed to the user. Users could view a larger version of the image with caption (translated into Italian) by clicking on the image title. A similar manner of browsing results was also used in the Eurovision cross-language image retrieval system (Clough and Sanderson 2006).

A final example is Arctos, an interactive CLIR system in which 'Document Thumbnail Visualizations' of the retrieved documents were presented with colour highlighting to indicate the position of translated terms in the retrieved documents (Ogden and Davis 2000). Figure 4.13 shows an example visualisation which enables users to quickly scan the returned document set for instances of the query terms (in this case 'pope' and 'beatifications'), query term collocations and their distribution in and between documents. Ogden and Davis found that the thumbnail images could be used to support document selection without users having to know the document's language: the position and collocation of translated query terms was enough to decide whether documents may be relevant or not. Further examples of visualisation techniques can be found in Hearst (2009).

[19] http://babelfish.yahoo.com/


**Fig. 4.12** Clustering the search results using concept hierarchies for cross-language image retrieval (Petrelli and Clough 2006)

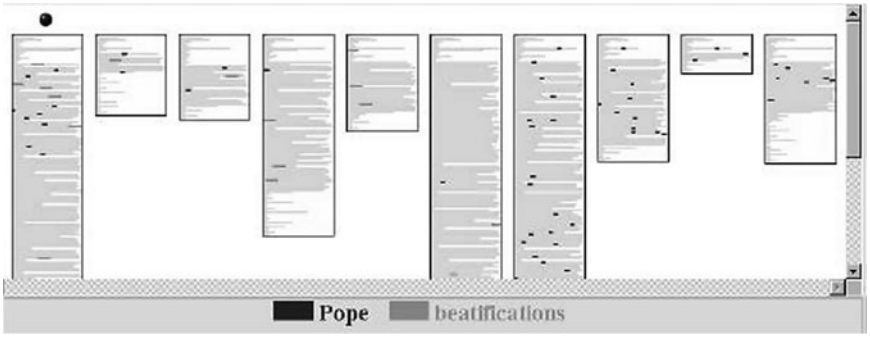


**Fig. 4.13** Results visualisation in the Arctos system. This screenshot is taken from Ogden and Davis (2000)

## 4.5 Designing Multilingual Search User Interfaces

There are at least two aspects to consider when designing user interfaces: handling user-system interaction and creating an interface that is appealing and usable. Designing the interface will involve developing components for navigation and searching (as discussed in Section 4.4), and labelling and positioning these in an intuitive manner on the interface to support interaction. A user's interaction with the search system can take various modes, such as formulating queries in command languages, selecting items from a menu, filling in fields on a form, following hyperlinks, clicking on interface objects (e.g., maps or radio buttons) or using a natural language dialogue.

The aim of MLIR/CLIR interface design is to match the multilingual search tasks that people do (and their goals) with the appropriate interface objects (or functionality). As the search interface acts as an intermediary between users and the search system itself, it must therefore be designed in a way that assists users in clarifying their information needs, formulating and translating suitable queries (possibly in languages they are unfamiliar with), refining their queries and interpreting the results (again, possibly in languages they are unfamiliar with). Resnick and Vaughan (2006) describe a set of best practices developed to assist specifically in the design of search user interfaces. These design principles are organised into five domains: the corpus, search algorithms, user and task context, the search interface and mobility. Best practices include the use of faceted metadata within a controlled corpus, the use of spell-checking during user input, hybrid navigational support through combined search and browse, the use of past queries to frame the search context, the provision of a large query box, the organisation of a large set of search results into categories, showing the keywords in context in search results and designing alternate versions of content specifically for mobile and handheld devices.

These best practices can be used in conjunction with existing guidelines for developing effective search interfaces, for example the framework proposed by Shneiderman et al. (1998), the design recommendations for IR systems by Ahmed et al. (2006) and general guidelines for *usability*[20] (Nielsen 1994). There are further issues to consider in providing effective multilingual information access and supporting cross-language user interaction. These can range from adapting existing information for use by local communities (Section 4.5.2) to providing cross-language search assistance (as discussed in Section 4.4). Based on a subset of the principles of

[20] This relates to the ease of use with which functionality can be accessed, i.e., the ease of use in which a user communicates with a system. Usability depends on characteristics of the user and characteristics of their tasks (or human processes). That is, if the functionality provided is easy to use, but does not address the task at hand, then the system is not successful. Jakob Nielsen's website (http://www.useit.com/) is a useful source of information providing links to resources on many different aspects of usability. A further helpful website that references and implements many proposed usability assessment schemes is Gary Perlman's 'User Interface Usability Evaluation with Web-Based Questionnaires': http://hcibib.org/perlman/question.html

Shneiderman et al. (1998), the following issues may be considered in the design of MLIR/CLIR user interfaces:

- *Strive for consistency in the use of terminology, layout and design*. Further consideration must be given to the choice of labels used in MLIR/CLIR user interfaces as users from different cultural backgrounds may be unfamiliar with colloquial terms, which may also prove difficult to translate and create localised versions of the interface.
- *Users should feel in control of the search process*. This is particularly important for functionalities of MLIR/CLIR systems, such as query translation, which can often seem like a 'black-box' to the user producing unexplained behaviour in the case of translation. However, there must be a trade-off between users being in control versus not being concerned with the inner workings of the system.
- *Users should be able to quickly see the breadth of their options, grasp how to achieve their goals, and do their work*. The options for users with varying language skills should be made clear and in a language that is accessible to them.
- *Effective interfaces do not concern the user with the inner workings of the system*. For MLIR/CLIR systems this may involve hiding the implementation of query translation functionality.
- *Reduce cognitive load*. The amount of processing or mental effort required by a user to complete a task, such as performing a search in an IR system should be as low as possible. As the cognitive effort required by users to search in languages unfamiliar to them will be higher than searching in their native languages, the functionality developed to implement cross-language searching should not impose additional burden on the user.
- *Reduce short-term memory load*. The use of functionalities to promote recognition over recall, such as auto-completion and the use of back-translation for query translation, reducing the need for scrolling, or providing examples of syntax use for query formulation.
- *Effective applications and services perform a maximum of work, while requiring a minimum of information from users*. A basic guideline applicable to monolingual, multilingual and cross-language applications.

An important mantra for user interface design is that people are better at recognising things they have previously experienced than recalling those things from memory. The use of menu-driven over command line interfaces is an example of implementing this theory of memory. The use of auto-term completion seen earlier in Figure 4.2 provides users with a list of possible queries as they submit individual query terms, which they can select as the query to use. This general rule is pertinent to multilingual search interfaces as users who are unfamiliar with a language may find it far easier to recognise something relevant to them than recall it. In fact recall may be virtually impossible as they do not have a prior vocabulary of terms to pick from and therefore providing users with possible translations of a term is a necessary mechanism to formulate queries.

### 4.5.1 User-Centred Design

One aspect that can assist with effective interface design is to adopt a user-centred design approach that focuses on the needs of the end user in an iterative cycle involving identifying users' needs and establishing requirements, and designing/evaluating various kinds of prototype system (Rubin 1994, Preece et al. 2002). The user gets involved at various stages in the life cycle (design, evaluate, re-design); exactly how and when is based on the individual application. The key to this approach is that systems are built to meet users' needs and their individual characteristics (e.g., language skills) are investigated as part of the development. A *persona*, a description of an invented character representative of a key user group, can be developed to aid designers with thinking about prospective end users of the CLIR/CLIR system (and their activities or tasks).

In the case of developing search applications, adopting a user-centred approach to design means that the users' wider information seeking context is considered beyond information retrieval and fits closer to a user-oriented view of IR (Mulhem and Nigay 1996, Ahmed et al. 2006). Example projects that have adopted this development approach for developing MLIR/CLIR systems include Clarity and MultiMatch. In the case of the Clarity project, the following user-centred design cycle was used (Petrelli et al. 2002, Petrelli et al. 2004):

1. *Preliminary requirements*[21] *specification* – informal definition of users' needs gathered through discussions with representative end users.
2. *Scenarios and preliminary design* – scenarios were developed based on the initial requirements specification. This represented the designers' view of possible users, their tasks and their interaction with the proposed system.
3. *Formative evaluation* – the scenarios were used to design an initial user interface. Mock-ups were created to represent different aspects of the users' information seeking behaviour (e.g., query formulation and translation) and were judged by end users during a field study.
4. *Detailed requirements specification* – end users were observed in order to see real users at work in a realistic setting. Further data collection methods were used (e.g., interviews and participatory design) to develop a more complete specification of the proposed system.
5. *Main design phase* – the results from data collection were integrated and used to create prototypes and, through multiple iterations, a final working system.

[21] The process of determining requirements is to gather or capture what a system should do (not how) by: identifying users' needs; generating a set of stable requirements. The first step aims to understand as much as possible about the users, their work, and the context of their work so that the system being built will meet their goals. The second step aims to produce from the needs identified a set of requirements which provides a foundation from which to continue with the design stage. Requirements may include functional and non-functional requirements.

A common technique often employed in the earlier stages is *task analysis* (Hackos and Redish 1998, Preece et al. 2002): learning about ordinary users by observing them in action and understanding how and why users perform certain tasks, including information seeking tasks. This helps to understand issues such as what the user's goals are (i.e., what they are trying to achieve); what tasks users perform to achieve their goals; what personal, social and cultural characteristics users bring to the tasks; how users are influenced by their physical environment; how users' previous knowledge and experience helps; what users value most that will make a new interface (and system) satisfying for them and how a new system can help people do things better, such as searching for information. Studying users can answer questions such as the following (Hackos and Redish 1998):

- What are the individual characteristics of the user that may affect their behaviour with the system being designed? For example, users will have different learning styles which will affect the way in which an interface is used and information is managed.
- What experiences and knowledge do they bring with them to perform the tasks the job requires? For example, what language skills do they have? How long have they been doing the tasks and how did they learn to perform their tasks?
- What do they know about the subject matter and what tools are currently used to perform the tasks today?
- What is their experience at using existing tools and technologies? What tools do they currently use? Are they happy with those tools and how would they like to see the tools extended?
- What are their actual jobs and tasks? What types of searches do they perform? How do they go around looking for information to fulfil their tasks?

Common ways of describing tasks include *scenarios* and *use cases*.[22] A scenario is an informal narrative description (i.e., a story) of human activities or tasks. This is a more natural way for people to describe their goals and tasks and typically does not include information about particular systems or technologies to support the task. These descriptions can then be analysed to extract requirements of the proposed system and build up models of the domain under study (see Chapter 5 for example scenarios). If the design has included creating personas, then multiple scenarios might be developed for each persona.

Many approaches of data collection can be used, such as interviews and questionnaires, to gather quantitative and qualitative data to inform the elicitation of user requirements. In addition, more recently user-system interactions as

[22] A use case is also focused on user goals, but with an emphasis on user-system interaction rather than the task itself. A use case is a "*case of using the (prospective) system*" (used to specify user's functional requirements) and a scenario specifies a flow of events. These can be written in natural language or expressed in graphical form. The results can be fed into the design of the application and inform the specification of system requirements to ensure the provision of functionality to support the users' search tasks and information needs.

captured in transaction log files are being used to gather data to establish the kinds of searches users pose to a retrieval system. Throughout the user-centred design process *formative* evaluation can be used to verify design choices and highlight areas for further revision. At the end of the design process *summative* evaluation can be conducted to evaluate the final design (see Chapter 5 for more information about conducting user-oriented evaluation of MLIR/CLIR systems).

### *4.5.2 Internationalisation and Localisation*

In addition to the design of a cross-language search user interface one must also consider its implementation. A key area for consideration in multilingual interfaces is internationalisation and localisation (Savourel 2001). *Internationalisation* is the process of developing a product in such a way that it works with data in different languages and can be adapted to various target markets without engineering changes, i.e., developing an architecture that is able to accommodate multiple languages. Representing, storing, processing, inputting and displaying languages not based on the Roman alphabet poses a number of difficulties. For example Arabic is written from right to left; Chinese contains thousands of logograms (each one representing a different concept), which prevents a one byte coding; in Thai and many Indian languages, the sequence of characters does not correspond to its phonetic equivalent and one character may even be drawn encircling others, and in Korean, characters are fused to make syllables. In addition to this, languages will differ in the way they are segmented, the ordering in which they should be sorted alphabetically, and the varying use of date and time conventions.

To enable the storage and display of multiple languages, an appropriate character[23] encoding and font set must be utilised (see Chapter 2 for further information on character encoding). As already stated, most multilingual applications are actually using a universal character set, such as UTF-8 or UTF-16 based on Unicode. The correct character encoding must be used otherwise the characters will not appear correctly on the screen. In addition to representing the characters a suitable font set must also be used to map between the codes and the character images (glyphs) that appear on the screen. That is, not only must characters be represented but scripts too (e.g., Latin, Semitic, etc.). The font set must also apply rules for composite characters, ligatures and other script-specific features.

*Localisation* is the subsequent process of translating and adapting a product to a given market's cultural conventions. Localisation can involve customisation of numeric, date and time formats, currency usage, keyboard usage, symbols, icons and colours, legal requirements, rules for sorting and re-designing any references to culturally-specific ideas. For example, 11/02/74 would be interpreted as

[23] Characters include letters, ideographs, digits, punctuation marks, diacritic marks and mathematical symbols.

2nd November 1974 in the US, but the rest of the world would mostly interpret this as 11th February 1974. The use of colours is also an important consideration when tailoring interfaces for certain cultures. For example, red is associated with 'death' in Egypt, 'happiness' in China, 'danger' in Japan and 'freedom' and 'peace' in France. Further useful information about globalising web content can be found in Chapter 13 of McCracken and Wolfe (2004).

Localisation may also involve selecting (or generating) content that is only relevant to specific locations, such as job advertisements and the announcement of events. The process of internationalisation and localisation involves dealing with issues, such as character encoding and the translation of content into multiple languages, and MT technologies are often employed at this stage. Multilingual versions of a website (or IR application) may exhibit different degrees of parallelism, ranging from a collection of monolingual sites at one extreme to a completely parallel site with identical structure, navigation and content at the other. Consideration must also be given to how much content on a site is static and how much is produced dynamically and thereby requires translation at run-time. Approaches to translate an interface may include creating different language versions of labels used on the interface and applying the appropriate language version at run-time.[24]

A multilingual interface must support the input, storage and presentation of different languages. For example, the interface must provide support for entering queries that consist of non-ASCII characters in the search box and this is especially pertinent when developing interfaces for non-Western languages, such as Chinese and Arabic. Many users do not have access to keyboards that support the input of non-ASCII characters; an alternative approach is to offer users a virtual keyboard. An example of such an input mechanism for entering Cyrillic characters for searching Google is shown in Figure 4.14.

Implementation will often involve separating out the elements to be translated when localising the interface, e.g., command prompts or labels that provide information to the user. The use of colloquial expressions, jargon or ambiguous terms should be avoided when designing text for the interface as these can be easily mistranslated (if translated at all). The use of (multilingual) controlled vocabularies of preferred terms can help with reducing translation errors (as well as aid browsing and navigation as discussed in Section 4.4.4).

Typically the user will be able to manually select the interface language. For example, Figure 4.15 shows the user interface for Europeana, a portal that enables access to the digital resources of Europe's museums, libraries, archives and audio-visual collections. Users can select from a range of languages (note that the language is presented in its native form rather than English and with a three letter language code). Selecting Italian, for example, provides a localised version of the interface in that language (Figure 4.16). The language choice is then stored in a

[24] See http://www.multilingual.com/articleDetail.php?id=594 for further details on reusing text in multilingual user interfaces.

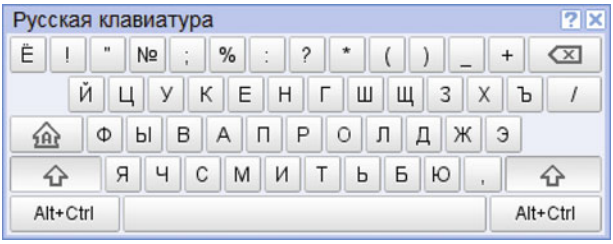


**Fig. 4.14** Example of non-ASCII keyboard character input (Google.com). This screenshot is copyright of Google

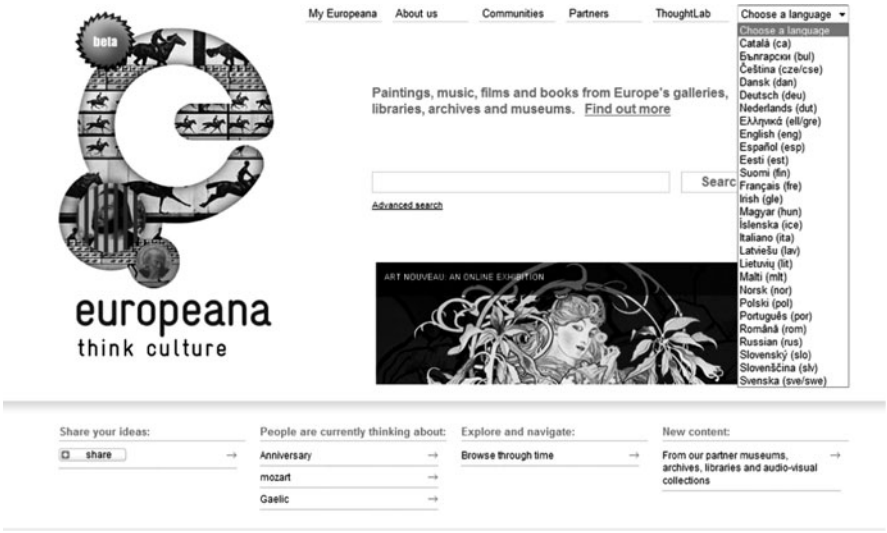


**Fig. 4.15** Example of selecting the interface language from Europeana (http://europeana.eu/)

cookie[25] and the interface is automatically switched to the language that was previously chosen. Taking into account the translation of menu items and the text of interface objects, such as buttons, is important to provide a fully localised interface.

[25] A cookie is a piece of text stored on a user's computer by their web browser. A cookie can be used for authentication, storing site preferences or tracking users through a session, e.g., paying for goods from an online shop. The cookie can be used to add persistence to a web application.


**Fig. 4.16** Example of Europeana interface localised to Italian (http://europeana.eu/)

Making an interface appealing for a particular audience will involve more than just translation; it will also include recognising and adhering to cultural conventions, e.g., the way in which dates and times are represented, or the use of particular icons. These issues will become important if adapting an IR application to a particular cultural background (Shen et al. 2006).

### *4.5.3 Case Study: CLIR in Google's Web Search*

To demonstrate the various aspects to consider when designing multilingual search interfaces we consider the user interface for Google's translated web search. The simple user interface helps to reduce cognitive load and the cross-language functionality is implemented in such a way that it can provide users with a seamless monolingual search experience, from search through to the translation of original texts. The cross-language functionality in Google's web search includes the following.[26]

**User interface**: The user can select the interface language (locale) and save the settings (personalisation). The localisation of the interface includes the correct page layout for languages that display from right to left and the handling of

[26] This is the functionality provided as of January 2011. Note that Google is given as a case study not because it is the only (or best) way of doing things, but because it exhibits many of the ideas discussed in this chapter.

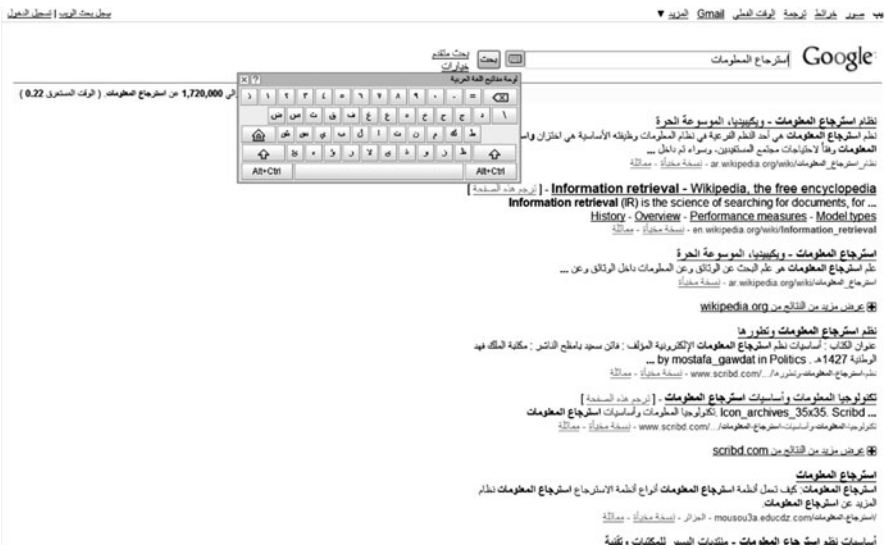


**Fig. 4.17** Example user interface for Google web search in Arabic (http://www.google.ae/). This screenshot is copyright of Google

non-Western scripts (e.g., Arabic), as well as the translation of user input fields and buttons and instructions. Figure 4.17 shows an example of Google with the interface language set to Arabic highlighting the correct page layout, the translation of textual strings and the provision of a virtual keyboard for inputting queries in non-Latin script languages without using language-specific keyboards.

**Query formulation and translation**: Query translation is invoked when the user enters a query and selects the 'translated foreign pages' link. The user's query is translated automatically and results shown immediately to the user; there is no interactive query translation step to begin with. The user's source language defaults to the interface language but can be manually selected under the option 'My language'. Initially Google automatically selects the target languages but the user can specify up to five languages – these can be manually added and removed. Each time the user selects another target language a search is initiated immediately and results updated. The estimated number of hits is shown per target language, which may help guide the user's search.

**Document selection and examination:** A single ranked list is produced with results from the different target languages interleaved (Figure 4.8). Surrogates of each search result are generated and consist of a query-biased summary, title of the web page and URL. The interface also indicates which target language the result is in ('Translated from . . .'). The surrogates are translated into the user's source language and presented in the results list, but a link is also provided to the original text ('Show original text'). Query terms in the source language are highlighted in the translated version of the surrogate and the translated version of

the query in the original text. For each item in the results list users can select a URL to navigate to the original page. This will invoke Google Translate which translates the original page into the source language (the original page can also be viewed). From this point on, each link the user selects will invoke Google Translate which creates a smooth searching experience in the user's source language.

**Query reformulation and refinement:** The user can edit the translated query in each target language. The virtual keyboard appears when necessary. When a new target language is selected, the search results are immediately updated with the results for the added target language interleaved with the results from the current target languages. Through the search customisation options (under 'search settings') the user can specify to only display results written in specified language(s).

## 4.6 Summary and Future Directions

This chapter has focused on the design of multilingual and cross-language searching systems from the perspective of the user. It is vital to consider this broader view so that retrieval systems supporting a user's tasks and adaptable to specific situations and contexts can be developed. This user-oriented view is complementary to system-oriented development as it seeks to understand the human, or user, role in accessing information and is now seen as a vital factor in producing IR systems that meet the needs of the end user and therefore succeed in the market-place: "*most IR systems are used by people and we cannot design effective IR systems without some knowledge of how users interact with them*" (Robins 2000, p. 57). Interactive IR provides frameworks which help to understand the users and their characteristics, the users' information needs and their context when accessing information and the users' interactions with information and search systems. This complements a more system-oriented perspective of information retrieval. The design of search assistance for the user interface will depend on the users' search context (e.g., their work tasks) and technical issues, such as how the IR application has been implemented, which may also involve interaction with designers of the system.

The level of support provided by an IR application will vary depending on a number of contextual factors such as the individual characteristics of the user (e.g., language skills), their familiarity with the search task in hand and their past experiences. The situation is compounded in cross-language search where the users' language skills and cultural background will affect their experience and interactions with the search system. In addition to providing aids to assist with query formulation and translation, document selection and examination and query reformulation, the user interface should be designed to accommodate best practices in general user interface design and usability guidelines. The role of localisation is particularly important as user interfaces are adapted to meet the languages specific to a particular community or culture. This may include consideration of issues, such

as character encoding and the design of informative labels suitable for a multilingual audience and for translation.

Future directions for cross-language search interfaces include the development of more naturalistic human-computer dialogue mechanisms, improved real-time translation, in-depth and broader studies of users to develop use cases for cross-language search, further studies to investigate the effects of language skills and cultural differences on interactive search and the design of cross-cultural, as well as cross-language, search interfaces. Advances in semantic enrichment of documents will facilitate new forms of information categorisation, structuring, visualisation, browsing and supporting more exploratory forms of user information behaviour as highlighted in Wilson et al. (2010). The use of automated categorisation, summarisation and visualisations may all aid users in multilingual searching (Chung 2008).

## 4.7 Suggested Reading

One of the most comprehensive resources for the general design of search user interfaces is *Search User Interfaces* by Marti Hearst (Hearst 2009).[27] This provides a comprehensive overview of the topic with many of the insights applicable to developing user interfaces for multilingual information access systems. Two further helpful resources are *Information Seeking in Electronic Environments* by Marchionini (1995) and *Interactive Information Retrieval* by Ruthven (2009). Both discuss broader issues, such as information seeking and behaviour, interactivity in search, and users' individual differences and their effect on searching behaviour. A helpful discussion of interaction in information searching and retrieval is provided by Beaulieu (2000). From a more practical perspective, *Information Architecture* by Morville and Rosenfeld (2006) provides a summary of general design principles for large-scale websites and web-based information services or applications. In particular the book highlights the use of knowledge structures to organise the contents of websites and design interfaces that incorporate search and browsing functionalities into applications. The overview article by Wilson et al. (2010) entitled 'From keyword search to exploration: designing future search interface for the Web' provides a comprehensive summary of current and future developments in search assistance, particularly in relation to developments in the Semantic Web and Natural Language Processing.

The study of interactivity in CLIR ranges from studying aspects of the search process such as document selection (Oard et al. 2004), query translation (Oard et al. 2008), presentation of search results (Ogden et al. 1999, Petrelli and Clough 2006); to the entire search process (Petrelli et al. 2002, 2004, Ogden et al. 1999, Ogden and

[27] This is also available for free at http://searchuserinterfaces.com/book/

Davis 2000, Capstick et al. 2000, Peňas et al. 2001). Participation in the Cross-Language Evaluation Forum (CLEF) interactive or iCLEF[28] track (Oard and Gonzalo 2002) has shown some interesting search behaviours from users such as adopting terms from relevant documents during query refinement (thereby confirming the need for document translations and consistency of translation resources used) and different strategies for query formulation (He et al. 2003). The report by Braschler and Gonzalo (2009) provides a summary of work conducted at the Cross-Language Evaluation Forum (CLEF) and highlights both system-oriented and user-oriented best practices for developing multilingual information access systems.

Many resources discuss human computer interaction and interface design, such as *Human computer interaction* by Dix et al. (2004). The classic 'golden rules' of interface design by Shneiderman et al (1998) are as applicable to the design of cross-language and multilingual search interfaces as for monolingual ones. In addition, Nielsen (1994) provides five attributes of usability that can be used to assess how usable an interface or system is: learnability, memorability, efficiency, errors (accuracy) and subjective satisfaction. This is important as studies continue to show that users still find IR systems difficult to learn, use, and remember (Ahmed et al. 2006). There are issues which are specifically related to internationalising user interfaces and del Galdo and Nielsen (1996), De Troyer and Casteleyn (2004) and McCracken and Wolfe (2004) provide guidelines for exposing user interfaces to international audiences.

## References

Ahmed Z, McKnight C, Oppenheim C (2006) A user-centred design and evaluation of IR interfaces. J. of Librariansh. and Inf. Sci. 38(3):157–172

Aula A, Kellar M (2009) Multilingual search strategies. In: Proc. 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '09). ACM Press, Boston, MA, USA, pp. 3865–3870

Barry CL, Schamber L (1998) Users' criteria for relevance evaluation: a cross-situational comparison. Inf. Process. and Manag. 31(2/3):219–236

Beaulieu M (2000) Interaction in information searching and retrieval. J. of Documentation 56 (4):431–439

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci. Am. 35–43

Braschler M, Gonzalo M (2009) Best practices in system-oriented and user-oriented multilingual information access. TrebleCLEF Deliverable 3.3. Available at http://www.trebleclef.eu/getfile.php?id=254. Cited 15 Feb 2011

Broder A (2002) A taxonomy of web search. SIGIR Forum 36(2):3–10

Capstick J, Diagne AK, Erbach, G, Uszkoreit H, Leisenberg A, Leisenberg M (2000) A system for supporting cross-lingual information retrieval. Inf. Process. and Manag. 36(2):275–289

Chung W (2008) Web searching in a multilingual world. Commun. of the ACM 51(5):32–40

[28] http://nlp.uned.es/iCLEF/

Clough P, Sanderson M (2006) User experiments with the Eurovision cross-language image retrieval system. J. of the Am. Soc. for Inf. Sci. and Technol. 57(5):697–708

Clough P, Al-Maskari A, Darwish K (2007) Providing multilingual access to Flickr for Arabic users. In: Proc. of Evaluation of Multilingual and Multimodal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006). LNCS 4730: 205–216

Clough P, Marlow J, Ireson N (2008) Enabling semantic access to cultural heritage: A case study of Tate online. In: Proc. of the ECDL 2008 Workshop on Information Access to Cultural Heritage, IACH2008, Aarhus, Denmark. Available at http://ilps.science.uva.nl/IACH2008/proceedings/proceedings.html. Cited 15 Feb 2011

Clough P, Eleta I (2010) Investigating language skills and field of knowledge on multilingual information access in digital libraries. Int. J. of Digit. Libr. Syst. 1(1):89–103

Collier N, Kawazoe A, Jin L, Shigematsu M, Dien D, Barrero R, Takeuchi K, Kawtrakul A (2006) A multilingual ontology for infectious disease surveillance: rationale, design and challenges. Language Resources and Evaluation, 40(3–4). DOI: 10.1007/s10579-007-9019-7

Colowick S (2008) Multilingual search with PanImages. MultiLingual 19(2): 61–63. Available at http://turing.cs.washington.edu/PanImMultilingual.pdf. Cited at 15 Feb 2011

Del Galdo EM, Nielsen J (1996) International user interfaces. New York: John Wiley & Sons

De Luca EW, Nürnberger A (2006) LexiRes: A tool for exploring and restructuring EuroWordNet for information retrieval. In: Proc. of the Workshop on Text-based Information Retrieval (TIR-06). In conjunction with the 17th European Conference on Artificial Intelligence (ECAI'06). Riva del Garda, Italy

De Troyer O, Casteleyn S (2004) Designing localized web sites. In: Proc. of the 5th International Conference on Web Information Systems Engineering (WISE2004): 547–558

Dix AJ, Finlay JE, Abowd GD, Beale R (2004) Human-computer interaction, 3rd edition, Prentice Hall

Hackos J, Redish J (1998) User and task analysis for interface design. John Wiley & Sons, New York

Hansen P, Karlgren J (2005) Effects of foreign language and task scenario on relevance assessment. J. of Documentation 61(5):623–639

He D, Wang J, Oard D, Nossal M (2003) Comparing user-assisted and automatic query translation. In: Proc. of 3rd Workshop of the Cross-Language Evaluation Forum (CLEF 2002). LNCS 2785: 400–415

Hearst M (2006) Clustering versus faceted categories for information exploration. Commun. of the ACM 49(4):59–61

Hearst MA (2009) Search user interfaces. Cambridge University Press

Ingwersen P (1996) Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. J. of Documentation 52(1): 3–50.

Ingwersen P, Järvelin K (2005) The turn: integration of information seeking and retrieval in context. Springer, Dordrecht, The Netherlands

Jansen BJ, Spink A, Pedersen J (2005) A temporal comparison of AltaVista web searching: research articles. J. of the Am. Soc. for Inf. Sci. and Technology 56(6):559–570.

Koll M (2000) Track 3: Information retrieval. Bull. of the Am. Soc. for Inf. Sci. 26(2). Available at http://www.asis.org/Bulletin/Jan-00/track_3.html. Cited 15 Feb 2011

Kralisch A, Berendt B (2004) Cultural determinants of search behaviour on websites. In: Proc. of the Sixth International Workshop on Internationalisation of Products and Systems. Vancouver, Canada, 8–10 July 2004

Lazarinis F, Vilares J, Tait J, Efthimiadis EN (2009) Current research issues and trends in non-English web searching. Inf. Retr. 12(3):230–250

Lorigo L, Pan B, Hembrooke H, Joachims T, Granka L, Gay G (2006) The influence of task and gender on search and evaluation behavior using Google. Inf. Process. and Manag. 42 (4):1123–1131

Marchionini G (1992) Interfaces for end-user information seeking. J. of the Am. Soc. for Inf. Sci. 43(2):156–163

Marchionini G (1995) Information seeking in electronic environments. Cambridge University Press

Marchionini G (2006) Exploratory search: from finding to understanding. Commun. of the ACM 49(4):41–46

Marlow J, Clough P, Dance K (2007) Multilingual needs of cultural heritage website visitors: A case study of Tate Online. In: Proc. of International Cultural Heritage Informatics Meeting (ICHIM07). Available at http://www.archimuse.com/ichim07/papers/marlow/marlow.html. Cited 15 Feb 2011

Marlow J, Clough P, Cigarrán Recuero J, Artiles J (2008a) Exploring the effects of language skills on multilingual web search. In: Proc. of the 30th European Conference on IR Research (ECIR'08). LNCS 5478: 126–137

Marlow J, Clough P, Ireson N, Cigarrán Recuero J, Artiles J, Debole F (2008b) The MultiMatch project: Multilingual/multimedia access to cultural heritage on the web, In: Proc. of Museums on the Web Conference (MW2008). Available at http://www.archimuse.com/mw2008/papers/marlow/marlow.html. Cited 15 Feb 2011

McCracken DD, Wolfe RJ (2004) User-centred website development: A human-computer interaction approach. Pearson Education Inc.

Minelli S, Marlow J, Clough P, Cigarran J, Gonzalo J, Oomen J (2007) Gathering requirements for multilingual search of audiovisual material in cultural heritage. In: Proc. of Workshop on User Centricity – state of the art (16th IST Mobile and Wireless Communications Summit), Budapest, Hungary

Morville P, Rosenfeld L (2006) Information architecture for the world wide web: Designing large-scale web sites. 3rd Edition. O'Reilly

Mulhem P, Nigay L (1996) Interactive information retrieval systems: from user centred interface design to software design. In: Proc. of the 19th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96). ACM Press. Zurich, Switzerland: 326–334

Nielsen J (1994) Heuristic evaluation. In: Nielsen J, Mack RL (eds.), Usability Inspection Methods. John Wiley & Sons, New York

Oard D, Gonzalo J (2002) The CLEF 2001 interactive track. In: Proc. of 2nd Workshop of the Cross-Language Evaluation Forum (CLEF 2001). LNCS 2406: 372–382

Oard D, Gonzalo J, Sanderson M, López-Ostenero F, Wang J (2004) Interactive cross-language document selection. Inf. Retr. 7(1–2):205–228

Oard DW, He D, Wang J (2008) User-assisted query translation for interactive cross-language information retrieval. Inf. Process. & Manag. 44(1):180–211

Ogden WC, Davis MW (2000) Improving cross-language text retrieval with human interactions. In: Proc. of the Hawaii International Conference on System Science (HICSS-33). Vol. 3

Ogden W, Cowie J, Davis M, Ludovik E, Nirenburg S, Molina-Salgado H (1999) Keizai: An interactive cross-language text retrieval system. Machine Translation Summit VII, Workshop on Machine Translation for Cross-Language Information Retrieval, Singapore

Peňas A, Gonzalo J, Verdejo F (2001) Cross-language information access through phrase browsing. 6th International Conference of Natural Language for Information Systems (NLDB'01), Madrid, Spain

Petrelli D, Hansen P, Beaulieu M, Sanderson M (2002) User requirement elicitation for Cross-Language Information Retrieval. New Rev. of Inf. Behav. Res. 3:17–35

Petrelli D, Hansen P, Beaulieu M, Sanderson M, Demetriou G, Herring P (2004) Observing users – designing Clarity: A case study on the user-centred design of a cross-language retrieval system. J. of the Am. Society for Inf. Sci. and Technology 55(10):923–934

Petrelli D, Clough P (2006) Using concept hierarchies in text-based image retrieval: A user evaluation, In: Accessing Multilingual Information Repositories. LNCS 4022: 297–306

Preece J, Rogers Y, Sharp H (2002) Interaction design: Beyond human-computer interaction. Wiley, New York

Resnik P (1997) Evaluating multilingual gisting of Web pages. In: Working Notes of the AAAI97 workshop on Cross-Language Text and Speech Retrieval
Resnick M, Vaughan M (2006) Best practices and future visions for search user interfaces. J. of the Am. Soc. for Inf. Sci. and Technol. 57(6):781–787
Rieh HY, Rieh SY (2005). Web searching across languages: Preference and behavior of bilingual academic users in Korea. Libr. & Inf. Sci. Res. 27(2):249–263
Rieh SY, Xie H (2006) Analysis of multiple query reformulations on the Web: The interactive information retrieval context. Inf. Process. & Manag. 42(3):751–768
Robins D (2000) Interactive information retrieval: Context and basic notions. Informing Sci. 3:57–62
Rubin J (1994) Handbook of usability testing: how to plan, design, and conduct effective tests. Wiley, New York
Ruthven I (2009) Interactive information retrieval. Annu. Rev. of Inf. Sci. and Technol. 42 (1):43–91
Savourel Y (2001) XML internationalization and localization. Sams Publishing, Indianapolis
Shen S, Woolley M, Prior S (2006) Towards culture-centred design. Interact. with Comput. 18 (4):820–852
Shneiderman B, Byrd D, Croft B (1998) Sorting out searching: a user-interface framework for text searches. Commun. of the ACM 41(4):95–98
Steinerová J (2008) Seeking relevance in academic information use. Inf. Res. 13(4) paper 380. Available at http://InformationR.net/ir/13-4/paper380.html. Cited 15 Feb 2011
Sutcliffe A, Ennis M (1998) Towards a cognitive theory of information retrieval, Interacting with computers. HCI and Inf. Retr. 10(3):321–351
Taylor R (1968) Question negotiation and information seeking in libraries. Coll. and Res. Libr. 29 (3):178–194
Tombros A, Sanderson M (1998) Advantages of query-biased summaries in IR. In: Proc. of the 21st ACM Conference of the Special Interest Group in Information Retrieval: 2–10
Vakkari P (2003) Task-based information searching. Ann. Rev. of Inf. Sci. and Technol. 37:413–464
Wildemuth BM (2006) Evidence-based practice in search interface design. J. of the Am. Soc. for Inf. Sci. & Technol. 57(6):825–828
Wilson TD (2000) Human information behaviour. Informing Sci. 3(2): 49–56. Available at http://inform.nu/Articles/Vol3/v3n2p49-56.pdf. Cited 15 Feb 2011
Wilson ML, Kules B, Schraefel MC, Shneiderman B (2010) From keyword search to exploration: Designing future search interfaces for the web. Found. and Trends in Web Sci. 2(1):1–97.
Zoe LR, DiMartino D (2000) Cultural diversity and end-user searching: An analysis by gender and language background. Res. Strateg. 17(4):291–305

# Chapter 5
# Evaluation for Multilingual Information Retrieval Systems

*"True genius resides in the capacity for evaluation of uncertain, hazardous, and conflicting information."*

Sir Winston Churchill, 1940s

**Abstract** This chapter discusses IR system evaluation with particular reference to the multilingual context, and presents the most commonly used measures and models. The main focus is on system performance from the viewpoint of retrieval effectiveness. However, we also discuss evaluation from a user-oriented perspective and address questions such as how to assess whether the system satisfies the requirements of its users. The objective is to give the reader a working knowledge of how to set about MLIR/CLIR system evaluation. In addition, we report on some of the ways in which evaluation experiments and evaluation campaigns have helped to achieve greater understanding of the issues involved in MLIR/CLIR system development and have contributed to advancing the state-of-the-art.

## 5.1 Introduction

The aim of any IR system is to provide users with easy access to, and interaction with, information that is relevant to their needs and enable them to effectively use this information; the aim of a MLIR system is to do this in a multilingual and/or cross-language context. The task of an evaluation action is to measure how successful the system has been at this. Success depends on a number of contributing factors including the retrieval mechanism used, how this mechanism processes the search request, the underlying information need, but also the way in which the results are presented and to what extent they meet the user's expectations. This adds complexity as system effectiveness needs to be assessed from diverse perspectives. Evaluation in IR also plays another very important role: most of the advances in the state-of-the-art of IR system research and development have been achieved as the result of experiments aimed at assessing the particular contribution of a given system feature or retrieval mechanism. In an MLIR system, depending on the languages that are handled by the system and the translation components deployed, diverse optimisation methods may be necessary. Evaluation is thus not only an

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_5, 

essential part of the MLIR/CLIR system building cycle but is also crucial in making progress towards building better systems.

There are two broad classes of IR evaluation: user-oriented and system-oriented. User-oriented evaluation measures the user's satisfaction with the system while system-oriented evaluation mainly focuses on indexing methods and algorithms that match documents against queries and rank the results. This implies that the validity of the system and components should be assessed from both usage and technology perspectives. However, as user-based evaluation tends to be expensive and difficult to do correctly, IR researchers have mostly focused on system evaluation. Figures 3.2–3.4 in Chapter 3 illustrated that evaluation in the MLIR context is particularly complex as, in addition to modules for indexing and matching, the MLIR system typically includes language-specific processing and translation procedures. Ideally an MLIR system evaluation protocol requires distinguishing methodological aspects (generic across languages) from linguistic knowledge (specific to particular languages). This would involve investigating not only overall system performance but also the effectiveness of single components and their behaviour with respect to the language(s) being processed.

In this chapter we outline what is implied by IR evaluation with particular reference to the multilingual context, and present the most commonly used measures and models. In accordance with current practice, the main focus will be on system performance from the viewpoint of retrieval effectiveness. However, we also discuss user-oriented questions, such as how to assess whether the system satisfies the requirements of its users. The objective is to give the reader a working knowledge of how to set about MLIR system evaluation. In addition, we report on some of the ways in which evaluation experiments have helped to achieve greater understanding of the issues involved in MLIR/CLIR system development and have contributed to advancing the state-of-the-art. For a more exhaustive perspective on IR system evaluation the reader is referred to the list of suggested reading at the end of this chapter.

## 5.2 System-Oriented Evaluation

System-oriented evaluation focuses on measuring the performance of an IR system or retrieval strategy in an objective controlled setting. One of the primary distinctions made is between effectiveness and efficiency, where effectiveness basically measures the ability of the system to find the right information (i.e., discriminate between relevant and non-relevant documents) and efficiency measures the speed with which this is achieved and the associated storage requirements. Effectiveness is thus defined as how well the system ranks the information in order of relevance while efficiency is defined in terms of the time and space requirements of the ranking algorithm. Generally speaking, IR research has concentrated on improving search effectiveness; only when a technique is established as potentially useful does the focus shift to finding efficient

implementations (Croft et al. 2009). In this section, we discuss evaluation techniques that focus on measuring system effectiveness.

### 5.2.1 The Cranfield Tradition

The first proposals for IR system evaluation were made in the 1950s and the best known early work consists of the Cranfield studies that ran from the late 1950s to the mid-1960s. The aim for the Cranfield experiments was to create "*a laboratory type situation where, freed as far as possible from the combination of operational variables, the performance of index languages could be considered in isolation*" (Cleverdon 1967). The components of the Cranfield experiments were a small collection of documents, a set of test queries and, for each query, a set of relevance judgments with respect to the document collection. The performance of indexing and retrieval strategies were compared under various controlled conditions. The measures used in the Cranfield II experiments to measure the effectiveness of retrieval are precision and recall as defined in Chapter 2: precision is the proportion of retrieved documents that are relevant and recall is the proportion of relevant documents that are retrieved. Typically, returning more documents means potentially sacrificing precision, and guaranteeing precision means sacrificing recall. These measures have been the most widely used in academic IR literature and remain popular today.

This Cranfield test collection model has also remained popular and is still commonly used for comparative system evaluation. It has been refined and extended through the collective experience of IR researchers and is the main model employed in the best known IR evaluation campaigns described below. It is used to simulate in a very simple way the real world search situation. The model as it is understood today can be summarised as follows:

1. Retrieval strategies to be compared produce ranked lists of documents for each query;
2. The effectiveness of a strategy for a single query is computed as a function of the ranks of the relevant documents;
3. The effectiveness of the strategy on the whole is computed as the average score over the set of queries in the test collection;
4. The scores measured are used to indicate the performance of one system relative to another.

Note that the comparison is between relative rankings and not absolute scores; the assumption is that a similar relative performance will be observed on other test collections and in operational settings. The vast majority of test collection experiments assume that relevance is a binary choice, though the original Cranfield experiments used a five-point relevance scale.

In order to be viable in a laboratory setting the Cranfield paradigm makes several important assumptions: (1) that the relevance of one document is independent of the

relevance of others, that all relevant documents are equally important and that the user information need remains static; (2) a single set of judgments for a query is representative of the user population; (3) the lists of relevant documents for each query is exhaustive (Voorhees 2002). Although these assumptions clearly do not bear up in a real world search scenario where users will often refine their information needs during the search session and relevance is subjective and based on the specific interest of the individual searcher, they have been shown to be adequate for controlled laboratory experiments directed at comparative analysis. The focus of the Cranfield methodology is on the effectiveness of system performance in terms of finding and ranking relevant documents rather than with other aspects that can impact on user satisfaction such as ease of use, speed of response, or presentation of results.

Shareable and reusable test collections are recognised as being of crucial importance in IR research and development. They are used to compare the effectiveness of different search strategies, either to contrast the performance of different systems or to test different configurations of a single system. Following the Cranfield design, the main components of a typical IR test collection are: a set of documents representative of a real-world scenario; statements of information needs (sometimes denoted topics) expressed as narrative text or a set of keywords that represent realistic search requests; and a set of relevance judgments indicating which documents in the collection are relevant to each search request. The test collection should be accompanied by an appropriate evaluation protocol and guidelines for its adoption. Developers or researchers can test their own system by indexing the set of documents, submitting queries derived by their system from the topic statements provided and, for each query submitted, retrieving a ranked list of document identifiers. These lists are compared against the relevance judgments to identify the relevant documents found and the evaluation protocol is used to calculate the performance of each run. Although this methodology was conceived for use in a monolingual setting it has been demonstrated as effective also when testing the performance of cross-language IR systems.

### *5.2.2 Evaluation Campaigns*

Although it is possible for single research groups or applications to build their own test collection, this tends to be a demanding task as, depending on the size of the document collection used, the number of relevance judgments needed to build the collection is potentially extremely high.[1] Only by employing collaborative approaches does the production of relevance judgments for large test collections become feasible, both through using special techniques (such as pooling, see below)

[1] Of the order |documents| x |topics|

and by distributing the workload generated by the judging process.[2] This is where evaluation campaigns can play a useful role. The regular organisation of such campaigns helps to obtain a critical mass and limits the administrative overheads involved in managing document collections, dealing with copyright issues, and organising workshops in which the results can be discussed. Furthermore, as stated in Chapter 1, evaluation campaigns play an important role in promoting research into system development. There is a duality between research and evaluation. Good research is validated by evaluation and good evaluation environments stimulate future research. In addition, an evaluation campaign provides an important forum where researchers and system developers can come together, exchange ideas and experiences and discuss common problems.

The best known and most widely used test collections are thus those built by the large-scale evaluation campaigns such as the Text REtrieval Conference (TREC)[3] in the US, the NII Text Collection for IR Systems project (NTCIR)[4] in Japan, and the Cross-Language Evaluation Forum (CLEF)[5] in Europe. Recently, the Information Retrieval Society of India launched the Forum for Information Retrieval Evaluation (FIRE)[6] with the mandate of encouraging research in South Asian language information access. A brief history of the contribution that these evaluation initiatives have made to MLIR/CLIR system research and development is given in the Introduction to this book. A more detailed discussion of the results that have been obtained, in particular by CLEF, is given below in Section 5.2.4.

### *5.2.3 Building a Test Collection*

The methodology for test collection building originally introduced by TREC was also adopted by NTCIR, CLEF and FIRE and with appropriate adaptations can be applied by any individual or group to create their own evaluation corpus.

**Documents.** The document collections and also the documents used in the Cranfield experiments were small and for each query it was possible to judge the entire set of documents – nowadays, IR system evaluation normally uses very large test collections to best simulate actual search requirements. However, the acquisition and sharing of a suitably large document collection is not easy. In order to avoid copyright problems, clear agreements must be drawn up with the data providers establishing the conditions under which the use and distribution of the

[2] The recent introduction of alternative methods, such as the Amazon Mechanical Turk crowd sourcing service, can help considerably to reduce the effort required to create test collections, see Section 5.2.5.

[3] http://trec.nist.gov/

[4] http://research.nii.ac.jp/ntcir/

[5] http://www.clef-campaign.org/

[6] http://www.isical.ac.in/~clia/index.html

data is permissible. Newspapers and news agencies have proved to be generous sources; many of the best known test collections consist of this type of data; other good sources are non-profit organisations such as governmental and national archives. It is, however, still a significant challenge to obtain similar rights for the use of high-quality multimedia datasets as these often have very strong copyright restrictions. The availability of datasets has a direct impact on what can be evaluated and on future reusability of the test collection created.

Typically, there is no need for much pre-processing of the collections. Well-known standards should be adopted: documents are normally encoded using Unicode (usually UTF-8), XML is mainly used to provide a few basic tags, such as unique identifiers, document headers, and so on. Text is generally kept as close to the original as possible, no attempts are made to correct spelling or formatting errors or similar faults, in order to maintain authenticity.

The collection selected must be appropriate for the search task to be evaluated providing a sample of the kinds of texts that will be encountered in the operational setting of interest. When working in the multilingual context, a good way to create collections that are comparable over language boundaries is to acquire collections on the same topic and in the same genre or for the same time period in different languages, e.g., national newspapers or newswires from different countries for the same years, scientific documents for the same domain, European parliamentary transcripts. Initially, attention was mainly limited to a broad homogeneity of content and style of the collections, and newspaper and news agency documents were commonly used. However, more recently, efforts have been made to create collections that are better simulations of the intended setting of the IR systems to be evaluated. This has resulted in domain-specific tracks being set up by the major evaluation initiatives: the genomics and legal tracks at TREC, patent retrieval tracks at both CLEF and NTCIR, and tracks using social science texts and radiographic images in several languages at CLEF. Table 5.1 provides some statistics for sample collections used in TREC, NTCIR, CLEF and FIRE.

**Query Statements.** In addition to the documents, test collections must include sets of query statements or search requests in some form. These can be derived from the query logs of real world search sessions or can be created artificially as simulations of real information needs from which systems can derive query formulations. In TREC (followed by NTCIR, CLEF and FIRE), this second option has mainly been used. According to the TREC terminology, this type of simulation of a user's information need is known as a 'topic'. Typically TREC-style topics are structured in three fields: a brief title statement; a one sentence description; a more complex narrative. The title contains the main keywords, the description is a natural language expression of the concept conveyed by the keywords and the narrative adds extra syntax and semantics, stipulating the conditions for relevance assessment.

Queries can be constructed by the systems from one or more fields. The motivation behind the use of structured topics is to simulate possible query input for a range of different IR applications, representing keyword style input as well as natural language formulations. The latter potentially allows sophisticated systems

**Table 5.1** Some document collection statistics

| Collection | Language | Genre | No. of docs | Size |
|---|---|---|---|---|
| Cranfield II | English | scientific abstracts | 1,398 | 579 kB |
| TREC-AP | English | newswire | 242,918 | 0.7 GB |
| TREC-GOV2 | English | US government web pages | 25,205,179 | 426 GB |
| NTCIR-4 PATENT | Japanese, English | patent docs | 7,000,000 | 65 GB |
| CLEF-multi | Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish | newspaper and newswire for comparable time period (1994–1995) | 1,869,564 | 4.7 GB |
| EuroGOV (used at CLEF) | 25 European languages | European institutional web pages, for 27 domains | 3,600,000 | 100 GB |
| TEL-multi (used at CLEF) | English, French, German | library catalog records from The European Library | 2,869,553 | 3.8 GB |
| FIRE 2008 | Bengali, Hindi, Marathi, Punjabi, Tamil, Telugu | mainly news docs | 500,000 | 3.2 GB |
| RSNA 2010 (used in ImageCLEF) | English annotations | medical images from Radiographical Society of North America | 75,000 images | 18 GB Images 8 MB captions |

to make use of morphological analysis, parsing, query expansion and similar features.

The use of hand generated query statements or topics with little examination of query logs and user behavior in real world search situations has been criticised as being artificial. However, it is not always easy to obtain query logs from operational search systems and, even when available, their application is problematic. Due to the sparseness of most user queries, it is difficult to identify the underlying information need and, thus, which documents in the collection are really relevant. Jansen and Spink (2006) provide a useful survey of research on understanding user behavior from query logs.

Much of the work in TREC, NTCIR and CLEF has focused on informational queries in the so-called ad hoc tasks,[7] reflecting the roots of information retrieval in library environments. Informational queries tend to be exhaustive and simulate a user looking for all the information they can find on the given argument; this means that there is an unknown number of relevant documents in the target collection and many documents must be assessed in order to identify those relevant. When resources are limited, 'known item' tasks are often used. In this case, the task is to

[7] The term 'ad hoc' (often written 'ad-hoc') reflects the arbitrary subject of the search and its short duration; other typical examples of this type of search are web searches.

find a given, unique document in the collection. The assessor effort is thus concentrated on finding suitable candidate documents and creating appropriate query statements and the relevance assessment stage is eliminated.[8]

The construction of query statements for use in multilingual and cross-language testing scenarios is more challenging than in the monolingual context. The CLEF activity developed its own methodology for this. Working in a distributed setting, with a multilingual collection consisting of a number of separate language-dependent collections, topics are created on the basis of the contents of the collections involved. In order to reduce the workload, the same topics are used for monolingual, bilingual and multilingual tasks.[9] For each separate language, native speakers propose a set of topics covering a mix of events that, hopefully, will be covered by collections in all the languages. The topics are varied to provide a full range of cross-language testing possibilities for the participating systems by including names of locations (translatable or not), of people (where some kind of robust matching may be necessary, e.g., Eltsin, Ieltsin, or Yeltsin), important acronyms, some terminology (testing lexical coverage), syntactic and semantic equivalents. The goal is to achieve a balanced topic set accurately reflecting real world user needs while at the same time testing a system's processing capabilities to the full. This method has been used successfully for mixed language collections of social science documents, of library catalogues, and of news documents.

Once the final topic set has been selected it is prepared in all the collection languages by skilled translators translating into their native language. For example, in CLEF 2002, 50 topics were developed on the basis of a multilingual collection of news documents in eight target languages (Dutch, English, Finnish, French, German, Italian, Spanish, Swedish) and topic sets were produced in each of these languages plus Russian, Portuguese, Chinese and Japanese. An example of the English version of a CLEF 2002 topic is given below:

| | |
|---|---|
| Title: | AI in Latin America |
| Description: | Amnesty International reports on human rights in Latin America |
| Narrative: | Relevant documents should inform readers about Amnesty International reports regarding human rights in Latin America, or on reactions to these reports. |

Topics addressing events mostly covered in a single language often use vocabulary that is difficult to translate directly or invokes subtle cultural differences. Womser-Hacker (2002) gives some examples of such topics used in the early CLEF campaigns: consider the Dutch single word term 'Muisarm' (literally 'mouse arm'), which addresses a form of repetitive strain injury (RSI) linked to

[8] More details on Known Item Retrieval are given in Section 5.4.

[9] The document retrieval track in CLEF typically offers tasks for monolingual retrieval where the topics and the target collection are in the same language, so-called bilingual retrieval where the topics are in a different language from the target collection, and multilingual retrieval where topics in one language are used to find relevant documents in a target collection in a number of different languages.

the use of computer mice. The French translation of the topic containing this term used at CLEF reads 'ordinateur: souris et tensions musculaires' (literally 'computer: mice and muscle strain'), a much more complex paraphrasing of the term in the Dutch original, that may well have implications on how evaluation results are skewed.

A key question when creating a test collection is how many query statements are needed. It is not normally practical to use large numbers of query statements as each one will then require a set of relevance judgments which are costly to produce. The general feeling is that 50 is a good compromise.

This conclusion has been reached after considerable debate in the literature. Spärck Jones and van Rijsbergen (1975) asserted that for an ideal test collection a number of search requests below 75 is of little use, 250 is usually acceptable and 1,000 are sometimes needed. Since then, further studies have suggested that this figure could be considerably reduced. Studies by Voorhees (2000) and Sanderson and Zobel (2005) in the context of TREC propose 50 topics as a reasonable number saying that it is not a good idea to go below 25. In order to build large test collections for multilingual information retrieval tasks, CLEF has produced new topic sets for the same document collections over a number of years.

Voorhees and Buckley (2002) also discuss the effect of topic size on retrieval experiment error. They use TREC results to derive error rates based on the different number of topics used in a test and show that researchers need to take care when concluding one method is better than another when the number of topics used is low. They concluded that an absolute difference in Mean Average Precision (MAP) of 5–6% would be needed between two runs measured on 50 topics before one could be 95% confident that the ordering of systems measured on this topic set would also occur on a different set of 50 topics. Commenting on this result, Sanderson and Zobel (2005) remark that Voorhees and Buckley had not studied the impact of significance tests on error rates and claim that the use of such tests may reduce the difference in MAP required before experimenters can be confident that their result will hold when tested on other topic sets.

In conclusion, for anyone attempting to create their own test collection from scratch, our advice is to start with a minimum set of 50 query statements, eventually expanding the collection with the addition of a second set at a later stage.

**Relevance Assessment.** The concept of relevance is central for IR systems and for their evaluation. Relevance assessment traditionally involves a human assessor reading each document and deciding whether it is pertinent or not to the topic under consideration. This is known as ground truth creation. In TREC and CLEF relevance is normally determined on a binary basis; a document is either considered relevant or not. NTCIR has also experimented with multiple levels of relevance. The criteria that these initiatives use to judge for relevance is normally along the lines of "would any of the information contained in this document be useful for me if writing a report on the topic". In TREC, the person who created a topic is also responsible for the relevance assessment for that topic; this helps to ensure consistency of judgment. This is harder to achieve in the multilingual distributed scenario as it is necessary to ensure consistency in opinion between different assessors

working over language boundaries and on different target collections. The fact that human agreement on a binary relevance judgment is quite modest is one reason for not requiring more fine-grained relevance judging from the assessors.

The inherent subjectivity of assessment, caused by many factors (style and subject matter of the documents, specificity of the topic, experience and preferences of the assessor, etc.) has given rise to criticisms of this methodology. Schamber (1994) states that relevance is a "*. . .multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information need situations*." It is also a "*dynamic concept that depends on users' judgments of the quality of the relationship between information and information need at a certain point in time*." It is hardly surprising that doubts as to the reliability of relevance judgments made by assessors have frequently been raised.

In order to allay such fears, a number of studies have investigated variations in inter-assessor agreement between two or more judges, at varying levels (binary, scalar, weighted, order of documents, etc.), and to what extent this can impact on the results. An early study by Lesk and Salton (1969) using a small test collection of 1,268 abstracts in the field of library science, with 48 queries and binary assessment, investigated the effect of variations in judgments when measuring average precision and recall values for three systems. Agreement levels averaged 0.31 but, despite this low level of agreement, they found no differences in the relative performance of the different systems. Cleverdon (1970) did find some slight differences in the ranks of 19 different indexing methods when using four different sets of relevance judgments but there was a very high correlation between the rankings in each case. These two studies were both on very small test collections where the entire collection was judged for each query, and there were only a small number of relevant documents. Voorhees (2000) describes an investigation of the effect that changes in relevance assessments have on the evaluation of retrieval results for two TREC collections which were many times larger (approximately 800,000 documents each) and where the pooling methodology, in which only a selected sample of the documents in the collection are judged for any one topic, had been applied. The overall conclusion is that differences of opinion can affect the absolute scores quite considerably but in general have little impact on the relative effectiveness ranking of different retrieval strategies. She also stated that a minimum number of topics are required in order to obtain an acceptable correlation – as previously mentioned, her estimate was at least 25 for the collections she used.

A more recent study concluded that the Cranfield method remains fairly robust to variations in relevance judgments although it is important to obtain assessments from judges with a good understanding of the task (Bailey et al. 2008).

**Pooling.** Manual relevance assessment is a laborious process. As the size of the test collections grew, it rapidly became impossible to judge all documents exhaustively for each query. The pooling system, first proposed by Spärk Jones and van Rijsbergen (1975), was thus introduced in order to reduce the assessment load to manageable proportions.

Using the pooling strategy, only a fraction of the document collection is assessed for any given topic. The aim is to create a small subset of documents that will

hopefully hold the majority of the relevant documents for that topic in a collection. It operates as follows: for each query the top $k$ results of the rankings of the runs submitted by different search systems or retrieval techniques are merged in a pool with the duplicates removed. This implies eliminating all those documents from consideration that were not retrieved by any participant with a high rank in their list of results. The reasoning behind this strategy is discussed in detail in Voorhees (2002), and boils down to a decreasing probability of 'missing' a relevant document as the number of different retrieval systems contributing to the pool increases. The document pool thus created is then judged for relevance. The number of result sets per participant that are used for pooling, and the establishing of the separation between the top $k$ or the 'highly ranked' and other documents (the so-called 'pool depth') are dictated to some extent by practical needs (i.e., available resources for assessment). It is considered important that the pool contains runs coming from systems using different techniques. The pool can be 'strengthened' in various ways, for example, by the addition of some manual runs or pilot searches, or by including all documents marked as relevant by assessors during the topic creation stage in order to ensure that as many relevant documents as possible are included. When operating in a multilingual context, the situation is complicated by the need to render the pool sufficiently exhaustive for all the target collections and languages involved.

An important limitation when forming the pools is the number of documents to be assessed. Although initially a depth of at least 100 results per run was considered to be advisable, tests made on NTCIR pools in recent years have suggested that a depth of 60 is normally adequate to create stable pools as long as a sufficient number of runs from different systems are included. However, there is no hard and fast agreement as to what is a sufficient number and how many different systems/retrieval strategies need to be involved to create a stable pool. The question as to how many documents should be assessed appears to be linked to the number of queries available. Whether a 'wide and shallow' approach of minimally examining many queries is better than a deep and narrow examination of a small number of queries is discussed by Sanderson and Zobel (2005) with respect to the significance of results and the conclusion is that assessor effort would be better spent building test collections with more topics, each assessed in less detail. However, the number of judgments to be made is also affected by the evaluation measure used. For example, if the focus is on the average number of relevant documents in the top ten ranked (known as precision at 10) then only the first ten results of each run need to be judged.

The central importance of relevance assessment for the calculation of many popular evaluation measures means that the pooling methodology is not without critics. Concerns in this respect focus mostly on the coverage of the assessments. The problem of coverage arises from the fact that only the highly ranked documents included in selected results sets are judged for relevance. This implies that some relevant documents may go undetected if they have not been retrieved by any of the runs included in the pool. The original hypothesis is that a sufficient number of diverse retrieval strategies will find most of the relevant documents in their upper

ranks and that figures calculated on the basis of these limited assessments are a good approximation of theoretical figures based on complete assessments. This question of pool coverage has been studied by Zobel (1998) for early TREC collections and more recently, over a number of years, for TREC, NTCIR and CLEF by Tomlinson with similar findings to those of Zobel (Tomlinson 2010). The consensus is that, overall, collections created using pooling do provide reliable results (Zobel 1998, Harman 2005). It is also becoming common practice to supplement the pools with the addition of the results of some manual interactive searches, see Section 5.2.5 below.

A key question is the reusability of these test collections for the evaluation of a system that did not contribute to the pool of judged documents. The worry is that if such a system retrieves a substantial number of relevant documents that had not been included in the pool, it will be unfairly penalised when calculating the evaluation measures. This issue has also been studied by several groups. Braschler (2004a) reports an investigation on the CLEF 2003 multilingual collection using the technique proposed by Zobel (1998). The quality of the document pool is judged by the mean performance difference of a system that did not contribute runs to it. In order to measure this, the relevant documents found uniquely by a given group are removed and then the results for this group are recomputed – thus simulating the scenario that this group had not participated in the campaign. The smaller the change in performance observed, the higher the probability that the relevance assessments are sufficiently complete. Using this method, Braschler found that the pools used for the multilingual tasks in CLEF 2003 were sufficiently stable. The maximum change in performance scores was 0.77% for Multilingual-8 and 2.06% for Multilingual-4.[10] These small differences only influence direct comparisons between systems that have practically identical performances and where the original performance differences cannot be considered statistically significant. However, Voorhees (2006) reports that this 'leave-out-uniques' test can fail to indicate a problem with a collection if all the runs that contribute to the pool share a common bias – preventing a common bias is why a diverse run set is needed for pool construction. When pools are shallow with respect to the number of documents in the collection, the sheer number of documents of a certain type will fill up the pools to the exclusion of other types of documents. To produce an unbiased, reusable collection, traditional pooling with a limited number of topics requires sufficiently deep pools. However, as stated above, there is as yet no clear consensus as to what 'sufficiently deep' really means in terms of numbers.

**Evaluation Protocol.** The final component when using a test collection is the evaluation protocol to be adopted. A number of different measures have been proposed; these are discussed below in Section 5.2.6.

[10] Two multilingual tasks were offered in CLEF 2003; participants could test their system using queries in one language (from a choice of 12 languages) against multilingual collections in either four or eight languages.

### 5.2.4 *Promoting Research into Multilingual and Multimedia System Development via Evaluation*

The discussion so far has focused mainly on evaluation for textual document retrieval, ignoring questions regarding other media. However, in recent years, attention has also been given to creating benchmarks to evaluate multimedia systems (Smith 1998, Leung and Ip 2000, Müller et al. 2001), and a number of them have been used within comparative evaluation campaigns. Benchathlon[11] was the first large-scale event of this type that provided evaluation resources and promoted discussions throughout the multimedia research community. Subsequent events followed including TRECVid,[12] sponsored by TREC, and ImageEval,[13] financed by the French research foundation, which address different aspects of visual information retrieval evaluation. These initiatives have focused on image and video retrieval evaluation without paying any particular attention to language-dependent aspects.

CLEF has also included evaluation tasks for media other than text; the objective being to stimulate the development of multilingual search systems for any media. Underlying this is the strong conviction that IR indexing and matching procedures need to be tuned according to the languages for which a system is intended and also that today's systems increasingly include more than one media. In this section we briefly summarise the work at CLEF and report the main results.

Over the years the CLEF programme has expanded to include activities to test systems for various types of text retrieval, e.g., geographic IR, question answering, information filtering, on diverse genres, such as news, library catalogues, parliamentary archives, patents, and for other media, i.e., image, speech and video. Most of the activity has been on system-oriented experimentation but some studies of user behaviour in the multilingual context have also been conducted within tracks for cross-language interactive retrieval and for multilingual log analysis. For each new track, target collections and query statements have been provided, covering a wide range of languages. This activity has resulted in the creation of a large number of important and reusable test collections and important advances in the state-of-the-art.

**CLEF and Text Retrieval.** The Ad Hoc track for multilingual textual document retrieval was the core track of CLEF for the first 10 years. Different tasks were offered over the years presenting varying degrees of difficulty. The more basic tasks were designed to encourage inexperienced groups to experiment and increase their expertise, more challenging tasks invited researchers to address difficult issues and discover innovative solutions. It is probably true to say that this track has done much to foster the creation of a strong European research community in the CLIR area. It provided the resources, the test collections and also the forum for discussion

[11] http://www.benchathlon.net/

[12] http://www-nlpir.nist.gov/projects/trecvid/

[13] http://www.imageval.org/

and comparison of ideas and results. Groups submitting experiments over several years demonstrated flexibility in advancing to more complex tasks. Much work was done on fine-tuning for individual languages while other efforts concentrated on developing language-independent strategies. There is substantial proof of significant increase in retrieval effectiveness in multilingual settings by the systems of CLEF participants. Braschler (2004b) provides a comparison between effectiveness scores from the 1997 TREC-6 campaign and the CLEF 2003 campaign in which retrieval tasks were offered for eight European languages. While in 1997 systems were performing at about 50–60% of monolingual effectiveness for multilingual settings, that figure had risen to 80–85% by 2003 for languages that had been part of multiple evaluation campaigns.[14] In recent campaigns, we commonly see a figure of about 85–90% for most languages. CLEF participants tend to learn from each other and build up a collective know-how. The best systems tend to be a result of careful tuning of every component, and of a combination of different algorithms and information sources. Other text retrieval tasks have included cross-language question answering, cross-language geographic retrieval and cross-language patent retrieval.

**CLEF and Multimedia Retrieval.** CLEF tracks have also been set up with the objective of stimulating the development of strategies for multilingual information access in media other than text, i.e., speech, image and video. The ImageCLEF track has investigated the effectiveness of combining textual and visual features for cross-language image retrieval and building resources for benchmarking image retrieval systems. Retrieval from an image collection offers distinct characteristics and challenges with respect to one in which the document to be retrieved is text. For example, the way in which a query is formulated, the methods used for retrieval (e.g., based on low-level features derived from an image, or based on associated textual information such as a caption), the types of query, how relevance is assessed, the involvement of the user during the search process, and fundamental cognitive differences between the interpretation of visual versus textual media. For cross-language image retrieval the problem is further complicated (Clough et al. 2010b). The experiments conducted within ImageCLEF have done much to advance research in this sector. In particular, the medical image retrieval group has worked closely with medical professionals to provide realistic search tasks so that the effectiveness of systems in an operational setting can be estimated. Full details of ImageCLEF experiments and results can be found in Müller et al. (2010). Other tracks in CLEF have aimed at encouraging research in multilingual speech and video retrieval.

**Main Results.** CLEF has thus offered different entry points to the fields of MLIR and CLIR in order to support the creation and growth of a research community with diversified expertise. Research and development has been promoted through the exploration of a comprehensive set of MLIR-related issues which has

[14] A common way to measure the effectiveness of a CLIR system is to compare performance against a monolingual base-line.

encouraged component evaluation and has led to the development of specialised resources, tools and know-how covering the following areas:

- Language resources: the development and/or testing of language resources, such as stopword lists, dictionaries, lexicons, aligned and parallel corpora, etc., has been supported;
- Linguistic components: the development and/or evaluation of linguistic tools, such as stemmers, lemmatizers, decompounders, part of speech taggers, and so on, has been fostered;
- Translation approaches: different strategies for crossing language barriers, such as Machine Translation (MT), and dictionary-based, corpora-based, or conceptual network-based translation mechanisms have been tried;
- IR models: different models have been applied – Boolean, vector space, probabilistic, language models, and so on – to improve retrieval performances across languages;
- Advanced IR approaches: techniques, such as data fusion and merging or relevance feedback, have been adopted to address issues such as the need for query expansion to improve translation or the fusion of multilingual results;
- Interface issues: user requirements in the multilingual context have been investigated with respect to assistance in query formulation and results presentation;
- Evaluation measures: measures and metrics to analyse system behaviour in a multilingual setting and compare performances across languages and tasks have been developed and employed.

CLEF is a valuable source for test collections suitable for multilingual and multimedia system evaluation mainly for European languages. Other important sources are TREC for Arabic, NTCIR for Chinese, Japanese and Korean, and FIRE for Indian languages such as Bangla, Hindi, Marathi, Punjabi, Tamil and Telugu.

### *5.2.5 Alternative Methodologies for Test Collection Construction*

As document collections are becoming ever larger, many studies have focused on proposing ways to create large-scale test collections with limited costs while still guaranteeing reliable and robust evaluation results. The focus is mainly on reducing the effort required to produce the relevance assessments. Cormack et al. (1998) examined two ways to create effective test collections with greatly reduced judging effort: Move-to-Front (MTF) Pooling and Interactive Searching and Judging (ISJ).

Move-to-Front Pooling improves on standard pooling by using a variable number of documents from each retrieval system depending on their performance. For each run submitted to the pool, documents are assessed in their ranked order, under the assumption that the more relevant the document the higher its rank. Submissions that appear to locate more relevant documents for a particular query are assessed in more depth, while fewer documents are judged for the others.

Cormack et al. (1998) tested this approach on the TREC-6 test collection, building a set of relevance assessments judging only half the number of documents examined by the TREC assessors. They used this set of assessments to measure the Mean Average Precision (MAP) of each system that participated in TREC-6 and ranked them. They then repeated the process using the full set of relevance judgments created by the TREC assessors. The two sets of system rankings were correlated using Kendall's tau. The correlation found was 0.999. When repeating the ranking using just one tenth of the MTF pool the correlation was 0.990 which is still perfectly acceptable.

The objective of the Interactive Searching and Judging (ISJ) approach is to create a reasonable pool of judged documents with minimal effort. A group of searchers is requested to find and judge as many relevant documents as possible in a brief period of time. These judgments can be created independently and ahead of the participants' runs. In judging the documents, it is suggested that the searchers place them into three categories: relevant, not relevant, and borderline. Queries are constructed manually to find potentially relevant documents. Documents are judged until it is felt that it is unlikely that further relevant documents will be found. For most topics, several queries are submitted to investigate different aspects of the topic. In this way, the manual effort can be reduced considerably. The effectiveness of this method was assessed against the TREC-6 test collection. It was used to create an independent set of relevance judgments. Although the number of relevant documents contained in this set was similar to that of the official set, only one-quarter as many documents were judged. Despite the considerable reduction in effort, collection effectiveness did not appear to be compromised.

Since it was first proposed, ISJ has become quite popular as a way to facilitate test collection construction. There is, however, the risk that the relevance judgments may be biased by the search system used to create them. For this reason it has often been applied as a supplementary method to improve the quality of test collections created using the traditional pooling approach. For example, manual interactive searches have been used in selected tasks in ImageCLEF in order to supplement automatically created shallow pools. In these cases, ISJ found relevant images that the standard pooling had missed (Clough et al. 2004). The organisers of NTCIR also perform their own manual runs to supplement their pools (Kando et al. 1999).

Recently, there has been growing interest in techniques for low cost (in terms of judgment effort) evaluation. One of the approaches proposed is 'crowd-sourcing'.[15] A very simple form this approach has been employed in several of the CLEF multilingual retrieval tracks where no funding was available to support ground truth creation. Groups participating in the tracks were thus requested to produce search requests and perform part of the relevance assessment work. Alonso et al. (2008) describe how this approach can be used to distribute relevance judging tasks to a large group of potential assessors on the Internet. They used the Amazon

[15] This term was coined to describe tasks that were outsourced to a large group of people rather than being performed by an individual on-site (Howe 2006).

Mechanical Turk (MTurk) service which is part of the Amazon web services platform. The web services API is used to submit tasks to the MTurk website. People come to the site to search for and complete tasks and receive payment. These authors state that they uploaded an experiment requiring thousands of judgments and it was completed in just 2 days, far more quickly than they could have hoped to have it performed on-site. The cost was very low and this made it possible to get both many and multiple judgments.

Of course there is a strong down-side to this approach. There was already concern in CLEF with respect to the quality of the assessment work when performed by a large group of people, with different levels of competence, over which it was not easy to maintain tight control or quality checks. However, by involving campaign participants in ground truth creation you are using a group of people that have a strong vested interest in providing a high quality outcome. The problem of quality control is much exasperated when using an apparently random collection of strangers. How do you know that they have sufficient competence to perform the task? How do you know that they will perform it to the best of their ability? The first issue is addressed by a qualification procedure included in MTurk. Workers must complete an appropriate qualification task before being accepted. In order to address the second problem, Alonso et al. proposed to have more than one person judge each query-document pair and then aggregate the scores in some way. They assert that depending on the number of relevance judgments obtained and the strictness of the aggregation scheme adopted, it should be possible to guarantee a high level of quality control.

### *5.2.6 Performance Measures*

Many different measures for evaluating the performance of information retrieval systems on the basis of test collections have been proposed. All common measures described here assume the notion of relevancy as described above: every document is known to be either relevant or non-relevant to a particular query.[16] When evaluating multilingual or cross-language retrieval runs, essentially the same measures can be used as in the monolingual case.

The subjectivity of relevance and the variations in representation of the same information mean that it is not possible to simply validate the correct implementation of the retrieval algorithm (e.g., 'does the system return all documents containing all search terms?'). Different indexing and matching strategies, and in the multilingual case different translation strategies, will lead to different subsets of the relevant documents being retrieved. Two main criteria for assessing the performance of the information retrieval system are effectiveness (does the system

[16] All these measures are only defined for the cases when there is at least one relevant document; it is meaningless to assess ranking performance when no document is relevant to the query.

retrieve the 'right' documents?) and efficiency (does the system retrieve the documents the 'right' way, i.e., quickly and using few resources?). Our discussion of performance measures will be restricted to measures for retrieval effectiveness, as the measurement of retrieval efficiency does not present many IR specific peculiarities.

**Precision and Recall.** As already stated in Section 2.2.2, these are the basic measures for quantifying the effectiveness of an IR system. The two measures are based on two assumptions: that the user wants to:

1. Retrieve as many relevant items as possible, and
2. Retrieve as few irrelevant items as possible.

They are set-based measures: they operate on a set of results and ignore its ordering.

To compute precision and recall for a set of documents, the following definitions are used:

Precision = number of relevant documents in the set/total number of documents in the set;

Recall = number of relevant documents in the set/total number of relevant documents in the collection.

Perfect precision is achieved when only relevant items are contained in the set. This can obviously be the case irrespective of the recall associated with the same set. Perfect recall is achieved when all the relevant items are contained in the set. As this condition is independent of the size of the set, it is possible to achieve perfect recall by simply returning the whole collection of items for every search request. Obviously, such a retrieval strategy would lead to very low precision.

As the above considerations with respect to perfect precision and recall indicate, there is usually an inverse relationship between the two measures. Improved precision is attained by decreasing the number of irrelevant items returned. This implies a conservative indexing and matching strategy: only the most exact matches are returned. As a consequence, relevant items that require additional effort for retrieval are no longer returned. The recall drops. Improving recall, on the other hand, dictates the use of mechanisms that increase the overall number of matches, returning items based on looser criteria. Tools such as aggressive stemming and query expansion are often used. Consequently, the number of irrelevant items in the retrieved set often increases, and the precision decreases.

Often more weight is given to one measure over the other; for example in the case of web search the focus is typically on obtaining high precision by finding a small number of highly relevant documents at the top of the ranking. However, there are domains, such as patent search, where the focus is on finding all relevant documents through an exhaustive search (high recall).

The two measures need to be adapted for use with the ranked list output produced by an IR system. To this end, the measures can be computed continuously after each rank, starting with the top-ranked document (set of size 1), and then expanding the size of the set by one document in every step for the computation. This way, a pair of (precision, recall)-values is obtained at each

rank. Using a two-dimensional graph, where the y-axis denotes the precision and the x-axis denotes the recall, these values can be visualised. Such visualisation is typically referred to as a Precision/Recall (or Recall/Precision) graph.

Formally, to compute a precision/recall graph, we proceed as outlined in the following. For a summary of the notation used, refer to Table 5.2.[17]

Given a set of documents $D$, a set of test queries $Q$, and for every query $q_i$, a set of relevance assessments identifying the set of relevant documents $D^{rel}(q_i)$[18] and non relevant documents $D^{non}(q_i)$ for that query,[19] let RSV be a function that assigns retrieval status values (scores) to documents with respect to a query (as determined by the system's indexing method and weighting scheme). For every query $q_i \in Q$, the documents $d_j \in D$ are ranked in decreasing order of the values $\mathrm{RSV}(q_i, d_j)$. A recall and precision graph shows how many relevant and how few irrelevant documents are contained in the top ranked documents. Typically, users will inspect the documents in the order they are presented. At any given time, the first $r$ documents of the ranked list have been inspected. This set of $r$ documents is the answer set $D_r(q_i)$. The value for $r$ is determined by the user's preferences. Choosing a low value for $r$ implies that the user is interested in few, high-precision

**Table 5.2** Notation used for discussion of performance measures

| |
|---|
| $D$: set of documents |
| $Q$: set of queries |
| $q_i$: single query ($q_i \in Q$) |
| $d_j$: single document ($d_j \in D$) |
| $D^{rel}(q_i)$: set of documents that are relevant with respect to query $q_i$ ($D^{rel}(q_i) \subset D$) |
| $D^{non}(q_i)$: set of documents that are non-relevant with respect to query $q_i$ ($D^{rel}(q_i) \cup D^{non}(q_i) = D$) |
| $\mathrm{RSV}(q_i, d_j)$: retrieval status value of document $d_j$ with respect to query $q_i$, as determined by a given retrieval method; used for ranking |
| $D_r(q_i)$: set of top $r$ ranked documents from $D$ with respect to query $q_i$ |
| $D_r^{rel}(q_i)$: set of relevant documents in the top $r$ ranked documents from $D$ with respect to query $q_i$ ($D_r^{rel}(q_i) := D^{rel}(q_i) \cap D_r(q_i)$) |
| $\rho_r(q_i)$: recall value for set $D_r(q_i)$ |
| $\pi_r(q_i)$: precision value for set $D_r(q_i)$ |
| $\Pi_i(\rho)$: interpolated precision value at recall level $\rho$ for query $q_i$ |
| $\Pi(\rho)$: mean of interpolated precision values at recall level $\rho$ for all queries $q_i \in Q$ |
| $\mathrm{AP}_i$: average precision for query $q_i$ |
| MAP: mean average precision over all queries $q_i \in Q$ |
| GMAP: geometric mean average precision over all queries $q_i$ |

[17] The following paragraphs draw extensively from the description in (Schäuble 1997). We thank Peter Schäuble for his kind permission to base our description on his work.

[18] Assumed to be non-empty.

[19] In practice, $D^{non}(q_i)$ is rarely determined by an exhaustive relevance assessment process. After an approximation of $D^{rel}(q_i)$ is obtained using the pooling method, it is instead assumed that $D^{non}(q_i) = D \backslash D^{rel}(q_i)$

documents, whereas a high value for $r$ means that the user intends to conduct an exhaustive search.

For each answer set $D_r(q_i)$, a pair of recall and precision values is computed:

$$\text{Recall } \rho_r(q_i) := \frac{|D_r^{rel}(q_i)|}{|D^{rel}(q_i)|} \quad \text{and} \quad \text{Precision } \pi_r(q_i) := \frac{|D_r^{rel}(q_i)|}{|D_r(q_i)|}.$$

where $D_r^{rel}(q_i) := D^{rel}(q_i) \cap D_r(q_i)$ is the set of relevant documents contained in the answer set. Table 5.3 shows recall and precision values for the top ten ranks of a sample query. In order to be able to compute the recall values, we assume that for this query $|D^{rel}(q_i)| = 5$.

A plot of the corresponding precision and recall values results in a saw tooth curve (see solid line in Figure 5.1).

**Table 5.3** Precision/recall figures for a sample query $q_i$ and corresponding relevance assessments

| Rank $r$ | Relevant to $q_i$ | $\rho_r(q_i)$ | $\pi_r(q_i)$ |
|---|---|---|---|
| 1 | Yes | 0.20 | 1.00 |
| 2 | No | 0.20 | 0.50 |
| 3 | No | 0.20 | 0.33 |
| 4 | Yes | 0.40 | 0.50 |
| 5 | No | 0.40 | 0.40 |
| 6 | No | 0.40 | 0.33 |
| 7 | Yes | 0.60 | 0.43 |
| 8 | Yes | 0.80 | 0.50 |
| 9 | No | 0.80 | 0.44 |
| 10 | Yes | 1.00 | 0.50 |

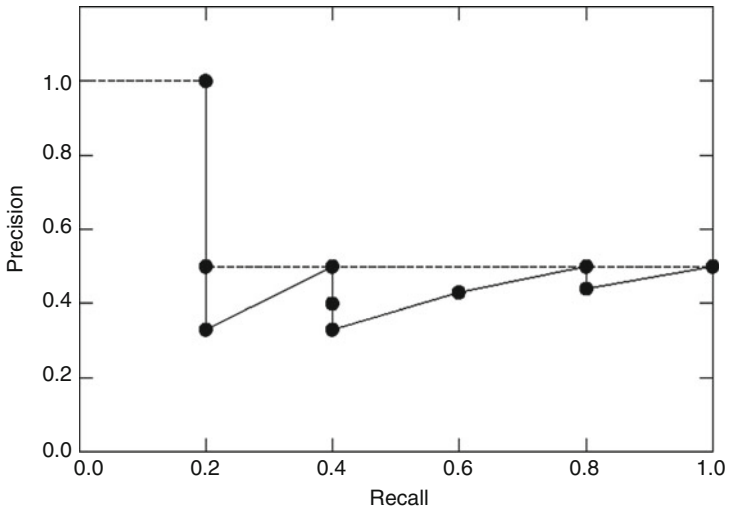


**Fig. 5.1** Interpolation of the recall/precision values of Table 5.3

Precision is commonly used as an effectiveness measure for IR system output by choosing result sets of fixed size (Precision @ $n(q_i) := \pi_n(q_i)$, or Mean Precision @ $n := \frac{1}{|Q|} \sum_{q_i \in Q} \pi_n(q_i)$ for a set of queries $Q$): most commonly, the top 10, 20 or 100 documents are used for computation (Precision @ 10, Precision @ 20, and Precision @ 100 measures, respectively). The first two alternatives are chosen to model the typical user of web search services, who has been shown to rarely look past the first page of results (which typically contains 10, and at most 20, results). Note that the Precision @ $n$ measure is not able to differentiate between two systems that retrieve the same number of relevant documents in the top $n$ ranks for a query, even if one system returns the relevant documents at higher ranks than the other. For example, consider two systems returning a single relevant document each. Both systems would be assigned Precision @ $10 = 0.1$, even if System One returns the relevant match at rank 1, and System Two returns it at rank 10. This can be a serious drawback of the measure. Furthermore, there are issues for small values of $n$ when a query has a large number of relevant documents, i.e., if $\left|D^{rel}(q_i)\right| \gg n$. For such queries, it is in practice often hard to separate systems that only retrieve few of the relevant documents ('the easy matches') at high ranks from systems that achieve high recall, with optimal scores being returned in both cases. While Precision @ $n$ is a frequently cited measure, the analogous Recall @ $n$ measure is used comparatively rarely.

The 'saw tooth' curves obtained through the computation outlined above are typically replaced by monotonically decreasing curves, thus assigning each recall value a single precision value. To this end, in the next step of computation, for every query $q_i$, a function $\Pi_i$ is defined that assigns a precision value for every recall value $\rho \in [0,1]$ as follows:

$$\Pi_i(\rho) := \max\{\pi_r(q_i) | \rho_r(q_i) \geq \rho\}$$

Using this 'interpolation step', we obtain the desired monotonically decreasing curve. The ceiling interpolation $\Pi_i(\rho)$ shown by the dotted line in Figure 5.1 looks at locally optimal answer sets for which recall and precision cannot be improved simultaneously by inspecting further documents (Schäuble 1997).

When evaluating a system with a set of queries, an averaging step is introduced that produces the final recall/precision curve:

$$\Pi(\rho) := \frac{1}{|Q|} \sum_{q_i \in Q} \Pi_i(\rho)$$

where $|Q|$ denotes the number of queries.

Typically, to plot a graph, the function $\Pi(\rho)$ is evaluated at eleven levels $\rho = 0.0, 0.1, \ldots, 1.0$.

Figure 5.2 shows the recall and precision graph for the top five groups in the CLEF 2000 Italian monolingual retrieval ad hoc task. For this recall and precision

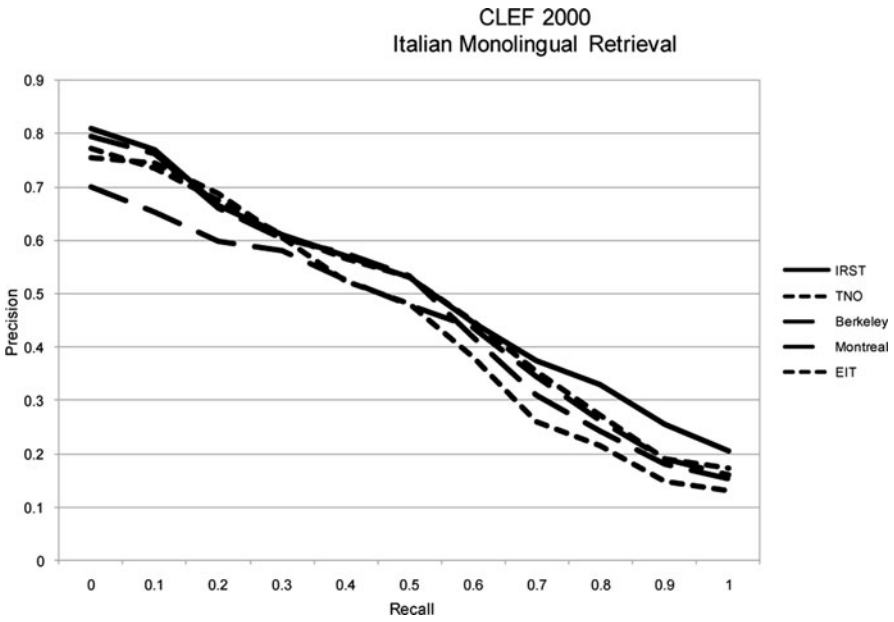


**Fig. 5.2** Top five systems for the CLEF 2000 Italian monolingual retrieval ad hoc task (as determined by mean average precision). This figure is adapted from (Peters 2000)

graph, the function $\Pi(\rho)$ was calculated and evaluated for eleven recall values as described above. A precision/recall graph enables the reader to compare at different recall levels. Indeed, in this specific case, the figure nicely illustrates how no single system delivers best performance for all recall levels.

In cases where a single value effectiveness measure is required average precision is often used. There are competing definitions for this value:

In its simplest form, average precision of a query $q_i$ is the average of three interpolated precision values, which loosely represent a precision-oriented, a 'balanced', and a recall-oriented user.

$$\mathrm{AP}_i := \frac{1}{3}(\ \Pi_i(0.25) + \ \Pi_i(0.50) + \ \Pi_i(0.75))$$

Based on a precision/recall graph, average precision can alternatively be calculated over 11 levels of interpolated precision.

$$\mathrm{AP}_i := \frac{1}{11}\left(\sum_{\rho=0.0\ldots1.0} \Pi_i(\rho)\right)$$

To avoid problems with interpolation, average precision is most commonly calculated today as the average of the exact precision values as determined after

each relevant document is inspected. For relevant documents that are not retrieved at all, a precision of 0 is assumed; i.e., the calculation is as follows:

$$\mathrm{AP}_i := \frac{1}{|D^{rel}(q_i)|} \sum_{r=1}^{|D|} \rho_r(q_i) * \mathrm{rel}(r)$$

where $\mathrm{rel}(r) = 1$ if the document at rank $r$ is relevant to query $q_i$, and $\mathrm{rel}(r) = 0$ otherwise.

The Mean Average Precision (MAP) is a very frequently cited one-value measure, and is defined as the mean of the average precision values for each query $q_i$ in the set of queries $Q$.

$$\mathrm{MAP} := \frac{1}{|Q|} \left( \sum_{q_i \in Q} \mathrm{AP}_i \right)$$

MAP has become one of the primary measures used in many evaluation exercises as well as a large quantity of published IR research. Often people prefer single value measures such as MAP to more complex performance indicators, e.g., the recall/precision graph. The advantage of single value measures lies in easy comparison, their danger in too much abstraction: Mean Average Precision (MAP) provides a single-figure measure of precision across recall levels and *multiple queries*, i.e., MAP is the arithmetic mean of average precision values for individual queries ('a mean of averages'). If relying exclusively on a single value such as MAP, the ability to judge a system's effectiveness for different user preferences, such as exhaustive search or high-precision results, or for individual queries, is lost.

**R-Precision.** Due to the problems inherent with the Precison @ $n$ measure discussed above, 'R-Precision' is sometimes used as a substitute. The measure is defined as:

$\text{R-Precision}(q_i) = \text{Precision @ } R_i(q_i) = \pi_{R_i}(q_i) = \frac{1}{R_i} \left| D^{rel}_{R_i}(q_i) \right|,$
which is actually equivalent to
$\text{Recall @ } R_i(q_i) = \rho_{R_i}(q_i) = \frac{1}{R_i} \left| D^{rel}_{R_i}(q_i) \right|.$

The value $R_i$ is defined as $R_i = \left| D^{rel}(q_i) \right|$, i.e., the total number of relevant documents for query $q_i$. Notably, for sets of size $R_i$, precision and recall values are equal, as demonstrated above. In contrast to the Precison @ $n$ measure, when averaging over queries, the number of ranks $R_i$ that is used for determining the R-Precision values typically differs from query to query. R-Precision has been shown to correlate highly with average precision (Aslam et al. 2005).

**GMAP.** As stated before, when evaluating multilingual or cross-language retrieval runs, essentially the same measures can be used as in the monolingual case. However, work by Mandl (2009) indicates that some measures may exhibit different behaviour for multilingual settings when compared to using them for monolingual runs. Specifically, in the multilingual case, MAP calculated over a set of topics tends to be dominated by the performance that systems obtain on the

'easiest' queries. This same behavior was not observed in related monolingual experiments. In cases where the 'lower end' of the average precision scale is of interest, i.e., where it is desirable that badly performing queries impact the evaluation measure, Geometric Mean Average Precision (GMAP) is a possible alternative (Robertson 2006).

The definition of GMAP is as follows:

$$\text{GMAP} := \sqrt[|Q|]{\prod_{q_i \in Q} \text{AP}_i} = \exp\left(\frac{1}{|Q|} \sum_{q_i \in Q} \log(\text{AP}_i)\right)$$

Since the average precision for some queries may be zero in cases when no relevant documents are retrieved,[20] such queries are assigned a small value ε as their average precision (e.g., $\varepsilon = 10^{-5}$).

Simplifying things, it can be noted that when using Mean Average Precision, a change in average precision from 0.3 to 0.35 will have the same effect on the overall MAP value as does a change from 0.05 to 0.1 (a gain of 0.05 in both cases), whereas when using GMAP, an AP change from 0.3 to 0.6 has the same effect as a change from 0.05 to 0.1 (a doubling of AP in both cases). In the case of evaluating multilingual or cross-language runs, it has been shown that using GMAP, topic difficulty has indeed much less influence on overall system comparison. These observations can be taken as strong indications of the benefits of using *both* MAP and GMAP when 'robustness', i.e., solid retrieval effectiveness even for hard topics, is a concern. In any case, it may be beneficial to consider both measures when many languages are used for multilingual retrieval: in such cases, the correlation between the two measures is very low (Mandl et al. 2008), in contrast to usually high correlations for monolingual and bilingual experiments.

**Mean Reciprocal Rank**. Mean Reciprocal Rank is most commonly used to measure the success of search tasks where just one relevant document is required. Such tasks include known item retrieval, question answering and navigational search. The value is calculated as the reciprocal of the rank at which the first relevant document is retrieved. For example, if the first relevant document is found at rank position 2, the reciprocal rank score is $1/2 = 0.5$. Mean Reciprocal Rank (MRR) is the average score of the reciprocal ranks across a set of queries $Q$.

**Graded Relevance Measures.** The measures discussed above only differentiate between relevant and non-relevant documents ('binary judgments'). If a finer grading of relevance is desired, the Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (nDCG) measures originally proposed by Järvelin and Kekäläinen (2000) may be used. Assuming that grades of relevance

[20] Note that theoretically this is not an issue if systems produce exhaustive result lists that rank every item in the document collection; in such cases, inevitably, under the assumption that every query has at least one relevant document, the relevant documents will turn up at some point in the ranking, leading to an average precision which may be very small, but is greater than zero.

can be transformed to numerical values (by defining a function rel($i$) for the relevance of the $i$-th document in the ranking, e.g., rel($i$) = 2 for highly relevant documents; rel($i$) = 1 for partially relevant documents; and rel($i$) = 0 for non-relevant documents), the DCG measure for the top $n$ ranked documents is defined as follows:

$$\mathrm{DCG}(n) := \sum_{i=1}^{n} \frac{2^{\mathrm{rel}(i)-1}}{\log_2(i+1)}$$

An arbitrary number of levels of relevance can be used with this definition. Broadly speaking, the measure assesses the usefulness or gain from examining a document. DCG discounts the relevance values progressively using a log-based discount function to simulate users valuing highly ranked relevant documents over the lower ranked ones.

DCG numbers are averaged across a set of queries at specific rank values. For the Normalized Discounted Cumulative Gain (nDCG) measure, DCG values are normalised against an ideal ordering of the relevant documents (i.e., ordered by their rel($i$) value). This normalisation step makes averaging across queries with different numbers of relevant documents easier.

$$\mathrm{nDCG}(n) := \mathrm{DCG}(n)/\mathrm{optimalDCG}(n)$$

**Bpref.** The measures discussed so far were all originally designed with completeness of relevance judgments in mind – i.e., it is assumed that every document is judged with respect to every query. As discussed in Section 5.2.3 above, complete relevance judgments are no longer practical as test collections grow to include millions of documents. A pooling strategy is instead adopted, and only a small portion of the collection, which is believed to contain most relevant documents, is in practice assessed for each query. All the documents that remain unjudged, i.e., the vast majority of documents, are *assumed* to be not relevant. Buckley and Voorhees (2004) have analysed MAP, Precision @ 10 and R-precision, and have concluded that these measures are not robust to substantially incomplete relevance judgments. Instead, a new measure, 'bpref' (binary preference) is proposed. The basic idea is to use a notion of 'preference' by the user, i.e., the user is assumed to prefer any (judged) relevant document over any (judged) non-relevant document. Only those documents in a result list to be evaluated that have judgments are thus used for the calculation of the measure. The measure is defined as:

$$\mathrm{bpref} := \frac{1}{R} \sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

where the query being evaluated has $R$ (known) relevant documents, $r$ is a relevant document, and $n$ is a member of the first $R$ judged non-relevant documents as retrieved by the retrieval system. While more robust towards incomplete

judgments, bpref has been found to correlate highly with MAP when complete relevance judgments are available.

### 5.2.7 Statistical Significance Testing

It is common to find in practice that retrieval results (often called *runs*) are compared against each other (comparative evaluation) in academic literature; with the runs with the highest scores deemed as the 'best'. Note, however, that a high variability between retrieval effectiveness for different queries has often been observed, a variability that has been found to be generally higher than the variability between different systems or different well-tuned weighting schemes. It is thus important to consider whether the observed differences in effectiveness are in fact due to real differences between runs or generated by random chance (i.e., by picking a few non-representative queries that happen to perform better on one system). A common approach to gain a fuller understanding is to use one or more statistical significance tests. These tests estimate the probability $p$ of observing the data being examined given that a so-called null hypothesis is true, i.e., $P(Data|H_0)$. In the context of retrieval experiments, the null hypothesis states that the systems producing the two runs under examination have effectively the same retrieval characteristics and that any difference between the runs occurred by random chance. If $p$ is found to be below a threshold – typically either 0.05 or 0.01 – it is concluded that the null hypothesis is unlikely and consequently should be rejected.[21] In information retrieval research, three tests of statistical significance are most commonly used: the Student's paired *t*-test, the Wilcoxon signed rank test, and the sign test (Hull 1993). While the Student's *t*-test is an example of a parametric test, the latter two options are non-parametric tests.

These tests are not infallible and can make errors, which have been classified into Type I and Type II errors (Sanderson and Zobel 2005). Type I errors are false positives: leading to incorrectly rejecting *the null hypothesis*; Type II errors are false negatives: leading to incorrectly concluding that there is insufficient evidence to reject the null hypothesis. In a sense, Type I errors measure the precision of the test, Type II measure its recall. Different significance tests tend to produce a different balance between these two errors. For example, the sign test is known for its high number of Type II errors whereas the *t*-test is known for producing Type I.

It is important to consider the assumptions underlying the different tests. Non-parametric tests, such as the sign test and the Wilcoxon signed ranked test make few assumptions with respect to the data being tested, and are thus broadly applicable, but tend to generate many Type II errors. Tests are said to have 'power' according

[21] Please note that $P(Data|H_0)$ is not the same as $P(H_0|Data)$, i.e., the probability that $H_0$ is true given the data. As a consequence, if $P(Data|H_0)$ is high, we can only conclude that there is not enough evidence for a difference, not that the two runs behave the same.

to the number of Type II errors associated with them; the higher the power of a test, the fewer Type II errors it makes. Parametric tests, such as the Student's *t*-test make more assumptions, such as on the distribution of the data being tested. If these tests are applied to data in violation of their assumptions, Type I errors can result. However, Hull (1993) states that the *t*-test is relatively robust in this regard. According to Sanderson and Zobel (2005), the t- and the Wilcoxon tests allowed for more accurate prediction over the sign test when comparing retrieval runs, with the difference between the *t*-test and Wilcoxon being small. They also pointed out that even if one observed a significant difference between two runs based on a small number of topics ($\leq$25), one should not be confident that the same observed ordering of runs will be seen in the operational setting. However, the usefulness of statistical tests can also be questionable if samples are too large: with a large enough number of queries, any difference in effectiveness found when comparing two runs is likely to be significant.

### 5.2.8 System Effectiveness and User Satisfaction

When comparing two systems A and B the assumption is that if the effectiveness measure of A is greater than B then users would prefer system A. This is an important point as a test collection (including evaluation measures) acts as a simulation of a user's search and based on the results of the evaluation, the test collection is used to predict which retrieval system users would prefer. A number of studies have been undertaken to investigate the relationship between system effectiveness and user satisfaction with varying results. Broadly speaking results have indicated that user satisfaction does not always correlate with system effectiveness and that often users are able to adapt their search behaviour when presented with poor results by searching more intensively or by managing to exploit the relevant, but poorer, results.

A study performed by Sanderson et al. (2010) found clear evidence that effectiveness measured on a test collection did predict user preferences for one IR system over another. The strength of user prediction by test collection measures appeared to vary across different search tasks such as navigational or informational queries. The researchers studied this by recruiting users of Mechanical Turk (MTurkers) who were asked to compare the rankings of two search engines side-by-side (called a *preference judgment*) and indicate which set of results, from system A or B, they preferred for a given search topic. Various pairs of results from TREC's 2009 Web track were selected based on their relative differences and shown to users who selected the left or right sets of results as their preferred ranking. Different evaluation measures were compared including nDCG, MRR and P@10. It was found that measures such as nDCG correlated more highly than more widely used measures, such as P@10, which poorly modelled user preferences. This experiment would indicate that test collection results may well predict user satisfaction with search results and higher performance obtained on a

test collection would indeed lead to users being more satisfied in practice. User-oriented evaluation studies are discussed in the next section.

## 5.3 User-Oriented Evaluation

For many years IR evaluation has tended to focus on system-oriented evaluation, predominately through the use of standardised benchmarks or test collections in a laboratory-style setting as described in the previous section. However, IR systems are mainly used in an interactive way and this drives the need for user-centred evaluation to assess the overall success of a retrieval system as determined by the end users. Saracevic (1995) distinguishes six levels of evaluation for information systems: (1) engineering level, (2) input level, (3) processing level, (4) output level, (5) use and user level and (6) social level. Much of the current research in evaluating retrieval systems tends to focus on levels 1–4, but levels 5 and 6 are important in producing effective operational systems because they take into account factors other than just system performance, for example, how easy it is to use the interface, how quickly the system responds, how the results are presented. Many researchers have thus argued that a system-oriented laboratory-based IR evaluation framework is not sufficient and that alternative approaches must also be employed, e.g., Dunlop (2000) discusses an evaluation framework for evaluating interactive multimedia IR, Hansen (1998) discusses evaluation of IR user interfaces in web search, Petrelli (2007) discusses the role of user-centred evaluation within the context of interactive cross-language IR. Borlund (2003) asserts that *"... an information need ought to be treated as a user-individual and potentially dynamic concept and the multidimensional and dynamic nature of relevance should be taken into account ..."*.

Many of the basic tenets of information retrieval, traditionally focused on topic-based text documents, change when studying user behaviour in an interactive search session in a multilingual and multimedia context. A major difficulty is understanding to what extent language is situation-specific and dynamic. Another problem is tracking and understanding usage over time as the user learns and adapts during a session. Moving from text to other media entails new challenges with respect to the formulation of the information need; similarly, moving from monolingual to multilingual or cross-language information retrieval will change the way the system is able to match the formulated information need to document content and, most likely, change the way in which the users formulate their needs (Karlgren and Gonzalo 2010).

Evaluation of *Interactive IR systems* from a user-centred perspective typically forms part of an iterative design process. It can be formative (run at any point during the design process) or summative (run at the end of the design process); undertaken within a realistic setting (e.g., a heuristic evaluation with experts in the field) or under controlled conditions (e.g., experiments in the laboratory with users); used to evaluate a system as a whole or individual components (e.g., a formative evaluation of sub-components of an IR system such as query formulation, browsing or results

visualisation) and it can use a range of data collection techniques (naturalistic or qualitative methods such as content analysis of written data, or quantitative methods such as statistical analysis of log data or questionnaire results). However, user-centred evaluation is not only operational (testing how the system performs in practice) but can also be hypothesis-based (a research question is formulated in advance and proved/disproved through the evaluation).

### *5.3.1 Experimental Design*

In general, studies of human behaviour are cumbersome to set up and administer – instructing test subjects and ensuring adequate volume, reliability, and replicability of results is a challenge for any interactive study. Specifically, for information systems the challenge is twofold. Firstly, the coverage and breadth of the data source is one of the obviously important user satisfaction factors, the overhead effort of setting up a realistic test environment is a challenge in itself, and can seldom be practicably done for real-life tasks. This reduces most studies to mock-up or scaled-down environments. Secondly, the variation and breadth of information needs in the user population is immense. Thus interactive information retrieval studies tend to use simulated tasks – where test subjects are given a task to perform, seldom anchored in any real-life context.[22] The language skills of the testers constitute an additional factor to take into consideration when evaluating cross-language functionality, e.g., depending on whether the user has good, poor or no competence in the target language, the requirements on the user interface and the translation services offered may differ.

In order to alleviate the implicit dangers of the creation of artificial tasks, use cases are often employed. These are fairly informal specifications of the expected usage of the system, generally implemented via a set of specific 'scenarios', i.e., narrative descriptions of possible work situations. In order to define these, a study of the potential end users of the system and their typical work tasks/goals is needed. A controlled set of tasks simulating these scenarios can then be set up to evaluate the system within these pre-defined use cases. Users will be asked to complete these tasks and provide an assessment of their success/satisfaction with respect to various aspects such as ease of use, satisfaction with results, and usefulness of functionalities provided, generally also providing a numerical indication on a five or seven point scale. The results can be collected via interviews or questionnaires.

Table 5.4 shows the scenarios adopted in one user-centred evaluation.[23] Potential users of the system under development were requested to perform a sequence of

[22] See Borlund (2003) for examples of the application of simulated work task situations.

[23] These scenarios were used in the MultiMatch project which developed a multilingual multimedia search engine prototype for access, and personalised presentation of cultural heritage information (http://www.multimatch.eu/).

**Table 5.4** Usage scenarios for Cultural Heritage (CH) data

| User type | Task | Media and languages |
|---|---|---|
| CH professional | Searching for video footage on life and work of Pier Paolo Pasolini | Text, images, video<br>English, Dutch |
| Advertising agency | Looking for images of people drinking coffee that capture a feeling of relaxation | Text, images<br>English, Italian |
| Academic | Preparing a presentation on different artists' depictions of Don Quixote characters | Images, archives<br>English, Spanish |
| Cultural tourist<br>General user | Planning a visit to Turin, wants to know about museums and art galleries | Text, images, audio (podcasts)<br>English, Italian |

search tasks based on these scenarios; the tasks were carefully selected in order to highlight the diverse functionality of the system being assessed and to produce reliable results. In this type of evaluation, the use of a pre-defined set of tasks makes it possible to replicate the experiments in a controlled way.

In interactive evaluation, it is crucial that the experimental design is such that the presentation order for tasks, searchers and systems is balanced in order to facilitate comparison between systems and to remove bias. Experimental units should be assigned randomly and, if possible, there should be replication, i.e., the same tasks should be performed by more than one searcher but in a different order in order to ensure minimal user/topic and system/topic interactions.

Tague-Sutcliffe (1992) discusses how the use of standard experimental designs can help to alleviate this problem and describes the most commonly used experimental designs; the most popular of these is probably the Latin square design[24] employed in both the TREC and the CLEF interactive tracks. Table 5.5 shows the presentation order matrix used in the CLEF2002 interactive experiments. The basic design of the experiment consisted of:

- Two systems to be compared;
- A set of searchers, in groups of 4;
- A set of topic descriptions in one language;
- A document collection in a different language;
- A standardised procedure;
- A set of evaluation measures.

The minimum number of participants for any experiment was set at 4; additional participants could be added in groups of 8, with the same matrix being reused (Gonzalo and Oard 2003).

[24] A Latin square is an $n \times n$ table filled with $n$ different symbols in such a way that each symbol occurs exactly once in each row and exactly once in each column.

**Table 5.5** Example of Latin square design

| Searcher | Block 1 | Block 2 |
|---|---|---|
| 1 | System 1: 1–4 | System 2: 3–2 |
| 2 | System 2: 2–3 | System 1: 4–1 |
| 3 | System 2: 1–4 | System 1: 3–2 |
| 4 | System 1: 2–3 | System 2: 4–1 |
| 5 | System 1: 4–2 | System 2: 1–3 |
| 6 | System 2: 3–1 | System 1: 2–4 |
| 7 | System 2: 4–2 | System 1: 1–3 |
| 8 | System 1: 3–1 | System 2: 2–4 |

### *5.3.2 Evaluating Interactive CLIR Systems at CLEF*

CLEF is the only major evaluation campaign that is primarily focused on cross-language aspects of Information Access, and the only one that has run an interactive multilingual information access track (iCLEF) for many years (2001–2006, 2008–2009). The activity of this track has focused on studying how users address the problem of search over language boundaries and in investigating what system functionality can assist them. In 8 years, this track has addressed two main aspects of the problem: (1) document selection and results exploration; and (2) query formulation, refinement and translation. Both aspects have been addressed for various Information Access tasks (document retrieval, image retrieval, question answering...), from different methodological perspectives (hypothesis-driven, observational studies) and for different language profiles (i.e., different degrees of familiarity of the user with the target language/s). This activity has provided greater understanding of these issues and has played an important role in advancing the state-of-the-art.

For example, in the iCLEF 2008–09 campaigns the goal was generating a record of user-system interactions based on interactive cross-language image searches (Clough et al. 2010a). The level of entry to iCLEF was made purposely low with a default search interface and online game environment provided by the organisers. User-system interactions, such as queries, results, items clicked, selected query translations, query modifications, feedback from users and navigational actions and verbal input from users were recorded in log files for future investigation. This novel approach to running iCLEF resulted in logs containing more than two million lines of data. In total 435 users contributed to the logs and generated 6,182 valid search sessions (a session is when a user logs in and carries out a number of searches). In this version of the evaluation campaign, participants could recruit their own users and conduct experiments with the provided interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc. The logs provide a rich source of information for studying multilingual search from a user's perspective.

A detailed overview of iCLEF experiments is given in Braschler and Gonzalo (2009). We summarise the main findings listed in this report below:

- With respect to results selection over language boundaries, if possible, translating the whole document collection at index time pays off because it maps query reformulation and relevance feedback issues into their monolingual (much simpler) version. Using (appropriate) cross-language summaries can be a good solution, because they perform nearly as well as full documents, indexes take much less disk space, and are optimal for cross-language relevance assessment.
- In general, users are not comfortable choosing translations for their query terms in a foreign language; this task requires a high cognitive effort. Therefore, by default, query translation should be hidden from the user. However, the translation process is usually noisy and can lead to irrelevant results; in those cases, the system should be able to explain how query translation was performed, and to improve the translation with the help of the user.
- Feedback mechanisms that help to provide a better translation of the query without showing foreign language terms to the user can help to improve search effectiveness without adverse effects on the perceived difficulty of the task
- In image retrieval, combining text-based (even simple approaches) with content-based facilities can lead to better search effectiveness in less time and advanced presentation features (such as hierarchical clustering of search results) can be appreciated by users even if they do not lead to improved search effectiveness.

### *5.3.3 Alternative Performance Measures*

The Cranfield model described in the previous section is based on systematically run laboratory tests in which all variables are controlled. It commonly applies recall and precision (relevance-based measures) and focuses on measuring system effectiveness. As these measures have been mainly developed for batch searching environments, they present severe limitations in an interactive search environment or when the focus is on measuring user satisfaction in real world situations. Thus an important part of this type of evaluation is the definition of a suitable set of criteria and measures. A number of researchers have investigated alternative measures which take into account aspects other than relevance, e.g., utility, completeness, satisfaction, worth, time, cost, etc., in order to obtain a more complete picture of the overall performance of a system in the operational setting and taking into account the wider information seeking behaviour of users. Some of the measures, such as user confidence, usefulness, usability and satisfaction, are subjective; others, such as the time taken to complete a task, the number of queries issued, the number of search terms used, the number of errors in task results, the number of search result pages viewed, the number of relevant documents saved, the number of search result items clicked on, the rank of selected documents in the results and the number of results viewed are more objective measures of a user's behaviour.

Su (1992) selected a set of measures as important indicators of IR system performance and grouped them under four major evaluation criteria: relevance, efficiency, utility and user satisfaction. We have already considered measures for relevance extensively in Section 5.2.6. Here we focus on the other measures as defined by Su. Under efficiency, she lists three measures: search session time, relevance assessment time, search cost. Under utility she cites value of search results in terms of money, time, physical and mental effort expended, both considered separately and globally. She divides the user satisfaction slot into a number of categories regarding the searcher's understanding of the search and thoroughness in conducting it, and their satisfaction with its completeness and the precision of the results. However, the most important measure for Su was the user's overall judgment of the success of the system performance.

The measures listed by Su are based on two types of data: the scores given for each measure on a seven-point scale (quantitative data), and verbal data from post-search interviews (qualitative data). They can be collected via interviews, questionnaires, analysis of search logs and can still be considered valid today. A statistical analysis is proposed for the quantitative data and a content analysis for the verbal data.

Later work by Borlund (2003) proposes a framework for the evaluation of interactive IR systems and information searching behaviour through the introduction of an experimental setting which aims at measuring all the user's activities of interaction with retrieval and feedback mechanisms as well as the retrieval output. This is particularly important for multilingual systems where the system-user interaction enters into new and poorly understood dimensions as users are often querying target collections and trying to retrieve relevant documents in languages with which they are not well acquainted. The employment of an evaluation model of the type proposed by Borlund implies the adoption of alternative performance measures that take into account non-binary representations of relevance. She proposes two measures: Relative Relevance (RR), and Ranked Half-Life (RHL). The RR measure describes the degree of agreement between the types of relevance applied in evaluating IR systems in a non-binary context. On the other hand, the RHL indicator denotes the degree to which relevant documents are located on the top of a ranked result, thus measuring how well a system is capable of satisfying a user's need for information for a given set of queries at given precision levels. She also mentions the use of the graded relevance measures proposed by Järvelin and Kekäläinen (2000) described above in Section 5.2.6.

## 5.4 Evaluating Your Own System

There are various possibilities open to researchers or system developers who need to evaluate their own system. The first step is to decide whether they intend to perform a system or a user-oriented evaluation, i.e., whether they are most interested in measuring the effectiveness of the retrieval strategies adopted by the system or

whether they wish to assess the extent to which the system satisfies the user expectations. The reality is that both types of evaluation are needed. In both cases, performing a valid and reliable evaluation involves a certain degree of expertise on the part of the evaluator. Depending on the form the evaluation is to take, a background in IR techniques and methodology, system design and/or user studies is necessary. For large-scale testing, a team approach is probably desirable. However, much can be learned even from small-scale informal testing, perhaps by the system developer, as such tests may be closer to the actual context for which the system is intended. Tague-Sutcliffe (1992) provides useful guidance on many of the decisions that have to be taken when performing information retrieval experiments: how to choose the type of test needed (laboratory or operational), how to define the variables to be studied, how to choose the database, how to find and process the queries, how to collect and analyse the data, how to present the results. These are all issues that have to be taken into consideration when deciding to evaluate.

Individuals or groups wishing to test system performance have several possible alternatives: they can choose to participate in one of the well-known evaluation campaigns; they can reuse an existing test collection; they can build their own test collection.

A major advantage of participating in an evaluation campaign is the possibility offered to compare and discuss experiences and ideas directly with other groups working on the same problems. This is thus often the choice for researchers who want to acquire a deeper understanding and/or to test new ideas and discuss results with their peers. On the other hand, system developers may prefer to test performance effectiveness independently, reusing an existing collection and comparing their results against established baselines, without the constraints imposed by campaign participation. NTCIR makes test data available to non-campaign participants free-of-charge, TREC and CLEF test data can be obtained via independent distribution agencies: the Linguistic Data Consortium (LDC) in the case of TREC and the Evaluations and Language resources Distribution Agency (ELDA) for CLEF. The trec_eval package used in both TREC and CLEF can be obtained directly from the TREC website.[25]

When evaluating a cross-language system, it is always possible to extend an existing test collection by translating the query statements provided into different languages. However, depending on the specific evaluation task to be performed, no suitable test collections may be available. In this case, it is necessary to build your own test collection. The TrebleCLEF project has produced a set of useful guidelines for this purpose (Sanderson and Braschler 2009). They include a useful list of the questions that need to be considered and provide instructions for the acquisition and preparation of the components: documents; queries; relevance judgments; evaluation measures. They warn that before starting it is necessary to understand the purpose of the evaluation, e.g., whether comparing two different search strategies,

[25] The trec_eval package developed by Chris Buckley of Sabir Research and used in both TREC and CLEF evaluations is available free of charge: http://trec.nist.gov/trec_eval/

or attempting to optimise the strategies employed, or long term monitoring. The specific objective will impact on the design of the collection and the type of testing. Similarly, the type of searching typically conducted by the system must be taken into account, e.g., whether the main priority is high precision or high recall; whether the target collections are general purpose or of a specific domain, etc. In particular, they stress that it is necessary to consider the resources available to build the test collection; the creation of a large-scale test collection using the TREC methodology as described above requires significant human and financial resource.

If resources are scarce it may well be advisable to adopt one of the more simple evaluation options such as precision at ten or known item retrieval. If the evaluation is a comparison between the outputs of two systems or retrieval strategies, Carterette et al. (2008) suggest that it is possible to save time by only assessing the retrieved documents that will make a difference in the evaluation measure calculated on the two outputs. For example, if the evaluation measure is precision at ten, then only documents in the outputs that are unique to one of the systems need to be examined.

A known item search is a particular IR task in which the system component is asked to find a single target document in a given document set. The first $n$ documents in the ranked lists are assessed and the Mean Reciprocal Rank measure (MRR) described in Section 5.2.6 is used to calculate the results. An individual query will receive a score equal to the reciprocal of the rank at which the correct response was returned, or 0 if none of the responses contained a correct answer. The score for a submission is the mean of the individual queries reciprocal ranks. The reciprocal rank has several advantages as a scoring metric. It is closely related to the mean average precision measure used extensively in document retrieval. It is bounded between 0 and 1, inclusive, and averages well. A run is penalised for not retrieving any correct answer for a query but not unduly so (Voorhees and Garofolo 2005). However, with a known-item search it is crucial that the 'known-item' in question must be unique in its properties with respect to the retrieval methods to be tested; this means that a 'good' system will return documents with the known item rather than other documents. It is not easy to set up this kind of task. Another alternative which can be employed in order to reduce the effort needed to create the relevance assessments is the Interactive Searching and Judging technique described above in Section 5.2.5.

A user-centred evaluation will start with a study of the user context in order to identify an appropriate set of user requirements on which to base the system design. These requirements can be collected via interviews and questionnaires – or when feasible via the observation of the user interaction with an existing system.

When setting up a user-oriented evaluation, the correct design of the test setting is crucial and the points listed in Section 5.3.1 should be borne in mind.

**Summing Up:** Evaluation of an MLIR system is not an activity to be undertaken lightly. A number of decisions must be made when planning the activity. They have been well summed up by van Rijsbergen (1979) in three questions: (1) Why evaluate? (2) What to evaluate? (3) How to evaluate? We elaborate on these questions in the following points for consideration:

1. Scope: What is the main scope of the evaluation? Is it focused more on improving system effectiveness or user satisfaction? Is it a laboratory-based predominately hypothesis-driven theoretical activity, or an operational one, on a running system with real users?
2. Objective: What particular aspect of the system performance is the objective of the evaluation? Should the evaluation focus on the performance of a specific component, e.g., the indexing or ranking algorithms, the translation mechanisms, feedback modules, output presentation, interface usability?
3. Experimental design: The evaluation tasks must be designed appropriately, depending on the type of evaluation to be conducted and the dependent and independent variables that need to be taken into account?
4. Test Collection: What test collection will be adopted? Is there a suitable existing test collection that can be reused? Is it necessary to build an ad hoc collection specifically for the task? If so, what resources are available for this? What kind of query statements are needed, how will the ground truth be obtained, what evaluation measures will be employed?
5. Methodology: What methodology will be adopted to conduct the evaluation and compute the measurements?
6. Results: Whatever the type of evaluation and whether the main purpose was purely internal system testing or a desire to advance the state-of-the-art via scientific experimentation, it is important that the results are written up in a clear and exhaustive fashion in order to permit future replication and/or comparison.

Hopefully, this chapter has provided the reader with the necessary information on which to base their design choices, depending on the specific objective of their evaluation task.

## 5.5 Summary and Future Directions

This chapter has described the evaluation of systems for multilingual information retrieval. We have discussed evaluation from both system- and user-oriented perspectives and have provided the reader with guidelines on how to undertake their own evaluation activity, listing the points that need to be taken into consideration before starting out. However, the intent of the chapter has been twofold: the focus is not only on how to conduct system testing and benchmarking for MLIR/CLIR systems, but also on the importance of the role of evaluation in advancing the state-of-the art.

Evaluation is at the very heart of the IR field – and thus is also central to questions that address MLIR and CLIR. It is generally regarded as the prime instigator of progress in the field, stimulating the development of systems that give better results. This means that the models, methodologies, measures and metrics adopted in the experiments are subject to critical scrutiny. (ML)IR is no

longer just about text, today's content is overwhelmingly multimedia, and the role of the user is rapidly changing. From being – more or less – passive recipients in a typically academic context, today's users are entering into dynamic interaction with digital content for the most disparate motives: from educational to leisure, from business to entertainment, from work to play. For this reason, the Cranfield evaluation paradigm, described in Section 5.2.1 and generally recognised as having laid the foundations for research in the IR domain, has recently come under critical examination. It is felt that its scope is too limited to meet the new challenges, the model needs to be broadened and the user perspective needs to be considered in more depth.

These ideas emerged clearly at a recent SIGIR workshop on the future of IR evaluation where the limitations of current benchmark test collections to address the scale, diversity, and interaction that characterise information systems nowadays were discussed (Geva et al. 2009). There was a call for new hypotheses and models, more representative of the new searching paradigms, which enable a better understanding of the factors involved; results should be assessed not only from the standpoint of relevance but also with respect to aspects such as redundancy, novelty and diversity. In addition, new formal models of information seeking behaviour that can be subjected to empirical testing are needed. This is especially true in the multilingual context where the linguistic competence and the cultural expectations of the user navigating content in an unfamiliar language comport additional layers of complexity. Finally, there should be more in situ testing of user interaction with system components and new methods for gathering user data in an ongoing way must be devised. The main conclusion of the workshop was that there is more to IR than the evaluation of systems and their rankings, and that there is still a general disconnect between user-centred and system-centred research that must be overcome.

It seems that the assertion by Saracevic (1995) still holds true today: both system- and user-centred evaluations are needed but they should work together and feed on each other, working towards co-operative efforts and mutual use of results.

## 5.6 Suggested Reading

There are a number of authoritative publications providing detailed accounts of the issues involved in traditional IR evaluation. These include chapters in books by Salton (1968), van Rijsbergen (1979), Croft, Metzler and Strohman (2009), and Spärck Jones's (1981) edited articles on Information Retrieval experiments. The history of the TREC evaluation exercises is reported by Voorhees and Harman (2005), and special issues of Information Processing and Management (Harman 1992) and of the Journal of the American Society for Information Science (Tague-Sutcliffe 1996) were dedicated to this topic. A special issue of Information Retrieval (Braschler and Peters 2004) concentrated on cross-language evaluation.

Müller et al. (2010) contains a collection of texts centred on the evaluation of image retrieval systems including references to experiments in the multilingual context. The centrality of the concept of relevance both for information retrieval systems and by extension for their evaluation is discussed in an exhaustive review of the literature by Mizzaro (1998). Borlund (2003) discusses the evaluation of Interactive Information Retrieval (IIR) systems and proposes a framework to facilitate this and Kelly (2009) gives a detailed survey on the same topic with the goal of cataloguing all related literature. Borlund also includes a chapter on user-centred evaluation of IR systems in which she focuses on a cognitive-oriented IR evaluation approach using laboratory-based simulated work tasks (Borlund 2009). Finally, we recommend Stephen Robertson's history of IR evaluation (Robertson 2008) and Tefko Saracevic's much cited paper on the evaluation of evaluation in information retrieval (Saracevic 1995).

## References

Alonso O, Rose DE, Stewart B (2008) Crowdsourcing for relevance evaluation. ACM SIGIR Forum. 42(2): 9–15

Aslam JA, Yilmaz E, Pavlu V (2005) A geometric interpretation of R-precision and its correlation with average precision. In: Proc. 28th ACM SIGIR conference on research and development in information retrieval (SIGIR 2005). ACM Press: 573–574

Bailey P, Craswell N, Soboroff I, Thomas P, de Vries AP, Yilmaz E (2008) Relevance assessment: Are judges exchangeable and does it matter? In: Proc. 31st ACM SIGIR conference on research and development in information retrieval. ACM Press: 667–674

Borlund P (2003) The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. J Inf. Res. 8(3)

Borlund P (2009) User-centred evaluation of information retrieval systems. Chapter 2. In: Göker A, Davies J (eds.) Information Retrieval: Searching in the 21st Century. John Wiley & Sons: 21–37

Braschler M (2004a) CLEF 2003 – Overview of results. In: Comparative evaluation of multilingual information access systems. 4th workshop of the Cross-Language Evaluation Forum, CLEF 2003, Springer LNCS 3237: 44–63

Braschler M (2004b) Robust multilingual information retrieval. Doctoral Thesis, Institut interfacultaire d'informatique, Université de Neuchâtel

Braschler M, Gonzalo J (2009) Best practices in system and user-oriented multilingual information access. TrebleCLEF Project: http://www.trebleclef.eu/

Braschler M, Peters C. (eds.) (2004) Cross-language evaluation forum. J Inf.Retr. 7(1–2) 2004

Buckley C, Voorhees E (2004) Retrieval evaluation with incomplete information. In Proc. 27th ACM SIGIR conference on research and development in information retrieval (SIGIR 2004). ACM Press: 25–32

Carterette B, Pavlu V, Kanoulas E, Aslam JA, Allan J (2008) Evaluation over thousands of queries. In: Proc. 31st ACM SIGIR conference on research and development in information retrieval (SIGIR 2009). ACM Press: 651–658

Cleverdon CW (1967) The Cranfield tests on index language devices. In: Aslib Proc. 19(6): 173–192

Cleverdon CW (1970) The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical Report No. 3, Cranfield Institute of Technology, Cranfield, UK, 1970

Clough P, Sanderson M, Müller H (2004) The CLEF cross language image retrieval track (ImageCLEF) 2004. In: Image and Video Retrieval (CIVR 2004), Springer, LNCS 3115: 243–251

Clough P, Gonzalo J, Karlgren J (2010a) Creating re-useable log files for interactive CLIR, In SIGIR 2010 Workshop on the Simulation of Interaction (SimInt), Geneva, Switzerland, 23 July 2010

Clough P, Müller H, Sanderson M (2010b) Seven years of image retrieval evaluation. In: Müller H, Clough P, Deselaers Th, Caputo, B. (eds.) (2010) ImageCLEF experimental evaluation in visual information retrieval. The Information Retrieval Series. Springer 32: 3–18

Cormack GV, Palmer CR, Clarke CLA (1998) Efficient construction of large test collections. In: Proc. 21st ACM SIGIR conference on research and development in information retrieval (SIGIR 1998). ACM Press: 282–289

Croft B, Metzler D, Strohman T (2009) Evaluating search engines. In: Search engines: Information retrieval in practice 1st ed., Addison Wesley: 269–307

Dunlop M (2000) Reflections on MIRA: Interactive evaluation in information retrieval. J. Am. Soc. for Inf. Sci. 51(14): 1269–1274

Geva S, Kamps J, Peters C, Sakai T, Trotman A, Voorhees E (eds.) (2009) Proc. SIGIR 2009 workshop on the future of IR evaluation. SIGIR 2009, Boston USA. http://staff.science.uva.nl/~kamps/publications/2009/geva:futu09.pdf

Gonzalo J, Oard DW (2003). The CLEF 2002 interactive track. In: Advances in cross-language information retrieval. 3rd workshop of the Cross-Language Evaluation Forum, CLEF 2002. Springer LNCS 2785: 372–382

Hansen P (1998) Evaluation of IR user interface – Implications for user interface design. Human IT. http://www.hb.se/bhs/it

Harman DK (ed.) (1992) Evaluation issues in information retrieval. J. Inform. Process. & Manag. 28(4)

Harman DK (2005) The TREC test collections. In: Voorhees EM, Harman DK (eds.) TREC: experiment and evaluation in information retrieval, MIT Press, 2005

Howe J (2006). The rise of crowdsourcing. Wired, June 2006. http://www.wired.com/magazine/

Hull D (1993) Using statistical testing in the evaluation of retrieval experiments. In: Proc. 16th ACM SIGIR conference on research and development in information retrieval (SIGIR 1993). ACM Press: 329–338

Järvelin K, Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: Proc. 23rd ACM SIGIR conference on research and development in information retrieval (SIGIR 2000). ACM Press: 41–48

Jansen BJ, Spink A (2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs. J. Inform. Process. & Manag. 42(1): 248–263

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In Proc. First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition: 11–44

Karlgren J, Gonzalo J (2010) Interactive image retrieval. In Müller H, Clough P, Deselaers Th, Caputo, B. (eds.) ImageCLEF. Experimental evaluation in visual information retrieval. The Information Retrieval Series. Springer 32: 117–139

Kelly D. (2009) Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval. 228 p.

Lesk ME, Salton G. (1969) Relevance assessments and retrieval system evaluation. J Inf. Storage and Retr. 4: 343–359

Leung C, Ip H (2000) Benchmarking for content-based visual information search. In Fourth International Conference on Visual Information Systems (VISUAL'2000), LNCS 1929, Springer-Verlag: 442–456

Mandl T (2009) Easy tasks dominate information retrieval evaluation results. In Lecture Notes in Informatics. Datenbanksysteme in Business, Technologie und Web (BTW): 107–116

Mandl T, Womser-Hacker C, Di Nunzio GM, Ferro N (2008) How robust are multilingual information retrieval systems? Proc. SAC 2008 - ACM Symposium on Applied Computing: 1132–1136

Mizzaro S (1998) Relevance: The whole history. J. of Am. Soc. for Inf. Sci. 48(9): 810–832

Müller H, Müller W, Squire DM, Marchand-Maillet S, Pun T (2001) Performance evaluation in content-based image retrieval: overview and proposals. Pattern Recognit. Lett. 22(5): 593–601

Müller H, Clough P, Deselaers Th, Caputo, B. (eds.) (2010) ImageCLEF. Experimental evaluation in visual information retrieval. The Information Retrieval Series, Springer 32: 495 p.

Peters C (ed.) (2000) First results of the CLEF 2000 cross-language text retrieval system evaluation campaign. Working notes for the CLEF 2000 workshop, ERCIM-00-W01: 142

Petrelli D (2007) On the role of user-centred evaluation in the advancement of interactive information retrieval. J. Inform. Process & Manag. 44(1): 22–38

van Rijsbergen CJ (1979) Evaluation. In: Information Retrieval 2nd ed., Butterworths

Robertson S (2006) On GMAP: and other transformations. Proc. 15th ACM international conference on information and knowledge management (CIKM '06). ACM Press: 78–83

Robertson S (2008) On the history of evaluation in IR. J. of Inf. Sci. 34(4): 439–456

Salton G (1968) Automatic information organization and retrieval. McGraw Hill Text

Sanderson M, Braschler M (2009) Best practices for test collection creation and information retrieval system evaluation, TrebleCLEF Project: http://www.trebleclef.eu

Sanderson M, Zobel J (2005) Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proc. 28th ACM SIGIR conference on research and development in information retrieval (SIGIR 2005). ACM Press: 162–169

Sanderson M, Paramita M, Clough P, Kanoulas E (2010) Do user preferences and evaluation measures line up?In: Proc. 33rd ACM SIGIR conference on research and development in information retrieval (SIGIR 2010). ACM Press: 555–562

Saracevic T (1995) Evaluation of evaluation in information retrieval. In: 18th ACM SIGIR conference on research and development in information retrieval (SIGIR 1995), ACM Press: 138–146

Schäuble P (1997) Multimedia information retrieval: Content-based information retrieval from large text and audio databases, Kluwer Academic Publishers

Schamber L (1994) Relevance and information behaviour. Annual Review of Information Science and Technology, 29: 3–48

Smith JR (1998) Image retrieval evaluation. In: Proc IEEE Workshop on Content-based Access of Image and Video Libraries:112–113

Spärck Jones K (1981) Information Retrieval Experiment. Butterworth-Heinemann Ltd. 352 p.

Spärck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development report 5266. Cambridge: Computer Laboratory, University of Cambridge

Su LT (1992) Evaluation measures for interactive information retrieval. J. Inform. Process & Manag. 28(4): 503–516

Tague-Sutcliffe J (1992) The pragmatics of information retrieval experimentation, revisited. J. Inform. Process & Manag. 28(4): 467–490

Tague-Sutcliffe J (ed.) (1996) Evaluation of information retrieval systems, J. Am. Soc. for Inf. Sci. 47(1)

Tomlinson S (2010) Sampling precision to depth 10000 at CLEF 2009. In: Multilingual information access evaluation Part I: Text retrieval experiments: 10th workshop of the Cross-Language Evaluation Forum, CLEF 2009, Springer, LNCS 6241: 78–85

Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. J Inform. Process. & Manag. 36(5): 697–716

Voorhees EM (2002) The philosophy of information retrieval evaluation. In: Evaluation of cross-language information retrieval systems: 2nd workshop of the Cross-Language Evaluation Forum, CLEF 2001, Springer, LNCS 2406: 355–370

Voorhees EM (2006) Overview of TREC 2006. Fifteenth text retrieval conference (TREC 2006) Proc. NIST Special Publication SP500-272. http://trec.nist.gov/pubs/trec15/

Voorhees EM, Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proc. 25th ACM SIGIR conference on research and development in information retrieval (SIGIR 2002). ACM Press: 316–323

Voorhees EM, Garofolo JS (2005) Retrieving noisy text. In Voorhees EM, Harman DK (eds.). TREC. experimentation and evaluation in information retrieval. MIT Press: 183–197

Voorhees EM, Harman DK (eds.) (2005) TREC: experiment and evaluation in information retrieval. The MIT Press, Cambridge, MA

Womser-Hacker C (2002) Multilingual topic generation within the CLEF 2001 experiments. In: Evaluation of cross-language information retrieval systems: 2nd workshop of the Cross-Language Evaluation Forum, CLEF 2001, Springer, LNCS 2406: 255–262

Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Proc. 21st ACM SIGIR conference on research and development in information retrieval (SIGIR 1998). ACM Press: 307–314

# Chapter 6
# Applications of Multilingual Information Access

*"In a global community, users are looking for online information access systems or services that can help them find and use information presented in native or non-native languages."*
Chen and Bao 2009

**Abstract** The continually growing number of applications creating and/or using collections of natural language digital documents in diverse media and languages means that there is increasing demand for technologies to access the information contained in these archives. When the documents are comprised of media other than text and in languages unfamiliar to the user of the application, then a more complex approach is required than that used for monolingual textual document retrieval. Technologies employed for Multilingual Information Access (MLIA) must be integrated with others in a seamless fashion. In this chapter, we discuss ways in which the technologies developed to implement systems for Multilingual and Cross-Language Information Retrieval (MLIR/CLIR) are adopted in areas that go beyond textual document search such as multimedia retrieval and information extraction over language boundaries. We also describe a range of practical application domains which employ these technologies, thus helping to motivate the need for further developments in MLIR/CLIR.

## 6.1 Introduction

So far in this book, we have focused on the design, development and evaluation of multilingual information access systems, with the main focus on text retrieval. However, 'The Grand Challenge' cited in Chapter 1, i.e., the demand for fully *multilingual multimedia* retrieval systems first formulated in 1997, is increasingly relevant today as the global networks become vital sources of information for both professional and leisure activities. The problem is that much Internet content is effectively inaccessible as it is stored in media files and in languages that are not searchable without specific technical know-how.

In this chapter, we first provide an overview of research that applies multilingual information access technologies in areas that go beyond textual document search such as (1) the retrieval of information from archives containing material in one or more audio or visual media (e.g., image, speech or video), and (2) the handling of

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0_6, 

retrievable items other than documents (e.g., information extraction or question answering). Solutions to these retrieval problems integrate a number of technologies. For example, access to a multilingual collection containing electronic text, handwritten documents, and/or speech and images could require techniques from Information Retrieval (IR), Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), Content- and Text-based Image Retrieval (CBIR/TBIR) and Machine Translation (MT). Developing some kind of information extraction system generally involves the adoption of techniques from the Natural Language Processing (NLP) domain. One important consequence of these research activities has been the study and development of many types of language-specific processing tools, such as speech recognisers, syntactic and semantic parsers and analysers, dictionaries and translation systems. Another result has been the adapting and tuning of systems initially tested in one language to handle others, thus enabling the global dissemination and access of information over media and language boundaries.

In the second part of the chapter, we describe a number of domains in which currently available multilingual and multimedia information access technologies are being studied and employed. Each domain has its own unique characteristics and specific challenges which must be recognised and catered for if MLIA technologies are to be successfully employed in practical contexts.

## 6.2 Beyond Multilingual Textual Document Retrieval

In this section, we provide an overview of the current state of research in a number of areas that go beyond textual document retrieval. The aim is to give the reader an idea of what is involved when MLIR/CLIR technologies are integrated in other kinds of applications. There are situations where the object to retrieve is not a full-text document but instead an image (maps, photographs, graphics and clip art), sound (music, speech and audio clips) or perhaps a combination of information in more than one media (e.g., scenes from a video). Multimedia content can be either static, in the case of digitised images such as photographs or paintings, or temporal, when comprising audio and/or video files. Retrieval from multimedia archives thus raises a number of issues with respect to content processing, browsing and presentation. For example, the fact that different media adopt different standards for encoding is a major problem as it makes interoperability difficult. There are also situations when, rather than receiving a set of documents potentially containing relevant information in response to a query, the user needs to extract, filter or categorise specific information from the document archives. An interesting example of this type of requirement is when a precise answer is required as the result of a search. In order to provide this, a question answering system will generally perform a series of syntactic and semantic analyses on both queries and documents. In this case, the indexing processes are considerably more complex than with textual document retrieval systems. The introduction of multilingual content and all that

this implies (further encoding, indexing, matching and presentation issues) adds an extra dimension of complexity.

In this section we discuss research activities in these areas (multimedia and question answering) where the focus is on providing solutions that are able to overcome language boundaries. Much of the discussion refers to studies and experiments undertaken in the context of the Cross-Language Evaluation Forum (CLEF).[1] The mission of CLEF has been to promote the development of systems that satisfy the requirements of the Grand Challenge mentioned above.

### *6.2.1 Image Retrieval*

Cross-language image retrieval is an area of research that is receiving increasing attention from the research community (Clough and Sanderson 2006, Müller et al. 2010). Images by their very nature are language independent, but are often accompanied by semantically related text (e.g., free-text captions, structured metadata and, in the case of web pages, anchor text), and thus offer a clear use case for adopting CLIR in practice. As Oard (1997) commented: "*an image search engine based on cross-language free text retrieval would be an excellent early application for cross-language retrieval because no translation software would be needed on the user's machine*". This is because the relevance of images retrieved with respect to a given query can often be judged regardless of the user's linguistic abilities and foreign language skills.

Approaches for retrieving visual objects depend on the information that is associated with the object (Enser 1995, Rasmussen 1997, Goodrum 2000). In one approach, techniques taken from the field of computer vision are applied to extract low-level features, such as colour, shapes and texture, for indexing and retrieval (Del Bimbo 1999, Smeulders et al. 2000). This method is commonly referred to as Content-Based Image Retrieval (CBIR). In this form of retrieval the typical *modus operandi* is that users will submit a visual exemplar (e.g., a similar image to that sought or a sketch of the desired image), the system will compare features extracted from the query with those extracted from archived images and will rank the results on the basis of dissimilarity/distance measures between the feature representation of the exemplar and feature representations of images from the collection. Subsequent user interaction can refine the query on the basis of relevance feedback where the users indicate which images in the results are relevant (or not relevant).

Despite the many advances in content-based approaches a number of significant challenges still exist. One problem is the 'semantic gap' – the gap that exists between the automatically extracted low-level image features, which are ultimately represented as a sequence of numeric values (i.e., in binary form), and the high-

[1] http://www.clef-campaign.org/

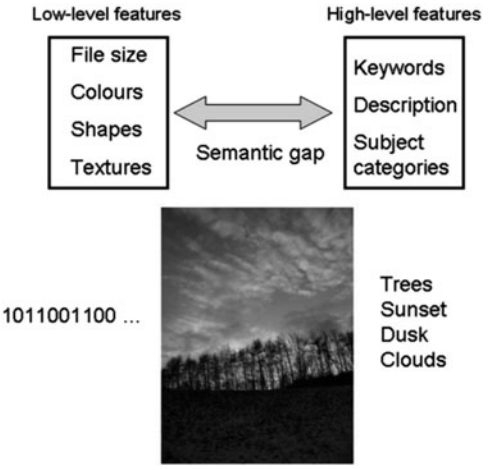


**Fig. 6.1** The 'semantic gap' in image retrieval is the gap between the low-level features used to represent an image (expressed as numeric values) and associated text that could be used to describe the contents of the image

level concepts (the semantics) depicted in an image. In fact, CBIR systems are often not intuitive for users accustomed to issuing natural language queries rather than visual exemplars.

The semantic gap is depicted in Figure 6.1 which shows an example image that consists of pixels from which, at one level, low-level features, such as colours, shapes and textures in the image, can be extracted and, at another level, can be represented by a list of keywords describing the semantic content of the image (e.g., visual properties, locations, people, emotions, etc.). The keywords can be added to describe the content of the image, either manually or automatically. The gap may also widen when keywords that describe semantics not easily derived from the visual contents, such as the location of the image, or labels that express emotion or feeling, are added.

The problem of the semantic gap is further compounded in MLIR/CLIR where images may represent different concepts in different cultures. Fortunately, in practice many images are accompanied by textual metadata (e.g., assigned as terms from a controlled vocabulary, or as user-created tags or captions), which describe properties of the image, such as its visual content or creation history. This associated text can be used to provide simple text-based access to images and is often the user's preferred method for image retrieval interaction (Eakins et al. 2004). Such approaches are often referred to as concept-based or Text-Based Image Retrieval (TBIR) and are generally the most popular form of image retrieval, particularly on the Web, as standard IR techniques can be used. Figure 6.2 shows an example image from the St Andrews University Library (Reid 1999) and its associated metadata, which is separated into eight distinct fields, all (or a subset) of which can be used for text-based image retrieval. In web image search, where there is no evident metadata, the text on the web page surrounding an image is often used for retrieval; in this case care must be taken to filter out non-relevant text during indexing in order to limit the risk of noise, i.e., the retrieval of irrelevant images.

| | |
|---|---|
| Record ID: | JV-A.000460 |
| Short title: | The Fountain, Alexandria. |
| Long title: | Alexandria. The Fountain. |
| Location: | Dunbartonshire, Scotland |
| Description: | Street junction with large ornate fountain with columns, surrounded by rails and lamp posts at corners; houses and shops. |
| Date: | Registered 17 July 1934 |
| Photographer: | J Valentine & Co |
| Categories: | Columns unclassified, street lamps – ornate, electric street lighting, streetscapes, shops |
| Notes: | JV - A460 jf/mb |


**Fig. 6.2** Example image with associated metadata. This image is copyright of St Andrews University Library

The quantity and quality of any text associated with an image will affect retrieval performance as it will impact on the possibility of finding a match between the search terms entered by the user and the text used to describe the image. This leads to a number of problems associated with text-based image retrieval:

- The manual annotation of images can be time consuming and expensive;
- Manual annotation is subjective and can suffer from low agreement between individuals (and groups);
- The meaning of an image can be difficult to interpret and express in a written form (especially visual or emotive concepts);
- Many images are associated with short texts, and the criteria used to judge relevance may differ from that used for longer text documents (e.g., the quality and size of an image may be included).

Often combining the outputs from retrieval based on low-level features and associated textual descriptions can improve the overall performance of image retrieval systems. This combination of different modalities is known as data or information *fusion*[2] and can help to overcome the weaknesses of using an approach based on a single modality alone. There are several techniques for information fusion in image retrieval that vary depending on when the fusion takes place in the retrieval process (Hsu and Taksa 2005, Depeursinge and Müller 2010):

1. At the input of an IR system by submitting multiple queries or using query expansion;
2. Within the IR system itself by using multiple algorithms to retrieval documents;

[2] Chapter 3 discusses the use of data fusion for combining the results from searching multilingual document collections.

3. At the output of the IR system by combining multiple results lists (e.g., the results from CBIR and TBIR systems).

Early academic cross-language image retrieval systems were purely text-based, using some kind of translation mechanism to match user queries against text associated with the images across languages. One example was the Eurovision system (Sanderson and Clough 2002, Clough and Sanderson 2006). This system operated on a historic image archive, a collection of historic images from St Andrews University library, and used machine translation for the translation of user's queries, the user interface and search results. More recently, the PanImages cross-language image search engine provided access to images on the Web, utilising dictionary-based query translation rather than using MT (Colowick 2008). The system included features, such as auto-completion, and also allowed users to add their own (preferred) translations. The FlickrArabic application (Clough et al. 2008) translated a user's query from Arabic into English, which was then used as an *interlingua*, or pivot language, to translate into French, Spanish, German, Italian or Dutch. Chapter 4 shows examples of screen shots from these three systems.

Nowadays, the implementation of cross-language image retrieval systems is typically a combination or fusion of standard cross-language information retrieval and image retrieval techniques. The academic community has shown that leveraging the use of both content- and text-based techniques offers the most effective form of cross-language image retrieval across multiple search requests.

Events such as ImageCLEF,[3] which has evaluated cross-language image retrieval systems in a standardised manner in annual campaigns since 2003, have allowed comparison between the various approaches and have helped to systematically investigate the effectiveness of different algorithms (Müller et al. 2010). Reviewing the ImageCLEF results for systems that adopt combination techniques (in total 62% of papers in ImageCLEF submissions from 2003 to 2009 mixed TBIR and CBIR techniques), Depeursinge and Müller (2010) note that fusion at the output of the system (point 3 above) is by far the most widely used fusion strategy (over 60% of the papers describing fusion techniques). However, they also note that combining textual and visual information is not devoid of risk and can degrade the retrieval performance if the fusion technique is not well adapted to the information retrieval paradigm. The trick is to be able to make the most of both modalities.

Much of the academic work in cross-language image retrieval has been system-oriented and focused on improving the effectiveness of retrieval algorithms. However, in the 2008 and 2009 campaigns, the interactive track of the Cross-Language Evaluation Forum (CLEF), known as iCLEF,[4] focused on cross-language image retrieval from the user's perspective (Karlgren and Gonzalo 2010). The organisers

[3] http://www.imageclef.org/

[4] http://nlp.uned.es/iCLEF/

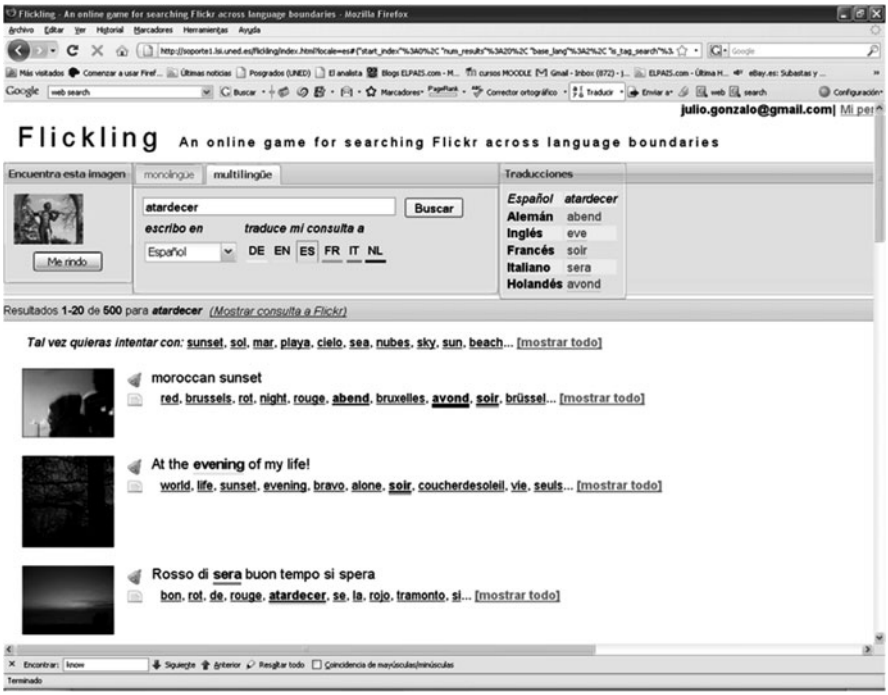


**Fig. 6.3** An example of the interface provided for the iCLEF 2008–09 experiments (Flickling)

provided cross-language information access to Flickr, a large-scale online photograph sharing tool, using a default multilingual search system called Flickling (Fig. 6.3). Participants were assigned a set of simple (known-item) interactive image search examples and the interaction of the users with the system was recorded in log files which were then analysed to investigate the effects of language skills on users' search behaviour. This activity has resulted in a valuable set of data for studying multilingual image search from the user perspective.

### *6.2.2 Speech Retrieval*

Speech retrieval uses a combination of Automatic Speech Recognition (ASR) and information retrieval technologies. A speech recogniser is applied to an audio stream (the spoken document) and generates a time-marked transcription of the speech. The transcription may be phone- or word-based in either a lattice (probability network), *n*-best list (multiple individual transcriptions), or more typically, a one-best transcript (the most probable transcription as determined by the recogniser). The recogniser is typically trained on a large corpus of labelled acoustic speech data. Spoken words not included in the vocabulary of the speech

recognition system may not be indexed correctly. Other factors will impact on the performance of the system such as poor articulation, spontaneous speech issues, e.g., pauses, break-offs, repetitions, and acoustic channel noise (Jones 2000). The transcribed data is indexed and can then be searched by a retrieval system. The result returned for a query is a list of temporal pointers to the audio stream ordered by decreasing similarity between the content of the speech being pointed to and the query (Garofolo et al. 1997).

Most initial research on speech retrieval focused on news broadcasts. These broadcasts have specific characteristics – clearly articulated speech, well-formed sentences, easily recognised breaks between topics – that facilitated studies. Retrieval on this kind of speech data is known as Spoken Document Retrieval (SDR).

The performance of a speech recognition system is generally measured in terms of the word error rate percentage. For state-of-the-art broadcast news transcription, word error rates below 10% are reported for widely studied languages such as English; whereas for spontaneous speech, word error rates between 40% and 60% are by no means exceptional (de Jong et al. 2008).

In recent years, automatic speech recognition technology has progressed to a level that is sufficient to make speech recognition transcript-derived indexing features effective for text retrieval. Experiments on information retrieval from speech transcripts at the Text Retrieval Conference (TREC)[5] found that as long as speech recognition has a word error rate better than 35%, then IR is only 3–10% worse than with perfect text transcriptions (Hauptmann et al. 2002).

While the first studies on speech recognition focused on English, later work has been extended to many other languages. Unfortunately, spoken content indexing has a high entry threshold, and collections or language-specific recognition systems are time consuming and expensive to develop. For each new language, new manually coded training collections and language models must be built. This tends to make speech retrieval research a niche area. Work directed towards affordable access to spoken content collections specifically targets the challenge of "*maximizing the potential of the collection while minimizing development costs*" (Ordelman et al. 2006).

The first experiments in Cross Language-Spoken Document Retrieval (CL-SDR) were reported in 1997 for a cross-language speech retrieval system finding German spoken news documents in response to user queries entered as French text (Sheridan et al. 1997a). The speech retrieval module was based on a speaker-independent phoneme recogniser indexing spoken documents by *n*-grams of phonemic features. The performance was evaluated over a collection of 30 hours of spoken news and the cross-language speech retrieval results were measured at about 45% of monolingual speech retrieval. Jones (2000) reported somewhat better performance using a Hidden Markov Model (HMM)-based recogniser for a small

[5] http://trec.nist.gov/

collection of 5 hours of spontaneous English spoken video mail messages. A number of experiments were made using French search requests with the best results varying from about 86% of monolingual performance using machine translation and 68% using a bilingual dictionary look-up system.

TREC ran an SDR track for 4 years from 1997 to 2000 using collections of American English broadcast news and demonstrated very good performance levels (Garofolo et al. 2000). In 2002 CLEF decided to exploit the test collections created during the final 2 years of this track (approx 550 hours of broadcast news from different sources) and extend them to a cross-language dimension by building topics in different languages. The TREC-8 and TREC-9 SDR topics were thus translated into French, German, Italian, Spanish and Dutch. The aim was to run a pilot study in order to establish a baseline for CL-SDR retrieval. The best performances gave results ranging between 75% and 80% of monolingual performance. CLEF 2003 and 2004 again offered this track using the same collection. As was expected, in all cases the cross-language results of all participants were downgraded with respect to the monolingual results. As with traditional CLIR tasks, it was noted that the degree of degraded performance seemed to depend on the translation resources used (Federico and Jones 2004) and the difficulty of the particular task (Federico et al. 2005).

Most initial research on spoken document retrieval focused on news broadcasts. However, a number of studies have also been conducted in the cultural heritage domain. This typically includes collections containing oral history interviews, lectures, talk shows and studio discussion as well as user generated podcasts. Cultural heritage content differs from broadcast news in that it is not necessarily structured into stories separated by easily recognised breakpoints and that the vocabulary used by the speakers and background conditions used for recording are significantly less predictable. In oral history interviews, for example, particular challenges include spontaneous speech, emotional speech, speech of elderly speakers, highly accented and regionally specific speech, foreign words, names and places (Byrne et al. 2004). This means that speech recognition word error rates for spoken audio content are highly variable (Huijbregts et al. 2007). In addition to speech recognition, audio segmentation and spoken content categorisation are two other areas that present particular challenges when providing access to this kind of content (Byrne et al. 2004). Creating appropriate segments is important not only for indexing, but also for display of spoken content results to the user in the interface in a way in which they can be easily skimmed (de Jong et al. 2008).

From 2005 to 2008, CLEF ran a Cross-Language Speech Retrieval (CL-SR) track on a cultural heritage collection derived from a carefully anonymised set of digitised and annotated interviews in English with Holocaust survivors made available by the Shoah Visual History Foundation. The distinction was made with spoken document retrieval as the collections used were spontaneous conversational speech – considerably more challenging for Automatic Speech Recognition (ASR) techniques than transcribing the speech of news readers and anchors. The aim in the first year was to create the first representative test collection for cross-language experiments on spontaneous speech and to provide a preliminary benchmark. This

track was repeated again in 2006 and 2007 with an additional test collection of 357 interviews in Czech used for monolingual tasks. The resulting English and Czech collections are the first standard information retrieval test collections for spontaneous conversational speech. These collections are now publicly available through the Evaluations and Language resources Distribution Agency (ELDA) (Pecina et al. 2008).

In summary, cross-language speech retrieval combines the challenges of monolingual speech retrieval and textual CLIR as both errors in speech recognition and translation reduce the effectiveness of retrieval. When these technologies are combined it can be expected that these errors will be accumulative. A key problem can be the lack of resources – in particular speech recognisers and training data – when 'new' languages are to be handled. Problems can also arise when the language translation and speech recognition system are drawn from entirely different sources and have to be treated as black boxes in which the operating parameters cannot be modified.

### *6.2.3 Video Retrieval*

Video is a rich source of information, with aspects of content represented both visually and acoustically, using language-independent and language-dependent features. Multilingual video retrieval can be seen to a large extent as an integration and extension of technologies adopted for multilingual text retrieval, speech retrieval and image retrieval. Typically a video retrieval system will index a video according to two different channels. On the one hand, the audio speaker segmentation and the speech recogniser will provide a segmented textual representation of the speech content. On the other hand, the visual signal is processed as follows. First, shots are segmented using a visual segmentation algorithm, then relevant visual descriptors (e.g., colours, texture, and shape) are extracted and semantic concepts are derived on the basis of a fixed vocabulary or ontology[6] using image processing and classification techniques. Finally information from both audio and visual signals can be merged based on the audio segmentation to produce a semantic content representation. This representation can then be indexed (Jones et al. 2008). There is also a third potential source of information: text contained in the video frames. This can be detected, extracted and recognised using Optical Character Recognition (OCR) techniques. However, text detection and extraction in video frames is more problematic than text segmentation in digital documents due to complex background, unknown text colour, and degraded text quality caused by lossy compression, and different language characteristics (Lyu et al. 2005). Most attempts at multilingual or cross-language retrieval on video have

[6] Typical examples of such semantic concepts are objects such as 'car', persons such as 'Barack Obama', scenes such as 'sea', or events such as 'football match'.

concentrated on the language-dependent features, i.e., transcriptions of the spoken content and/or textual metadata.

One of the earliest of these efforts was the Multilingual Informedia project (Hauptmann et al. 1998). This system performed speech recognition on a collection of news broadcasts in English and Serbo-Croatian, segmenting them into stories and indexing the data. A keyword-based translation module transformed English queries into Serbo-Croatian, allowing a search for equivalent words in the joint corpus. Serbo-Croatian news broadcasts were tagged with English topic labels in order to assist users in judging the relevance of particular news clips returned as results.

An early European research effort was the ECHO (European CHronicles Online) project sponsored by the European Commission. ECHO developed a digital library service for historical films belonging to large national audiovisual archives in English, French, Dutch and Italian. The ECHO film archives consisted of language-dependent (speech, text) and language-independent (video) media. Thus, although users querying over collections in different languages might not understand the spoken dialogue, they could still identify useful documents (or parts of documents) via the images. All the videos in the collections had been supplied with rich metadata by the content providers. This facilitated the implementation of a relatively simple multilingual search interface. The approach adopted was to implement online cross-language search tools based on the use of standard metadata formats and mechanisms that provide a mapping between controlled vocabularies agreed between the content providers. Access is provided by local site interfaces in the local languages, but a common user interface in English is also maintained on the project website for external access (Savino and Peters 2004).

The MultiMatch project[7] has already been cited in previous chapters of this book. The aim of MultiMatch was to enable users to explore and interact with online internet-accessible cultural heritage content, across media types and language boundaries. This was achieved through the development of a search engine targeted on the access, organisation and personalised presentation of the data. The MultiMatch search engine combined automatic classification and extraction techniques with semantic web compliant encoding standards in order to facilitate search across languages. Instead of returning documents in isolation, MultiMatch provided complex search results that associated documents of various media types displaying them not as isolated individual items, but as richly connected entities. The MLIR/CLIR functionality provided focused on the problem of using a request in one language to retrieve documents from a collection in multiple languages. Textual data from each media source (full-text, metadata, speech transcriptions) was indexed so that it could be matched against queries entered by users.

The SemanticVox project is an example of a cross-language video indexing and retrieval system based on both speech transcription and video analysis. The system

[7] http://www.multimatch.eu/

manages French, English, Spanish, German and Arabic queries and documents. The audio and video signals are both processed as described above and the resulting information is merged to produce a semantic content XML-based representation which is then indexed (Delezoide and Le Borgne 2007). Retrieval is via textual queries in any of the five languages.

Research in video retrieval has been promoted by the TRECVid benchmarking activity which began as a track within TREC in 2001 but then expanded to become an independent initiative. Work has focused on three kinds of search (automatic, manual and interactive), shot boundary detection, detection of concepts or high-level features within shots, and automatic video summarisation (Smeaton et al. 2009). In 2008 CLEF offered a video track with the goal of studying ways to analyse multilingual video content. The focus was on dual language video, i.e., videos in which two languages are spoken but the content of one does not duplicate the other. The video data was derived from interviews and studio discussions on Dutch television and was provided by the Netherlands Institute of Sound and Vision. The extensive use of English and other languages mixed with Dutch on Dutch television programmes means that they are a rich source for this type of data. One task involved the classification of Dutch-language documentaries having embedded English content arising from interviews and discussions with non-Dutch speakers. Task participants were supplied with Dutch archival metadata, Dutch speech transcripts, English speech transcripts and ten thematic category labels, which they were required to assign to the test set videos. Participants collected their own training data. Results were delivered in the form of a series of Really Simple Syndication (RSS)-feeds, one for each category. Feed generation, intended to promote visualisation, involved simple concatenation of existing feed items (title, description, keyframe). A favourite strategy was to collect data from Wikipedia or to employ a general search engine to train classifiers (Support Vector Machine (SVM), Naive Bayes and k-Nearest Neighbour (k-NN) were used). A competitive approach was to treat the problem as an information retrieval task, with the class label as the query and the test set as the corpus. Both the Dutch speech transcripts and the archival metadata performed well as sources of indexing features, but no group succeeded in exploiting combinations of feature sources to significantly enhance performance. In addition to the main classification task, which was mandatory, VideoCLEF offered two discretionary tasks. The first was a translation task, requiring translation of the topic-based feeds from Dutch into a target language. The second was a keyframe extraction task, requiring selection of a semantically appropriate keyframe to represent the video from among a set of keyframes (one per shot) supplied with the test data (Larson et al. 2009).

VideoCLEF was repeated in 2009 using the same Dutch television data. Three tasks were offered. The first involved the automatic tagging of videos with the subject theme labels, and the second task concerned detecting narrative peaks in short-form documentaries. The third task focused on the multilinguality of the data and involved linking video to material on the same subject in a different language. Participants were provided with a list of multimedia anchors (short video segments) in the Dutch-language 'Beeldenstorm' collection and were expected to return target

pages drawn from English-language Wikipedia. The best performing methods used the transcript of the speech spoken during the multimedia anchor to build a query to search an index of the Dutch-language Wikipedia. The Dutch Wikipedia pages returned were used to identify related English pages (Larson et al. 2010).

In 2010 VideoCLEF became MediaEval, an independent benchmarking initiative dedicated in general to evaluating new algorithms for multimedia access and retrieval rather than concentrating specifically on multilingual issues. MediaEval extends the VideoCLEF focus on speech to the full range of aspects that go beyond the visual channel of video, by including users and context.

So far, most efforts at multilingual retrieval on video data exploit available textual data. The difficulties to be addressed are thus similar to those described for multilingual speech retrieval in the previous section. The quality of both the audio source and the speech recogniser employed will impact on retrieval performance and the results of a cross-language search are dependent on the coverage of the translation resources employed. In order to overcome these problems, multilingual video retrieval systems should aim at merging the information extracted from audio data with additional conceptual information derived from the language-independent visual features. The results of an initial spoken or written query can be improved or extended using relevance feedback mechanisms exploiting the visual descriptors. Recent research has shown that content-based features can improve traditional text search on transcripts and that visual examples contribute slightly more than concept queries (Rautiainen et al. 2006).

### *6.2.4 Question Answering*

The document retrieval systems discussed in Chapters 2 and 3 of this book will typically return ranked lists in response to a user's information need expressed in natural language or as keywords. The goal of Question-Answering (QA) systems, on the other hand, is to provide direct answers (rather than documents containing relevant information) to questions that users pose to the system. More precisely, "*To answer a question, a system must analyse the question, perhaps in the context of some ongoing transaction; it must find one or more answers by consulting on-line resources; it must present the answer to the user in some appropriate form, perhaps with some justification or supporting materials*" (Hirschman and Gaizauskas 2001). Some QA systems work on specific domains and are known as *closed-domain*; other systems are domain-independent and known as *open-domain*. Developing open-domain systems is much harder than developing domain-specific ones as it involves dealing with natural language texts on different topics and written in varying styles. For this reason, much of the recent research in this area has focused on open-domain systems (e.g., the question answering tracks at TREC, NTCIR and CLEF and systems designed for question answering on the Web).

Compared to document retrieval, question answering generally relies more heavily on using techniques from natural language processing and a wide variety

of linguistic resources as questions are usually classified into particular types before attempting to find and extract documents from a target collection and this requires the application of syntactic and/or semantic analysis procedures.[8] A number of question types can be distinguished. Examples include: fact-based or *factoids* (e.g., 'What was the name of the first Russian astronaut to do a spacewalk?'), *opinions* (e.g., 'What do people think about Manchester United?'), *definitions* (e.g., 'What is an atom?') and *summaries* (e.g., 'What are arguments for and against capital punishment'). There are also *Yes*/*No* questions (e.g., 'Was Obama born in the United States of America?'). Some questions require single exact answers (e.g., 'Who wrote Harry Potter and the Half-Blood Prince?'); others may have multiple distinct answers, commonly known as *list* questions (e.g., 'Name all the airports in London, England'). Some questions require simple short answers; others may need longer and more verbose answers that may be generated by combining multiple sentences from multiple documents.

The architecture of a typical open-domain question-answering system consists of three main modules: question analysis/processing, document or passage retrieval and answer extraction. A generic architecture for a typical question-answering system is shown in Figure 6.4.

The purpose of the *question analysis*/*processing* module is to parse the natural language question posed by the user and determine the type of semantic entity expected in the answer. For example, in the factoid question 'Where was Mozart born?' the type of entity expected as an answer would be a location (as indicated by the question word 'where'), whereas with 'Who was the first man to walk on the moon?', 'who' indicates that the expected answer is a person's name. However, determining the semantic type of answer entities can often be difficult due to the

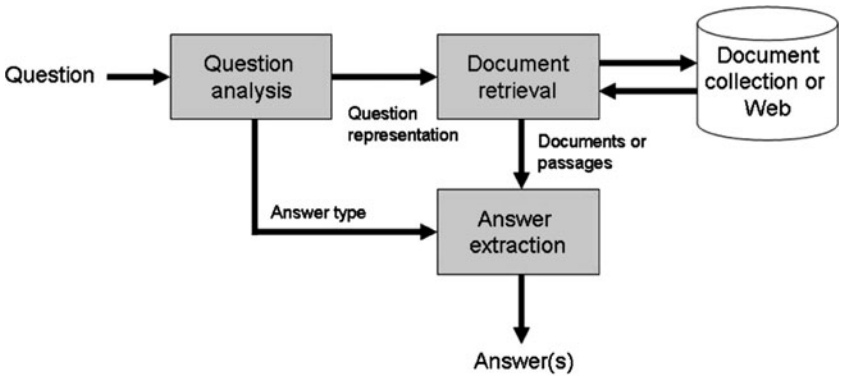


**Fig. 6.4** Generic architecture for a question-answering system

[8] A notable exception are the experiments at Microsoft Research which capitalise on the redundancy of information available on the Web as their knowledge base and use simple query transformations, passage ranking and *n*-gram extraction techniques (Dumais et al. 2002).

way in which the syntax of questions can vary, for example in the question 'What famous communist leader died in Mexico city?' the answer type is a person's name, while with 'Who is Nelson Mandela?' the expected answer will provide some biographical information concerning the ex-President of South Africa and Nobel prize winner. Techniques commonly employed in the question analysis stage include part-of-speech tagging, syntactic parsing, word sense disambiguation and keyword extraction. These help to deal with issues such as the same question being asked in multiple ways, e.g., 'What is Mozart's date of birth?' is equivalent to 'When was Mozart born?' and complex questions, e.g., 'How far is the capital of England from New York?' which could be expressed as two successive questions 'What is the capital of England' and 'How far is London from New York?' Systems normally also refer to knowledge bases, taxonomies or ontologies to find necessary external knowledge. For example, in the 'communist leader' question, an ontology will reveal that 'leader' is-a 'person'.

The *document retrieval* module (sometimes called *candidate document selection*) will typically be an implementation of an IR system and will return documents that are likely to contain answers to the question. It is common to also find passage retrieval utilised at this stage to reduce the document size to fixed-length text segments (or sentences/paragraphs) that contain terms from the search string as provided by the question analysis/processing module (sometimes called *candidate document analysis*). For example, the search system can be set to retrieve paragraphs or single sentences rather than whole documents.

The final module, *answer extraction*, is used to extract candidate answers from the documents or passages that have been retrieved from the document (and passage) retrieval stage(s). The candidate answers are often ranked according to their probability of correctness. This final stage will often include named entity recognition to identify, e.g., names of people, locations, dates, times and organisations, which can then be matched against the expected answer type. Answer extraction will also include *answer validation* – given a candidate answer, determine whether it is correct for the question posed or not. Of course, it is not always possible for the system to find an answer to a question, and some questions may not have an answer (e.g., 'Who will be the first person on Mars'). In such cases, the correct response would be 'I don't know', or an equivalent phrase.

Question-answering in languages other than English presents challenges similar to non-English document retrieval as discussed in Chapter 2 – representing, storing and managing different character sets (and scripts) for different languages, and tokenising and reducing terms to root forms. However, there are also significant additional problems due to the reliance on natural language processing, such as identifying named entities in multiple languages and developing effective taggers, syntactic parsers and sense disambiguators for text written in languages other than English (Katz et al. 2007). In cross-language question-answering one assumes that answers to a question may be found in languages different to the language of the question (Magnini et al. 2006). Approaches to cross-language question–answering are similar to cross-language document retrieval: translation of the question (i.e., query translation) or translation of the document collection (document translation).

In addition, depending on the user's needs, the system may also translate the answers into the user's source language. The approaches described in Chapter 3 for translating queries and documents are commonly employed. Clearly, a translation module has to be added to the architecture outlined above.

A major part of the research activity on question answering systems has been carried out in the context of global evaluation initiatives. The evaluation of monolingual question-answering systems targeting document collections in English was carried out on a large scale at TREC[9] between 1999 and 2007 – systems were asked to retrieve small snippets of text that contained an answer for open-domain, closed-class questions (i.e., fact-based, short-answer questions that can be drawn from any domain) (Voorhees and Tice 2000, Voorhees 2001). By the end of this period the best systems were achieving very good results. In 2008 this activity was continued at the Text Analysis Conference[10] which offered a more challenging opinion question answering track.

Following in the steps of TREC, both NTCIR[11] and CLEF[12] introduced question-answering tracks – NTCIR in 2001 and CLEF in 2003. The aim of both initiatives was to encourage groups to work on the development of QA systems in their own languages and also to investigate to what extent language-dependent factors could impact on system performance. Almost all the studies of multilingual question answering systems so far have been within the context of these two campaigns. Both have gradually increased the complexity of the tasks offered over the years.

NTCIR began by offering monolingual tasks for Japanese systems in NTCIR-3 and NTCIR-4. The main objective was to encourage the development of open domain QA systems with particular focus on research into user interaction and information extraction (Nomoto et al. 2004). This activity continued in NTCIR-5 and NTCIR-6 in the QAC tracks (Question Answering Challenge) with a more complex task in which special attention was given to the correct interpretation of questions within a dialogue, and systems were tested for their context processing abilities, such as anaphora resolution and ellipses handling for Japanese.[13] Systems were requested to answer a series of related questions via a simulated interaction. Each question, apart from the first, contained some kind of anaphoric expression. The systems were requested to return a single list containing correct answers (Fukimoto et al. 2007).

NTCIR-5 also offered a multilingual track QA track: the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1), with both monolingual (Chinese)

[9] http://trec.nist.gov/data/qa.html

[10] http://www.nist.gov/tac/

[11] http://research.nii.ac.jp/ntcir/

[12] http://www.clef-campaign.org/

[13] Anaphora (use of pronoun to refer to a preceding word or group of words) and ellipsis (the omission of one or more words from a sentence that are understood implicitly) are a common feature in natural languages.

and cross-language (Japanese-English and English-Japanese, Chinese-English and English-Chinese) sub-tasks. The target collections consisted of a 2-year (2000–2001) collection of news articles for each language. In order to reduce the complexity of the tasks, only questions about named entities were offered: person, location, organisation, artifact, date, time, money, and percent. This track was repeated in NTCIR-6, with the addition of Japanese and English monolingual tasks. In both campaigns, the performance of the cross-language QA systems was heavily degraded with respect to the monolingual systems – showing clearly that question answering over language boundaries is a complex activity. In NTCIR-6 the results of the best English to Japanese system were a little over 50% of the best Japanese monolingual performance while the best Chinese to English results were approximately 60% of Chinese monolingual performance (Sasaki et al. 2007).

The multilingual QA tasks at NTCIR-7 and NTCIR-8 were considerably tougher. Monolingual and cross-language tasks were offered for three target collections of news documents: Japanese, Simplified Chinese and Traditional Chinese. The goal was to test systems with complex questions (i.e., events, biographies/definitions, and relationships) as well as factoid questions. Both end-to-end and module-based evaluations for question type analysis, document retrieval and answer extraction were conducted (Mitamura et al. 2010). The performance of the cross-language systems was again measured against monolingual baselines with somewhat more encouraging results compared to those of previous years. For example, for the English to Chinese task one of the best groups reported that while in NTCIR-5 their best run only achieved 35% of monolingual retrieval effectiveness, in NTCIR-8 their cross-language result was almost 80% of monolingual performance (Min et al. 2010).

Evaluation of multilingual question-answering began at CLEF[14] in 2003 and has been important in encouraging the development of monolingual and cross-language QA systems for many European languages. Between 2003 and 2009, target collections were used for mono- and cross-language tasks in 11 European languages, with the consequent adoption and adaptation or development of the appropriate language processing tools. During this period, the main task was mostly focused on factoid, definition and closed list questions. It was made more challenging each year as different types of questions were proposed and different kinds of answer formats, ranging from paragraphs and snippets to exact answers, were required.

The main interest of the systems participating in the QA track at CLEF has always been in the monolingual tasks, with most teams concentrating on building a good monolingual system in their own language. The result is that there are now several QA research groups in many European countries. Since 2003 the performance of monolingual non-English systems at CLEF has improved considerably, to the extent that the best systems are now achieving results similar or close to those of the best English systems. Another important output has been the building of

[14] http://celct.fbk.eu/QA4MRE/index.php?page=Pages/pastCampaigns.php

multilingual test sets with their associated gold standard answers and document collections This material is available online allowing groups in future to reuse the data produced in order to develop and tune their systems (Forner et al. 2010).

The CLEF activity has also encouraged research into various distinct aspects of question answering in a multilingual context through the setting up of several pilot tasks. These tasks studied answer validation (Rodrigo et al. 2009), question answering on speech transcriptions (Turmo et al. 2009), question answering in real time (Noguera et al. 2007), and how to handle geographic questions (Santos and Cabral 2010). The goal of most of these experiments has been to attempt in-depth investigations of specific areas in order to have a better understanding of the role they play and the impact they can have on system performance as a whole. In particular, it has been shown that optimisation of the answer validation component can improve overall results significantly.

With respect to cross-language experimentation, although many possible candidate pairs of languages have been offered by CLEF, most teams have focused on experimenting with systems querying between their own language and English. The findings have been similar to those of NTCIR; the cross-language systems have not shown an improvement over the years comparable to that of monolingual ones. This finding differs from the findings for both campaigns with respect to cross-language document retrieval. The main reason seems to be the inadequacy of the translation mechanisms adopted to handle the more challenging demands of question answering. In particular, named entities, typically the focus of natural language questions, tend to pose a real problem; they are generally not included on most machine-readable dictionaries and their coverage is usually poor in machine translation systems. In their overview of the results of 7 years of multilingual question answering at CLEF, Forner et al. (2010) conclude that in order to improve performance for cross-language question answering, it is principally the translation component that needs to be strengthened.

At the moment it is not easy to provide useful information with respect to best practices in cross-language question answering, more research is needed. It does appear that the amount of training data available is important. Most QA systems in the literature use supervised-learning methods to train their models, and therefore the amount of training data has a significant impact on the system's performance. This is true for any system. For cross-language systems, it seems that focusing attention on managing named entity extraction and matching over languages and on improving strategies for answer validation can help considerably in improving performance.

## 6.3 Multilingual Information Access in Practice

Chapter 1 described a number of practical situations in which users are faced with the necessity of querying multilingual document collections in diverse media and where the availability of appropriate retrieval functionality technology would be

beneficial. In this section we describe a number of domains in which issues related to multiple languages arise and for which multilingual information access technologies are applicable. These include web search (Section 6.3.1), digital libraries and cultural heritage (Section 6.3.2), medicine and healthcare (Section 6.3.3), government and law (Section 6.3.4) and business and commerce (Section 6.3.5). The aim is to provide a sense of the diversity of areas in which the multilingual information access techniques discussed in this book are being used in practice. Areas of interest that have not been discussed here include cross-language plagiarism detection (Potthast et al. 2011), multilingual information access to user generated content, e.g., Wikipedia (Hecht and Gergle 2010, Udupa and Khapra 2010), and the use of cross-language search in learning and teaching contexts (Maeda 2004).

### *6.3.1 Web Search*

The rapid and continual growth of information and services on the Internet has popularised web search engines to the point that many web users begin their online activities by submitting a query to a search engine, such as Bing, Google and Yahoo!. Due to globalisation the textual content accessible to users on the Web is now available in multiple languages (see the discussion in Chapter 1) and subsequently search engines have evolved to support users with the implementation of multiple language features.[15]

Some web search engines have been specifically developed for users in a given country. For example, Baidu.com (shown in Figure 6.5) is a Chinese search engine that gathers and serves online content in a way that adheres to government regulations with respect to information dissemination in addition to supporting language-specific functions, such as the input of Chinese terms without the use of a Chinese keyboard. Ajeeb, on the other hand, is a search engine specifically developed for Arabic content and users. Larger search engines, such as Google, offer regional versions where the interface and results will be localised to the given region.[16] For example, searching the French version of Google (google.fr) will return links to documents that are written in French and offer a translation option to non-French documents; the interface is localised to French (and for cross-language

[15] Chau et al. (2008) describe in detail the implementation of a toolkit called SpidersRUs for developing multilingual search engines that deal with issues such as support for multi-byte characters, language identification and the parsing of documents in multiple languages.

[16] Some countries have multiple languages, e.g., Belgium has Flemish, French and German as official languages; Canada has English and French. Therefore, in the case of Google Canada (google.ca) the interface provides an option to switch between English and French and web pages in both languages are returned in the results by default. The Indian Union is an example of a very complex case; although the official language is Hindi, more than 20 languages actually have some official status.


**Fig. 6.5** Baidu – China's leading search engine (http://www.baidu.com). The interface includes language-specific functionality, such as allowing users to enter Chinese characters directly. This screenshot is copyright of Baidu

searches the source language is initially assumed to be the same as the interface language) and advertising is targeted at France.

A number of issues must be considered in non-English web retrieval with respect to indexing (language identification, word segmentation, stopword removal, stemming and lemmatisation) and searching. An analysis of query logs from multiple search engines across languages has shown that there is a variation between the queries formulated by searchers from different countries with respect to query length and morphology. A lack of consistency has been noted with some search engines when handling 'identical' queries across languages, e.g., query terms with and without the correct diacritics (Lazarinis et al. 2009). This can confuse the user.

Several studies have also compared the functionalities provided by search engines and web portals in different languages and for different regions. For example, Zhang and Lin (2007) investigated multiple language features in 21 internet search engines, including standard search engines (e.g., Google, MSN and Yahoo!), metasearch[17] engines (e.g., WebCrawler, EZ2Find and Hotbot) and visualisation search engines (e.g., Kartoo[18]). They employed five evaluation criteria for assessing multiple language support in a search engine: the number of languages a search engine can search and process, the visibility of multiple translation features support, translation ability, help file quality and interface design. Overall Google

[17] The main difference between metasearch and standard search engines, such as Google or Yahoo!, is that the former do not crawl the Web or maintain their own indexes. Instead they query multiple search engines (often in parallel), collate results and present them in a unified fashion to the user (Meng et al. 2002).

[18] No longer operational.

was rated as the top standard search engine providing features to support multiple languages; EZ2Find[19] was the top-rated metasearch engine. However, despite the provision of functionality to support multiple languages and the obvious benefits of offering cross-language options, very few search engines and portals provide such functionality.

Developing an effective multilingual web search engine is undeniably difficult as one not only has to deal with indexing various forms of content written in numerous ways, but must also support multiple character sets and encodings of varying standards and levels of quality. Pingali et al. (2006) describe a search engine for Indian languages called Webkhoj, which highlights a number of challenges, such as dealing with multiple scripts and transliteration of the multiple encodings found on the Web for Indian language content. More recently, in 2010, Google integrated cross-language functionality into their standard search interface (see Section 4.5.3).

### *6.3.2 Digital Libraries and Cultural Heritage*

One definition of a digital library is: "*A focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection*" (Witten and Bainbridge 2003). Digital libraries are amassing an increasing amount of digital content and information that must be made accessible and usable. This includes significant amounts of cultural heritage material offered by institutions, such as libraries and museums, through digital library portals and transcending national and linguistic boundaries. The result is that users and curators of digital libraries are being confronted with large and diverse quantities of digital materials that are multimedia, multi-cultural and multi-language. A fundamental goal of digital libraries is to provide universal access to the information being managed but this can only be realised if digital content is made more accessible and usable over time within online environments. Although digital libraries remove physical and spatial barriers in accessing information, the language barrier still remains due to the fact that many collections are in multiple languages and because of the linguistic diversity of users. For example, according to a report in 2010 by the Online Computer Library Centre (OCLC)[20] on WorldCat,[21] their global online library catalogue, nearly 197 million records for library items in 479 languages from more than 17,000 libraries in 52 countries are accessible to users. More than 57% of records in WorldCat are written in languages other than English.

[19] This metasearch engine has now ceased to exist (last accessed: 14/01/2011).

[20] http://www.oclc.org/

[21] http://www.oclc.org/news/publications/annualreports/2010/2010.pdf

Supporting MLIR and CLIR in digital libraries has long been recognised as important in providing universal access to digital content. This has been confirmed by many different types of user studies. For example, a study by Marlow et al. (2007) to investigate the potential need for multilingual access to the Tate Online, one of the UK's largest cultural heritage sites, found 31% of visitors came from outside the UK and their preferred language for searching and navigating the site was not English. Results from an online survey indicated that the provision of multilingual access would be welcomed by visitors who would prefer to interact with English content from Tate Online in their native (non-English) languages. The effects of differences in culture and in language skills have also been evidenced by studies by Bilal and Bachir (2007) on the usability of the International Children's Digital Library with Arabic-speaking children, and by Duncker (2002) who described the cultural issues of digital libraries for the Maori people. Pavani (2001) focused on usability issues of the Maxwell multilingual digital library in Brazil. Wu et al. (2010) studied the use of multilingual information access by students in Chinese academic libraries. Their findings highlighted the wide use of multilingual resources by Chinese students (most of which were in English and difficult for them to use) and the need for specialist technologies and translation resources for this domain (e.g., clustering and translation of abstracts or full-text for document examination). Clough and Eleta (2010) showed the potential benefit of multilingual information access for international students studying abroad, particularly in University departments such as Languages, Linguistics, Translation, Interpretation, Social Science, Anthropology, Politics and International Relations.

However, despite a number of important research prototypes,[22] the reality is that most digital libraries still do not provide multilingual access to their content. For example, an analysis by Chen and Bao (2009) of around 150 US digital library sites revealed that only five (or 3%) could be accessed using more than one language and no site employed cross-language information retrieval or utilised machine translation.

The situation in Europe is only slightly better. A fairly rare example of a digital library portal that does provide cross-language retrieval of library items through the OPAC (Online Public Access Catalogue) is the University Library from the Free University of Bozen in Italy. The site offers a multilingual search feature (shown in Figure 6.6) that enables users to submit queries in English, German and Italian. Queries are translated into all target languages (English, German and Italian) and results are grouped according to language (Bernardi et al. 2006). However,

[22] For example, Rydberg-Cox (2005) describes work by the Cultural Heritage Language Technologies Consortium to apply language technologies (including CLIR) to assist scholars and students working with texts written in Greek, Latin and Old Norse. The challenges involved when processing natural language in historical texts is also discussed by Koolen et al. (2006) who present a cross-language approach to historic document retrieval involving seventeenth century Dutch documents.

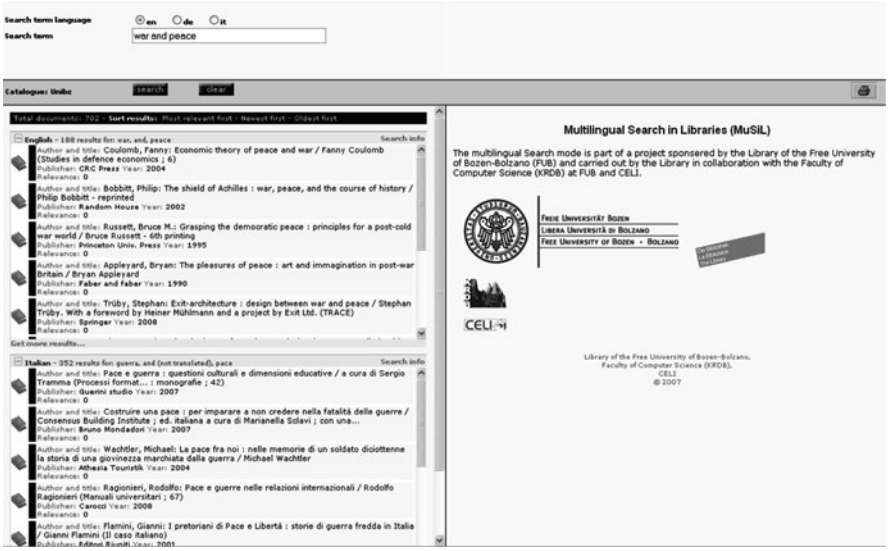


**Fig. 6.6** Cross-language retrieval functionality at the Free University of Bozen Library (http://www.unibz.it/en/library/welcome/default.html)

compared with the research prototypes discussed in Chapter 4, this multilingual search interface is basic: it does not allow interactive query translation or involve translation at document selection and examination stages.

In Chapter 1 we cited two of the most important European digital libraries: The European Library (TEL),[23] which offers free access to the bibliographical resources of many of Europe's national libraries, and Europeana,[24] which provides online access to much of Europe's cultural heritage. So far, both libraries provide multilingual interfaces but no cross-language search functionality. In 2008–2009, CLEF sponsored a set of experiments aimed at investigating retrieval of bibliographic records from The European Library through tasks testing intra-collection and cross-collection search on library catalogues derived from the British Library, the Bibliothéque nationale de France and the Austrian National Library. The aim of the exercise was to identify the most effective retrieval technologies for searching this type of sparse and inherently multilingual data, with particular focus on cross-language search, in order to provide useful feedback for the developers of both TEL and Europeana.

The data collection consisted of around three million records with free-text metadata consisting mostly of titles, subjects, and abstracts. In addition to the expected English, French and German, each collection contained records in many other languages. The tasks presumed a searcher, with a working knowledge of the

[23] http://www.theeuropeanlibrary.org/

[24] http://www.europeana.eu/

three main languages plus Spanish, who wants to find useful documents in any one of these four languages in any of the target catalogues. Queries were prepared in a number of languages including Chinese, Greek and Farsi. To the best of our knowledge, this is the only study of its kind to date.

Twenty-one groups participated in the experiments over the two year period, although only five groups participated both years. All the traditional approaches to mono- and cross-language retrieval described in Chapters 2 and 3 of this book were attempted: retrieval methods included language models, vector-space and probabilistic approaches; translation resources ranged from bilingual dictionaries, parallel and comparable corpora (including Wikipedia) and machine translation. Participants often used a combination of resources. Some groups attempted to address the multilinguality of the target collections in a systematic fashion, for example by using language detectors to identify the actual language of each record and then creating separate indexes for each language and applying language-specific stemmers, finally using some kind of fusion method to combine the separate partial results. It is noteworthy that the MLIR/CLIR technology already tried and tested for free-text document retrieval performed equally well on this type of semi-structured data with the best systems reporting state-of-the-art cross-language or cross-collection performance of well over 90% of the best monolingual or intra-collection runs (Ferro and Peters 2010). These research efforts have provided useful feedback for both applications, proving that the necessary MLIR technology is now in place and can produce good results.

Europeana is building on the experience of TEL. It has the ambitious goal of making Europe's cultural and scientific heritage accessible to the public and counts on the support of over 180 heritage and knowledge organisations and IT experts across Europe. In early 2011 Europeana contained well over 15 million digitised cultural objects and this number was increasing fast. The plan is to implement multilingual and cross-language functionality to access its multimedia archives. The first stage – multilingual interfaces for 28 European languages – has already been achieved.[25] The intention for multilingual search is to create language-dependent indexes. For this it is necessary to know the language of the documents (both metadata and contents). However, many of the Europeana records do not contain more than title and subject information and language detection on text of five words or less is still a challenge. Several studies are now ongoing with respect to the implementation of cross-language functions: both query and document translation are being experimented with. The query translation prototype is dictionary-based, with a lot of linguistic processing to cater for translation ambiguities (e.g., tokenisation, part-of-speech detection, lemmatisation, decompounding). Named entities are detected and not translated; multi-word phrases are detected and translated as such. A prototype for ten languages, both as source and target languages (English, German, French, Italian, Spanish, Polish, Swedish, Dutch,

[25] Screenshots of Europeana interfaces showing the multilingual function are given in Chapter 4, Figures 4.15 and 4.16.

Portugese, Hungarian) is being tested. All possible language pairs are catered for. At the same time, the Europeana office is experimenting with document translation and in January 2011 implemented a prototype with Bing Translator and Google Translate. The documents are translated from their source language to all possible target languages offered by Bing and Google. So far, all these studies are in the experimental phase. At the time of writing (mid-2011), no date has been established for producing a beta version to be made public.

The example of Europeana helps to explain the current lack of take-up of multilingual information access in the digital library world. The implementation of full cross-language service in a large-scale general domain digital library with collections in many languages is a very challenging endeavour as many language-specific processing tools and translation resources are needed. It is clear that this raises the complexity and the costs of the system considerably.

### *6.3.3 Medicine and Healthcare*

The medical/healthcare domain provides numerous opportunities for the application of multilingual information access (Lu et al. 2008, Kahn 2009). An example is the need to provide universal access to medical research literature found in large databases, such as PubMed[26] and Medline,[27] which contain English-only material making them difficult for non-English users to search, particularly when unfamiliar with medical terminology even in their native language.

This is not only the case for published academic research; the problem extends to finding public health and medical information online and conversing with medical professionals (Saha and Fernandez 2007). A recent study by Cleveland et al. (2008) showed that language is a serious barrier for Chinese communities in the Dallas-Fort Worth area in Texas trying to find and use quality online medical information, mostly published in English. Tran (2009) also showed that the barrier of limited English proficiency is one of the factors leading to increased outbreaks of the hepatitis B virus among the Asian American population in the US compared to English speaking citizens.[28]

Another example is the tourist on holiday abroad in countries where communication is impossible due to language barriers. Of course not all information is written in English; physicians will usually write notes regarding patient encounters in their native language.

In all of these cases the use of CLIR and MT could alleviate some of the problems with accessing information, particularly the problem of non-native

[26] http://pubmed.gov/

[27] http://www.nlm.nih.gov/databases/databases_medline.html

[28] This is not just a problem within the US; most countries have migrant populations that cause problems in providing adequate health services (http://www.bbc.co.uk/news/health-10951417).

English speakers accessing medical content only available in English. A common technique is to use a multilingual medical thesaurus or dictionary and this approach has been applied in many Cross-Language Medical Information Retrieval (CLMIR) applications. For example, Hersh and Donohoe (1998) describe enhancements to a concept-based medical retrieval system called SAPHIRE that maps terms from documents and queries to concepts in a resource called the Unified Medical Language System (UMLS) Metathesaurus which is derived from MeSH[29] (Medical Subject Headings), provided by the National Library of Medicine (NLM). The multilingual thesauri are then used to provide cross-language access to clinically-oriented web pages written in English (and assigned terms from MeSH) using queries in English, German, French, Russian, Spanish and Portuguese.

Liu et al. (2006) created BabelMeSH,[30] a cross-language retrieval interface to NLM databases, such as Medline, to enable users (e.g., healthcare providers or researchers) to express queries in their native language: Arabic, Chinese, French, German, Italian, Japanese, Portuguese, Russian and Spanish. This is available as a translation service and can be used in further CLMIR applications. Kahn (2009) used the BabelMeSH service to provide cross-language retrieval to the American Roentgen Ray Society (ARRS) Goldminer system, an indexed collection of over 200,000 radiological images with unstructured English captions. Lu et al. (2008) also make use of MeSH but show how exploiting multilingual web resources can be used to create a Chinese-English version of MeSH and allow retrieval of medical information (typically expressed only in English) in Chinese.

The importance of the medical area for CLIR applications has encouraged CLEF to include cross-language retrieval, classification and annotation of medical images as tasks since 2004. Datasets of images from radiology, pathology, endoscopy and nuclear medicine have been created in order to provide a number of re-usable benchmarks for evaluating medical retrieval systems (Hersh et al. 2006, Müller and Kalpathy-Cramer 2010). The CLEF results have shown the effectiveness of MLIR/CLIR techniques for medical images.

### *6.3.4 Government and Law*

There is considerable scope for applying multilingual information access technologies to legal information: in many cases legal discourse transcends national and language boundaries and there is an increasing requirement to make legal information accessible to all citizens within a country regardless of language skills (Peruginelli 2008). Sheridan et al. (1997b) discuss cross-language information retrieval in a multilingual legal domain in Switzerland and state: "*In many cases*

[29] http://www.nlm.nih.gov/mesh/

[30] http://babelmesh.nlm.nih.gov/

*important information is provided in all three majority languages, French, German and Italian, as is the case with Swiss federal law. In other cases, however, information which is of importance throughout the country is available only in one of the languages... Lawyers wishing to refer to decisions of the federal court however must have access to all decisions of the court, irrespective of the language in which they are recorded. Allowing Swiss lawyers to search all decisions of the federal court of justice in their own language is a perfect application for cross-language retrieval*."

Further examples include legal professionals wishing to compare the similarities and differences of legal systems across different countries (comparative law); corporations searching global databases for existing patents to avoid possible infringements; the general public accessing administrative documents held by government departments. Within the European Union (EU), the European Parliament[31] translates *all* parliamentary documents into the official EU languages as members of parliament have the right to speak in any official language of their choice but the information should be accessible to all citizens of the EU regardless of language. However, cross-language technologies are not utilised as parallel versions of the website are produced and accessible in each of the official EU languages.[32] Other international institutions, such as the United Nations and governments of multilingual countries (e.g., Belgium, Switzerland, Canada, India, Singapore) may also generate versions of the same document in multiple languages.

Providing multilingual access to legal documents is particularly challenging due to the complexities of legal language and differences in the meaning of legal concepts across cultures and languages. This motivates the need for translation resources developed especially for the legal domain. Sagri and Tiscornia (2004) describe the creation of multilingual semantic lexicons and thesauri for the legal domain where legal concepts are described and mapped between languages. A common approach has been to extend existing lexical resources, such as WordNet, with legal terminology and connections between languages. The EU-funded Lexical Ontologies for legal Information Sharing (LOIS) project, for example, connects legal WordNet in six languages to provide a multilingual legal knowledge base (on consumer law) that can be utilised for CLIR and MLIR (Peters et al. 2007, Dini et al. 2005).

Alternatively translation resources can be derived from parallel and comparable corpora and used within MT systems. For example, because documents from the

[31] http://www.europarl.europa.eu/

[32] To achieve this level of parallelism involves a large amount of manual effort and access to high quality translation resources *"To produce the different language versions of its written documents and to correspond with citizens in all the EU languages, the European Parliament maintains an in-house translation service able to meet its quality requirements and to work to the tight deadlines imposed by parliamentary procedures. It also has recourse to freelance professional translators for non-priority texts."* http://www.europarl.europa.eu/parliament/public/staticDisplay.do?language=EN&id=155

European Parliament are translated into multiple languages this provides a parallel collection of documents that can be used to derive cross-language similarity lexicons and bilingual word lists. The European Parliament Proceedings Parallel (Europarl) Corpus[33] is a collection of texts gathered from the European Parliament and offers parallel aligned corpora covering 11 of the official EU languages (10 languages aligned to English).[34] These documents have been used to train statistical MT systems (Koehn 2005). In a similar way, the EU Joint Research Centre (JRC) has produced the JRC-Acquis corpus,[35] approximately 464,000 documents covering the 22 official languages of the EU. JRC-Acquis consists of aligned parallel corpora constituting the Acquis Communautaire (AC), the total body of European Union law (legislative texts) applicable in the EU Member States (Steinberger et al. 2006). Sheridan et al. (1997b) also show how cross-language similarity thesauri can be successfully derived from collections of parallel and comparable documents for information retrieval within the legal domain.

Significant research efforts have also been carried out in the area of multilingual information access to intellectual property documents, such as patents. A Patent Retrieval Task has been conducted at NTCIR since 2001 addressing different problems in patent retrieval. For example, the task of carrying out a 'technology survey' consists of finding patents related to a specific technology; an 'invalidity search' task searches a collection of existing patents to find any that could invalidate a given patent application (Fujii et al. 2007). In the third NTCIR workshop (NTCIR-3) a technology survey task was addressed where search topics, available in five languages (Japanese, English, Korean, simplified/traditional Chinese), were used to search a document collection consisting of 5 years of Japanese patents (1993–1997) and English translations of the abstracts. In the NTCIR-4 Patent Retrieval Task, the problem of invalidity search was addressed. Search tasks were available in Japanese, English and simplified Chinese to enable investigation of cross-language retrieval (Fujii et al. 2004). Research on patent retrieval has continued at NTCIR and involved further tasks such as passage retrieval and classification (NTCIR-5 and NTCIR-6), machine translation of patent documents (NTCIR-7), and patent mining to create technical trend maps (NTCIR-8).

In 2009 a patent task, referred to as CLEF-IP,[36] was included in CLEF (Roda et al. 2010). The goal of CLEF-IP was to establish the patentability of a specific invention by attempting to find existing records in any language of similar or identical products that would constitute prior art for a given topic (a patent). A similar task was also run in 2010 but with a classification task added to the

[33] http://www.statmt.org/europarl/

[34] Similar parallel resources exist outside of Europe, for example the Canadian Hansards with parliamentary debates in English and French; the Hong Kong Hansards in English, Mandarin and Chinese; UN parallel texts in English, French and Spanish (see the Linguistic Data Consortium for further examples: http://www.ldc.upenn.edu)

[35] http://langtech.jrc.it/JRC-Acquis.html

[36] http://www.ir-facility.org/clef-ip

prior art search task. The data collection consisted of approximately 1.9 million XML documents in 2009 and over 2.6 million documents in 2010. The documents were derived from European Patent Office (EPO) sources with parts of the documents in the collection available in English, French and German (the same languages used for describing topics). Realistic patent search tasks at both NTCIR and CLEF have stimulated development of cross-language retrieval approaches for patent search. The domains of government and law will remain important areas for applying CLIR and MLIR techniques. As Sagri and Tiscornia (2004) state "... *corporate users, citizens, but especially professional organisations will benefit from crosslanguage linkages in order to retrieval legal documents from different European countries*."

### 6.3.5 Business and Commerce

As businesses seek to respond to the effects of globalisation, accommodating people with different language skills is becoming increasingly important for a variety of situations: exchanging knowledge and sharing resources within a multinational organisation; tapping into wider markets within e-commerce applications, optimising websites for search (multilingual Search Engine Optimisation or SEO); to widen the coverage of business and market intelligence activities; providing multilingual enterprise search within the organisation; and multilingual content management (Peterson 2002). O'Leary (2008) provides an insightful case study of multilingual knowledge management within the Food and Agriculture Organization (FAO), an agency of the UN, and its technology section, the World Agriculture Information Centre (WAICENT). The systems include a multilingual agricultural thesaurus called AGROVOC,[37] available in English, French, Spanish, Chinese, Arabic and Portuguese, to link resources across languages and facilitate cross-language retrieval. The EU-funded MONNET project[38] also deals with providing international access to networked knowledge within businesses by enriching ontologies with multilingual information (Montiel-Ponsoda et al. 2010). Two use cases are envisaged for technologies developed in the project: multilingual text analytics for business intelligence on companies and financial services and cross-language information access in the public sector, e.g., multilingual e-Government. Chung (2008) and Chung et al. (2004) also describe studies of three web search portals in Chinese, Spanish and Arabic specifically created for searching business information across languages for applications such as business intelligence.

As online shoppers increasingly rely on search engines to locate products and services, the use of international or multilingual Search Engine Optimisation (SEO)

[37] http://www.fao.org/agrovoc

[38] http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html

techniques[39] is another area of potential application of CLIR and MLIR technologies. The goal of multilingual SEO is to optimise websites to improve the volume of international traffic through understanding the ways that different cultures interact with the Web. For example, understanding which search engines are commonly used and getting ranked highly in them. This could involve translating key words or phrases describing products or services, using culturally-sensitive language and handling character sets correctly. However, translating content is not enough; that content must be findable from both technical and cultural perspectives and this is the overall goal of multilingual SEO.

In businesses the role of enterprise search is increasingly important to assist individuals with information retrieval and management. A report by Feldman and Sherman (2003) suggested that "*40% of corporate users reported that they cannot find the information they need to do their jobs on their intranets*." The need to therefore index and search a large variety of corporate sources (e.g., e-mails, blogs, web pages, files stores, contact lists, corporate records and databases, images, PDF files, presentations, etc.) effectively is paramount (see e.g., (Hawking 2004)). Many organisations have employees and customers based in regions around the world and therefore tools such as cross-language search and machine translation are of high importance to them. Many enterprise search tools exist from vendors, including Autonomy, Exalead, dtSearch, Funnelback, Google Search Appliance, Microsoft Search Server and Oracle Enterprise Search. Many of these now include support for internationalisation and cross-language search. For example, Google Translate is now available as a part of the Google Search Appliance; Microsoft has launched an API for Bing Translator[40] enabling integration with any search service including Microsoft's FAST Enterprise Search Platform (ESP).

Considering where business applications could go in the future, real-time voice translation is both an interesting and challenging problem. Being able to translate conversations in real time could open up a wide range of uses from tourism to business negotiations. In 2011 Google revealed a new proposed conversation mode using Google Translate for Android that would enable people to communicate fluently with a person in another language using a mobile device.[41] The first release was for English – Spanish conversation. Input is translated and the translation is read out aloud. At the time of release, the technology was in alpha status and it is too soon to evaluate the efficacy of this tool. However, it is to be expected that factors like regional accents, background noise, or rapid speech will all cause serious problems. Toshiba also announced a real-time mobile voice translation service for travelers between English, Chinese and Japanese that can handle around 30,000 words and phrases. Linking such products with mobile devices and

[39] http://econsultancy.com/uk/blog/2826-the-wonderful-world-of-multilingual-seo

[40] http://www.microsofttranslator.com/tools/#Dev

[41] http://www.telegraph.co.uk/technology/google/8255920/Google-Translate-and-the-future-of-voice.html and http://www.wired.com/epicenter/2010/02/googles-real-time-voice-translator-could-make-any-language-lingua-franca/

computer-based telephony applications (e.g., Skype) will put multilingual information access technologies at the forefront of people's interactions with all kinds of content.

## 6.4 Summing Up

This chapter has provided a summary of multilingual information access techniques that go beyond text retrieval and into multimedia. We have described a range of domains in which MLIR/CLIR technology is now being employed. Future applications will undoubtedly also involve information access on mobile devices, gaming and entertainment, and real-time conversational agents. Most applications will involve content in more than one media and combinations of language-dependent and language-independent features. The next generation of systems for multilingual information access will need to be able to combine and exploit both types of features in order to optimise retrieval performance.

Although there are obvious benefits to be gained by employing these techniques in a world that is fast becoming globalised and in which language barriers can inhibit information exchange and effective communication, the general consensus appears to be that, while the basic technology as described in this book is now in place, much work remains to be done. Despite the many academic advances, MLIR/CLIR technologies are still not widely used by the general public. This problem has already been discussed in some detail in Chapter 1. In our opinion there are two main areas where efforts should be focused. The first area involves overcoming what has been termed the 'resource bottleneck', i.e., finding ways to acquire, maintain and update the language processing tools, lexicons, thesauri and MT systems[42] needed in an easy and economic fashion. Building and sustaining these resources on a large scale is a costly enterprise and prohibitive for a single research project or a small or medium-sized business. The more languages involved, the more difficult the task. Collaborative efforts appear to be necessary. However, great care will be needed to guarantee quality and sustainability. We feel that action is needed in two directions in order to (1) build open source, freely available, high quality language processing tools and resources[43]; and (2) investigate how best to mix language-dependent and language-independent retrieval techniques to maximise MLIR/CLIR system performance on multimedia content. The second area is understanding and catering adequately for the requirements of the end user,

[42] Machine translation alone does not provide an optimal solution; research has shown that the best results are usually obtained by using a combination of different lexical resources and language processing tools.

[43] The European Commission has sponsored a number of projects aimed at producing lexical resources for European languages; however the results tend to be partial and demonstrative with few attempts at sustainability after the conclusion of a project.

especially with respect to the presentation of the results in a form that is useful and exploitable, and providing effective interaction regardless of the users' language skills. So far these questions have not received much attention from the academic community. More studies are needed into users' search behavior and expectations when interacting with content in multiple languages and to what extent language skills and cultural differences have an impact on these. Finally, progress towards more naturalistic human-computer dialogue involving real-time interaction will help to bring MLIA from the research prototype into people's lives.

## References

Bernardi R, Calvanese D, Dini L, Di Tomaso V, Frasnelli E, Kugler U, Plank B (2006) Multilingual search in libraries. The case-study of the Free University of Bozen-Bolzano. In: Proc. 5th International Conference on Language Resources and Evaluation - LREC 2006, Genoa

Bilal D, Bachir I (2007) Children's interaction with cross-cultural and multilingual digital libraries: I & II. Understanding interface design representations. J. Inf. Process. and Manag. 43: 47–80

Byrne W, Doermann D, Franz F, Gustman S, Hajĭc J, Oard D, Picheny M, Psutka J, Ramabhadran B, Soergel D, Ward T, Zhu W-J (2004) Automatic recognition of spontaneous speech for access to multilingual and historical archives. IEEE Trans. Speech and Audio Process. 12 (4): 420–435

Chau M, Qin J, Zhou Y, Tseng C, Chen H (2008) SpidersRUs: Creating specialized search engines in multiple languages. Decis. Support Syst. 45(3): 621–640

Chen J, Bao Y (2009) Information access across languages on the web: From search engines to digital libraries. In: Proc. Am. Soc. for Inf. Sci. and Technol. 46: 1–14

Chung W (2008) Web searching in a multilingual world. Commun. ACM 51(5): 32–40

Chung W, Zhang Y, Huang Z, Wang G, Ong TH, Chen H (2004) Internet searching and browsing in a multilingual world: an experiment on the Chinese business intelligence portal (CBizPort). J. Am. Soc. Inf. Sci. and Technol. 55(9): 818–831

Cleveland A, Pan D, Chen J, Yu X, Philbrick J, O'Neill M, Smith L (2008) Analysis of the health information needs and health related Internet usage of a Chinese population in the United States. J. Libr. and Inf. Service 52(3): 112–11

Clough P, Sanderson M (2006) User experiments with the Eurovision cross-language image retrieval system. J. Am. Soc. for Inf. Sci. 57(5): 697–708

Clough P, Eleta I (2010) Investigating language skills and field of knowledge on multilingual information access in digital libraries. Int. J. Digit. Libr. Syst. 1(1): 89–103

Clough P, Gonzalo J, Karlgren J, Barker E, Artiles J, Peinado V (2008) Large-scale interactive evaluation of multilingual information access systems - the iCLEF Flickr challenge. In: Proc. of Workshop on novel methodologies for evaluation in information retrieval, 30th European Conference on Information Retrieval, Glasgow, 30 March-3rd April 2008

Colowick S (2008) Multilingual search with PanImages. MultiLingual 19(2):61–63. Available at http://turing.cs.washington.edu/PanImMultilingual.pdf. Cited 15 Apr 2011

de Jong F, Oard DW, Heeren W, Ordelman R (2008) Access to recorded interviews: a research agenda. ACM J. Comput. in Cult. Herit. 1(1). Article 3

Del Bimbo A (1999) Visual information retrieval. Morgan Kaufmann Publishers Inc., San Francisco

Delezoide B, Le Borgne H (2007) SemanticVox: A multilingual video search engine. In: CIVR'07 Proc. of 6th ACM International Conference on Image and Video Retrieval: 81–84

Depeursinge A, Müller H (2010) Fusion techniques for combining textual and visual information retrieval. In: ImageCLEF: Experimental Evaluation in Information Retrieval. The Information Retrieval Series. Springer: 95–114

Dini L, Peters W, Liebwald D, Schweighofer E, Mommers L, Voermans, W (2005) Cross-lingual legal information retrieval using a WordNet architecture. In: Proc. of 10th International Conference on Artificial Intelligence and Law (ICAIL '05). ACM, New York: 163–167

Dumais S, Banko M, Brill E, Lin J, Ng A (2002) Web question answering: Is more always better? In: Proc. ACM SIGIR conference on research and development in informatiion retrieval, SIGIR 2002: 291–298

Duncker E (2002) Cross-cultural usability of the library metaphor. In: Proc. of 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '02). ACM, New York: 223–230

Eakins J, Briggs P, Burford B (2004) Image retrieval interfaces: A user perspective. In: Proc. CIVR 2004: 628–637

Enser PGB (1995) Pictorial information retrieval. J. of Documentation 51(2): 126–170

Federico M, Jones GJF (2004) The CLEF 2003 cross-language spoken document retrieval track. In: Proc. CLEF 2003. Comparative Evaluation of Multilingual Information Access Systems, Springer LNCS 3237: 646–652

Federico M, Bertoldo N, Levow G-A, Jones GJF (2005) CLEF 2004 Cross-language spoken document retrieval track. In: Proc. CLEF 2004. Multilingual Information Access for Text, Speech and Images, Springer LNCS 3491: 817–820

Feldman S, Sherman C (2003) The high cost of not finding information. Tech. Rep. 29127, IDC, April 2003

Ferro N, Peters C (2010) CLEF 2009 ad hoc track overview: TEL and Persian tasks. In: Proc. CLEF 2009. Multilingual Information Access Evaluation I. Text Retrieval Experiments, Springer LNCS 6241: 13–35

Forner P, Giampiccolo D, Magnini B, Peñas A, Rodrigo A, Sutcliffe R (2010) Evaluating multilingual question answering systems at CLEF. In: Proc. 7th International Conference on Language Resources and Evaluation. LREC 2010, Malta: 2774–2781

Fujii A., Iwayama M, Kando N (2004) Test collections for patent-to-patent retrieval and patent map generation in NTCIR-4 workshop. In: Proc. 4th International Conference on Language Resources and Evaluation. LREC 2004: 1643–1646

Fujii A, Iwayama M, Kando N (2007) Introduction to the special issue on patent processing. J. Inf. Process. and Manag. 43(5): 1149–1153

Fukimoto J, Kato T, Masui F, Mori T (2007) Overview of the 4th question answering challenge at NTCIR workshop-6. NTCIR-6 Workshop Proc., National Institute of Informatics, Tokyo. Available at http://research.nii.ac.jp/ntcir/ntcir-ws6/OnlineProceedings/NTCIR/75-revised-20070521.pdf. Cited 15 Apr 2011

Garofolo J, Fiscus J, Fisher W (1997) Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora, Proc. DARPA Speech Recognition Workshop, February 1997

Garofolo J, Auzanne CGP, Voorhees E (2000) The TREC spoken document retrieval track: A success story. In: Proc. RIAO 2000: 1–20

Goodrum A (2000) Image information retrieval: An overview of current research. Informing Sci. 3 (2): 63–66

Hauptmann AG, Scheytt P, Wactlar HD, Kennedy PE (1998) Multi-lingual Informedia: A demonstration of speech recognition and information retrieval across multiple languages. In: Information Retrieval Across Multiple Languages, BNTUW-98 Proc. of DARPA Workshop on Broadcast News Understanding Systems

Hauptmann AG, Jin R, Ng TD (2002) Multi-modal information retrieval from broadcast video using OCR and speech recognition. In: JCDL '02 Proc. of 2nd ACM/IEEE-CS joint conference on Digital Libraries: 160–161

Hawking D (2004) Challenges in enterprise search. In: Proc. Fifteenth Australasian Database Conference (ADC 2004): 15–24

Hecht B, Gergle D (2010) The tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context. In: Proc. 28th International Conference on Human Factors in Computing Systems (CHI '10). ACM, New York: 291–300

Hersh WR, Donohoe LC (1998) SAPHIRE international: a tool for cross-language information retrieval. In: Proc. 1998 AMIA Annual Symposium: 673–677

Hersh WR, Müller H, Jensen JJ, Yang J, Gorman PN, Ruch P (2006). Advancing biomedical image retrieval: development and analysis of a test collection. J. Am. Med. Inform. Assoc. 13: 488–496

Hirschman L, Gaizauskas R (2001) Natural language question answering: The view from here. Nat. Lang. Eng. 7(4): 275–300

Hsu DF, Taksa I (2005) Comparing rank and score combination methods for data fusion in information retrieval. J. Inf. Retr. 8(3): 449–480

Huijbregts MAH, Ordelman RJF, de Jong FMG (2007) Annotation of heterogeneous multimedia content using automatic speech recognition. In: Proc. Semantic and Digital Media Technologies 2nd International Conference on Semantic Multimedia (SAMT'07). Falcidieno B, Spagnuolo M, Avrithis Y, Kompatsiaris I, Buitelaar P (eds.). Springer-Verlag, Berlin, Heidelberg: 78–90

Jones GJF (2000) Applying machine translation resources for cross-language information access from spoken documents. In: Proc. MT 2000: Machine Translation and Multilingual Applications in the New Millennium, Exeter, U.K: 4-1–4-9

Jones GJF, Larson M, Marchand-Maillet S (2008). Multilingual/multimedia information retrieval. In: State-of-the-Art. Del. 1.1 MultiMatch Project. http://www.multimatch.eu/docs/publicdels/1.1.3.pdf

Kahn Jr CE (2009) Multilingual retrieval of radiology images. RadioGraphics 29: 23–29

Karlgren J, Gonzalo J (2010*)* Interactive image retrieval. In: ImageCLEF - Experimental Evaluation in Visual Information Retrieval. The Information Retrieval Series, (32). Springer, 117–138. ISBN 978-3-642-15180-4

Katz B, Borchardt G, Felshin S, Shen Y, Zaccak G (2007) Answering English questions using foreign-language, semi-structured sources. In: Proc. International Conference on Semantic Computing (ICSC '07). IEEE Computer Society, Washington, DC: 439–445

Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: Proc. of the Machine Translation Summit X: 79–86

Koolen M, Adriaans F, Kamps J, de Rijke, M (2006) A cross-language approach to historic document retrieval. In: Proc. 28th European Conference on IR Research (ECIR 2006): 407–419

Larson M, Newman E, Jones GJF (2009) Overview of VideoCLEF 2008: automatic generation of topic-based feeds. In: Proc. CLEF 2008 Evaluating Systems for Multilingual and Multimodal Information Access. Springer LNCS 5706: 906–917

Larson M, Newman E, Jones GJF (2010) Overview of VideoCLEF 2009: new perspectives on speech-based multimedia content. In: Proc. CLEF 2009 Multilingual Information Access Evaluation II. Multimedia Experiments, Springer LNCS 6242: 354–368

Lazarinis F, Vilares J, Tait J, Efthimiadis EN (2009) Current research issues and trends in non-English web searching. J. Inf. Retr. 12(3): 230–250

Liu F, Ackerman M, Fontelo P (2006) BabelMeSH: development of a cross-language tool for Medline/PubMed. In: Proc. AMIA Annual Symposium: 1012

Lu WH, Lin RS, Chan YC, Chen KH (2008) Using Web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. Decis. Support Syst. 45 (3): 585–595

Lyu MR, Song J, Cai M (2005) A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Trans. Circuits and Syst. for Video Technol. 15(2): 243–255

Maeda A (2004) Cross-language information access on the Web: A tool to help learning foreign languages. In: Cantoni L, McLoughlin C (eds.), Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004: 5449–5454

Magnini B, Giampiccolo D, Aunimo L, Ayache C, Osenova P, Penas A, de Rijke M, Sacaleanu B, Santos D, Sutcliffe R (2006) The multilingual question answering track at CLEF. In: Proc. 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy: 1156–1163

Marlow J, Clough P, Dance K. (2007) Multilingual needs of cultural heritage website visitors: A case study of Tate Online, In: Trant J, Bearman D (eds.). Proc. International Cultural Heritage Informatics Meeting (ICHIM07) http://www.archimuse.com/ichim07/papers/marlow/marlow.html. Cited 15 Apr 2011

Meng W, Yu CT, Liu K (2002) Building efficient and effective metasearch engines. ACM Computing Surveys 34(1): 48–89

Min J, Jiang J, Leveling J, Jones GJF, Way A (2010) DCU's experiments for the NTCIR-8 IR4QA Task. In: Proc. NTCIR-8 Workshop. National Institute of Informatics, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/08-NTCIR8-IR4QA-MinJ.pdf Cited 15 Apr 2011

Mitamura T, Shima H, Sakai T, Kando N, Mori T, Takeda K, Lin C-Y, Song R, Lin C-J, Lee C-W (2010) Overview of the NTCIR-8 ACLIA tasks: advanced cross-lingual information access. In: Proc. NTCIR-8 Workshop. National Institute of Informatics, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-CLQA-MitamuraT.pdf. Cited 15 Apr 2011

Montiel-Ponsoda E, Aguado-de-Cea G, Gómez-Pérez A, Peters W (2010) Enriching ontologies with multilingual information. J. Nat. Lang. Eng. Available on Cambridge Journals Online, 09 June 2010

Müller H, Kalpathy-Cramer J (2010) The medical image retrieval task. In: ImageCLEF: Experimental Evaluation in Information Retrieval. The Information Retrieval Series. Springer: 239–257

Müller H, Clough P, Deselaers T, Caputo B (eds.) (2010) ImageCLEF - Experimental evaluation of visual information retrieval. The Information Retrieval Series. Springer. http://dx.doi.org/10.1007/978-3-642-15181-1

Noguera E, Llopis F, Ferrandez A, Escapa A (2007) Evaluation of open-domain question answering systems within a time constraint. In: Proc. 21st International Conference on Advanced Information Networking and Applications Workshops, AINAW '07, Niagara Falls, Ontario: 260–265

Nomoto M, Fukushige Y, Sato M, Suzuki H (2004) Are we making progress? - An analysis of NTCIR QAC1 and 2. In: Proc. NTCIR-4 Workshop, National Institute of Informatics, Tokyo. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/. Cited 15 Apr 2011

O'Leary DE (2008) A multilingual knowledge management system: A case study of FAO and WAICENT. Decis. Support Syst. 45(3): 641–661

Oard D (1997) Serving users in many languages: Cross-language information retrieval for digital libraries. D-Lib Magazine Dec.1997. http://www.dlib.org/dlib/december97/oard/12oard.html. Cited 15 Apr 2011

Ordelman R, de Jong F, Heeren W (2006) Exploration of audiovisual heritage using audio indexing technology. In: Proc. 1st ECAI Workshop on Intelligent Technologies for Cultural Heritage Exploitation: 36–39

Pavani AMB (2001) A model of multilingual digital library. Ciência da Informação, Brasília 30 (3): 73–81

Pecina P, Hoffmannová P, Jones GJF, Zhang Y, Oard DW (2008) Overview of the CLEF-2007 cross-language speech retrieval track. In: Proc. CLEF 2007. Advances in Multilingual and Multimodal Information retrieval, Springer LNCS 5152: 674–686

Peruginelli G (2008) Multilingual legal information access: an overview. In: Chiocchetti E, Voltmer L, (eds), Harmonising Legal Terminology, Bolzano: EURAC, 2007, ISBN: 9788888906393: 6–34

Peters W, Sagri MT, Tiscornia D (2007) The structuring of legal knowledge in LOIS. Artif. Intell. and Law 15(2): 117–135

Peterson T (2002) The importance of being multilingual. Business Week September 4 2002. http://www.businessweek.com/bwdaily/dnflash/sep2002/nf2002094_2752.htm. Cited 15 Apr 2011

Pingali P, Jagarlamudi J, Varma V (2006) WebKhoj: Indian language IR from multiple character encodings. In: Proc. World Wide Web Conference Series – WWW 2006: 801–809

Potthast P, Barrón-Cedeño A, Stein B, Rosso P (2011) Cross-language plagiarism detection. Lang. Resour. and Eval. 45(1): 45–62

Rasmussen EM (1997) Indexing images. Annu. Rev. Inf. Sci. and Technol. 32: 169–196

Rautiainen M, Seppänen T, Ojala T (2006) Advancing content-based retrieval effectiveness with cluster-temporal browsing in multilingual video databases. In: Proc. International Conference on Multimedia Computing and Systems/International Conference on Multimedia and Expo: 377–380

Reid, NH (1999) Photographic archives: Aberdeen, Dundee and St Andrews. In: Coppock T (ed.) Making information available in digital format: perspectives from practitioners, The Stationery Office, Edinburgh: 106–119

Roda G, Tait J, Piroi F, Zenz V (2010) CLEF-IP 2009: retrieval experiments in the intellectual property domain. In: Proc. CLEF 2009. Multilingual Information Access Evaluation I. Text Retrieval Experiments, Springer, LNCS 6241: 385–409

Rodrigo A, Peñas A, Verdejo F (2009) Overview of the answer validation exercise 2008. In: Proc. CLEF 2008. Evaluating Systems for Multilingual and Multi-modal Information Accessl, Springer LNCS 5706: 296–313

Rydberg-Cox JA (2005) The cultural heritage language technologies consortium. D-Lib Magazine 11(5). http://www.dlib.org/dlib/may05/rydberg-cox/05rydberg-cox.html. Cited 15 Apr 2011

Sagri MT, Tiscornia D (2004) Semantic lexicons for accessing legal information. In: Proc. EGOV2004, LNCS 3183, Springer: 72–81

Saha S, Fernandez A (2007) Language barriers in health care. J. Gen. Intern. Medicine 22: 281–282

Sanderson M, Clough P (2002) Eurovision - an image-based CLIR system. In: Proc. Workshop of 25th ACM SIGIR Conference on Research and Development in Information Retrieval, Workshop 1: Cross-Language Information Retrieval: A Research Roadmap: 56–59

Santos D, Cabral LM (2010) GikiCLEF: expectations and lessons learned. In: Proc. CLEF 2009. Multilingual Information Access Evaluation I: Text Retrieval Experiments, LNCS 6241, Springer: 212–222

Sasaki Y, Lin C-J, Chen K-H, Chen H-H (2007) Overview of the NTCIR-6 cross-lingual question answering (CLQA) task. In: Proc. NTCIR-6 Workshop, National Institute of Informatics, Tokyo. http://research.nii.ac.jp/ntcir/ntcir-ws6/OnlineProceedings/NTCIR/72-revised-20070604.pdf. Cited 15 Apr 2011

Savino P, Peters C (2004) ECHO: a digital library for historical film archives. Int. J. on Digit. Libr. 4(1): 3–7

Sheridan P, Weschler M, Schäuble P (1997a) Cross-language speech retrieval: establishing a baseline performance. In: Proc. 20th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97): 99–108

Sheridan P, Braschler M, Schäuble P (1997b) Cross-language information retrieval in a multilingual legal domain. In: Peters C, Thanos C (eds.) Proc. First European Conference on Research and Advanced Technology for Digital Libraries (ECDL '97), Springer: 253–268

Smeaton A, Over P, Kraij W (2009) High-Level Feature Detection from Video in TRECVid: A 5-Year Retrospective of Achievements. In: A. Divakaran (ed.), Multimedia Content Analysis, Signals and Communication Technology, Springer Science + Business Media: 151–174

Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. Pattern Analysis and Mach. Intell. 22(12): 1349–1380

Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T, Tufiş D, Varga D (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proc. 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24–26 May 2006

Tran TT (2009) Understanding cultural barriers in hepatitis B virus infection. Cleveland Clinic J. of Medicine 73(3): 10–13

Turmo J, Comas PR, Rosset S, Lamel L, Moreau, Mostefa N (2009) Overview of QAST 2008. In: Proc. CLEF 2008. Evaluating Systems for Multilingual and Multi-modal Information Access, Springer LNCS 5706: 314–324

Udupa R, Khapra M (2010) Improving the multilingual user experience of Wikipedia using cross-language name search. In: Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10): 492–500

Voorhees EM (2001) The TREC question answering track. Nat. Lang. Eng. 7(4): 361–378.

Voorhees EM, Tice D (2000) Building a question answering test collection. In Proc. 23rd ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 2000: 200–207

Witten IH, Bainbridge D (2003) How to build a digital library. San Francisco, CA: Morgan Kaufman Publishers

Wu D, Luo B, He D (2010) How multilingual digital information is used: A study in Chinese academic libraries. In: Proc. International Conference on Management and Service Science (MASS 2010): 1–4

Zhang J, Lin S (2007) Multiple language supports in search engines. Online Inf. Review 31(4): 516–532

# Glossary of Acronyms

| | |
|---|---|
| AAAI | Association for the Advancement of Artificial Intelligence |
| AC | Acquis Communautaire (EU) |
| ACM | Association for Computing Machinery |
| AP | Average Precision |
| API | Application Programming Interface |
| ARRS | American Roentgen Ray Society |
| ASCII | American Standard Code for Information Interchange |
| ASR | Automatic Speech Recognition |
| CBIR | Content-Based Information Retrieval |
| CH | Cultural Heritage |
| CD | Compact Disc |
| CL | Computational Linguistics |
| CLMIR | Cross-Language Medical Information Retrieval |
| CL-SDR | Cross-Language Spoken Document Retrieval |
| CL-SR | Cross-Language Speech Retrieval |
| CLEF | Cross-Language Evaluation Forum |
| CLIR | Cross-Language Information Retrieval (or, less commonly now, Cross-Lingual Information Retrieval) |
| DARPA (US) | Defense Advanced Research Projects Agency |
| DBMS | Database Management System |
| DCG | Discounted Cumulative Gain |
| DT | Document Translation |
| DVD | Digital Versatile Disc (previously Digital Video Disc) |
| EC | European Commission |
| ECHO | European CHronicles Online (European Commission Project) |
| ELDA | Evaluation and Language Resources Distribution Agency |
| EMIR | European Multilingual Information Retrieval (European Commission Project) |
| EPO | European Patent Office |
| ESP | Enterprise Search Platform (Microsoft) |

C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0, 

| | |
|---|---|
| EU | European Union |
| EUROPARL | European Parliament Proceedings Parallel Corpus |
| FAO | Food and Agriculture Organisation (UN) |
| FAQ | Frequently Asked Questions |
| FIRE | Forum for Information Retrieval Evaluation (for Indian Languages) |
| GMAP | Geometric Mean Average Precision |
| HMM | Hidden Markov Model |
| IA | Information Access |
| IDC | International Data Corporation |
| IFLA | International Federation of Library Associations |
| IR | Information Retrieval |
| ISJ | Interactive Searching and Judging |
| ISO | International Organization for Standardization |
| IT | Information Technology |
| JRC | (European) Joint Research Centre |
| k-NN | k-Nearest Neighbour |
| LDC | Linguistic Data Consortium |
| LOIS | Lexical Ontologies for Legal Information Sharing |
| LSI | Latent Semantic Indexing |
| MAP | Mean Average Precision |
| MediaEval | Benchmarking initiative for multimedia evaluation |
| MeSH | Medical Subject Headings |
| MLIA | Multilingual Information Access |
| MLIR | Multilingual Information Retrieval |
| MRR | Mean Reciprocal Rank |
| MT | Machine Translation |
| MTF | Move-to-Front |
| MTurk | Mechanical Turk (Amazon) |
| MultiMatch | Multilingual/Multimedia Access To Cultural Heritage (European Commission Project) |
| nDCG | Normalized Discounted Cumulative Gain |
| NER | Named Entity Recognition |
| NERI | Named Entity Recognition and Identification |
| NII | National Institute of Informatics (Japan) |
| NIST | (US) National Institute of Standards and Technology |
| NLM | (US) National Library of Medicine |
| NLP | Natural Language Processing |
| NSF | (US) National Science Foundation |
| NTCIR | NII Text Collection for IR (Research infrastructure for comparative evaluation of information retrieval and access technologies of the National Institute for Informatics, Tokyo) |
| OCLC | Online Computer Library Centre |
| OCR | Optical Character Recognition |
| OOV | Out-of-Vocabulary (terms not found in dictionary being used) |

| | |
|---|---|
| OPAC | Online Public Access Catalogue |
| P@n | Precision at n |
| PDF | Portable Document Format |
| PRP | Probability Ranking Principle |
| QA | Question Answering |
| QAC | Question Answering Challenge |
| QBVE | Query-By-Visual-Example |
| QT | Query Translation |
| R&D | Research and Development |
| RHL | Ranked Half-Life |
| RR | Relative Relevance |
| RSNA | Radiological Society of North America |
| RSS | Really Simple Syndication |
| RSV | Retrieval Status Value |
| SDA/ATS | Swiss News Agency |
| SDR | Spoken Document Retrieval |
| SEO | Search Engine Optimisation |
| SERP | Search Engine Results Page |
| SIGIR | Special Interest Group on Information Retrieval |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |
| TALIP | Transactions on Asian Language Information Processing |
| TBIR | Text-Based Information Retrieval |
| TDIL | Technology Development for Indian Languages (Indian Government Programme) |
| TEL | The European Library |
| tf.idf | term frequency – inverse document frequency |
| TIDES | Translingual Information Detection, Extraction and Summarization (US DARPA Programme) |
| TrebleCLEF | Evaluation, Best Practice & Collaboration for Multilingual Information Access (European Commission Project) |
| TREC | Text REtrieval Conference |
| TRECVid | TREC Video Retrieval Evaluation |
| UCS | Universal Multiple-Octet Coded Character Set |
| UMLS | Unified Medical Language System |
| UN | United Nations |
| UNICODE | The standard for digital representation of the characters used in writing all of the world's languages. |
| UTF | UNICODE Transformation Format |
| WAICENT | World Agriculture Information Centre (UN) |
| XML | Extensible Markup Language |

# Index


**A**
Accumulator, 39
AltaVista, 2
Arctos, 112
ASR. *See* Automatic speech recognition
Audio segmentation, 179, 180
Austrian National Library, 193
Auto-completion, 97
Automated term suggestion, 108
Automatic speech recognition (ASR), 177–179

**B**
Babelfish, 11, 112
BabelMeSH, 196
Bag of tokens, 37
Bag of words, 36–38, 61–62
Bibliothèque nationale de France, 193
Bing, 189
Bing translator, 195, 200
Biocaster, 110
British Library, 193

**C**
Character encoding, 6, 25, 118
Checksum, 26
CL. *See* Computational linguistics
Clarity, 7, 98, 102, 105, 116
Classifiers
  k-nearest neighbour, 182
  Naive Bayes, 182
  support vector machine, 182
CLIR. *See* Cross-language information retrieval
Computational linguistics (CL), 5, 9
Copyright, 133
Cranfield, 131–132, 138, 160, 165
Cross-Language Evaluation Forum (CLEF), 9–10, 71, 80–81, 125, 133–137, 140–145
Cross-language information retrieval (CLIR), 5, 57–82
  best practices, 80–82
  commercial CLIR, 12, 14, 199–201
  implementation, 58–62
  system performance, 142
Cross-language medical information retrieval, 195–196
Crowd-sourcing, 144
Cultural differences, 87, 92–94, 192

**D**
DARPA. *See* Defense advanced research projects agency
Decompounding, 34–35
Defense Advanced Research Projects Agency (DARPA), 7, 8
Digital library initiative, 8
Document surrogates, 102–106, 122

**E**
EC. *See* European Commission
ECHO project, 181
Effectiveness, 130, 145
Efficiency, 130, 146
ELDA. *See* Evaluations and Language resources Distribution Agency
EMIR project, 7
Enterprise search, 12, 199–200
  Autonomy, 200


C. Peters et al., *Multilingual Information Retrieval*,
DOI 10.1007/978-3-642-23008-0, © Springer-Verlag Berlin Heidelberg 2012

Enterprise search (*cont.*)
dtSearch, 200
Exalead, 200
Funnelback, 200
Google search appliance, 200
Microsoft FAST, 200
Microsoft search server, 200
Oracle enterprise search, 200
EU. *See* European Union
EU-NSF Working Group on Multilingual Information Access, 7
Europeana, 8, 119, 193–195
European Commission (EC), 7
European Parliament, 197
European Union (EU), 7, 65, 197–198
Eurovision, 112, 176
EuroWordNet, 109
Evaluation, 129–166
design decisions, 162–164
experimental setup, 157–159
guidelines, 162
system oriented, 130, 156–157
usage scenarios, 158
user-oriented, 130, 156–157
Evaluation campaigns, 9–10, 132–133
Benchathlon, 141
CLEF (*see* Cross-Language Evaluation Forum)
CLEF-IP, 198
CLEF-QA, 187–188
CLEF-SDR, 179
CLEF-SR, 179
FIRE (*see* Forum for Information Retrieval Evaluation)
iCLEF, 125, 159–169, 176
ImageCLEF, 142, 144, 176
ImageEval, 141
MediaEval, 183
NTCIR (*see* NII Text Collection for IR)
NTCIR-QAC, 186
Text Analysis Conference, 186
TREC (*see* Text REtrieval Conference)
TREC-SDR, 179
TRECVid, 141, 182
VideoCLEF, 182
Evaluations and Language resources Distribution Agency (ELDA), 162, 180

**F**
Faceted browsing, 110–111
False friends, 60
Flickr, 177
FlickrArabic, 107, 176
Forum for Information Retrieval Evaluation (FIRE), 9–10, 133–134, 143

**G**
Globalisation, 189, 199
Google, 11, 12, 19, 52, 108, 189, 200
language tools, 11
translate, 11, 81, 93, 100–105, 195, 200
web search, 108, 121–123
Grand challenge, 1, 9–10, 171

**H**
Hash table, 39
Homonymy, 35

**I**
IA. *See* Information access
IDC, 12, 13
IFLA. *See* International Federation of Library Associations
Image retrieval, 173–177
combination techniques, 176
content-based techniques, 173
cross-language search, 173
manual annotation, 175
semantic features, 174
semantic gap, 173
text-based techniques, 176
Indexing, 20, 22, 24–36
document formation, 27–28
enrichment, 35–36
feature normalisation, 31–35
indexing features, 25, 28, 36
indexing terms, 28, 37
language identification, 26–27
low-level features, 173
parsing, 28–31
phonemic features, 178
pipeline, 25, 36
pre-processing, 25–26
segmentation, 28–31
semantic features, 181
tokenization, 28–31
visual descriptors, 180
Information access (IA), 5, 17
Information need, 18, 21, 86, 89
Information seeking, 85–95
context, 87, 92
individual differences, 92

process, 86, 87, 94
searching and browsing, 88, 109
search session, 88
search tasks, 90
user studies, 90
Interactive cross-language information retrieval, 86, 94–97, 103
Interactive information retrieval, 86
Interactive system evaluation, 159–160
International Children's Digital Library, 192
International Federation of Library Associations (IFLA), 6
Internationalisation, 6, 11, 118
Internet, 1–4
Intranet, 4
Inverted index, 38–40, 53
ISO 639, 6
ISO 5964, 6
ISO–8859, 26
ISO/IEC 10646, 6

**K**
Known item retrieval, 152, 156

**L**
Language skills, 4, 89, 92–94, 157, 177, 192
Latent semantic indexing, 65, 67
Latin square, 158
Lexical coverage, 34, 57, 71, 79
Linguistic Data Consortium, 162, 198
Localisation, 6, 11, 94, 118–121
Logograms, 29
LOIS project, 197
Lycos, 2

**M**
Machine translation, 71, 73
Matching, 20, 23, 36–52
basic algorithm, 39
Mechanical Turk (MTurk), 145, 155
Merging, 77–79
collection fusion, 73, 77
data fusion, 72, 77, 175
interleaving, 77
median rank, 73
raw score merging, 78
score normalisation, 78
MLIA. *See* Multilingual information access (MLIA)
MLIR. *See* Multilingual information retrieval (MLIR)
MONNET project, 199
Monoglots, 92
MULINEX project, 98
Multilingual agricultural thesaurus - AGROVOC, 199
Multilingual content management, 199
Multilingual digital libraries, 191–195
Multilingual information access (MLIA), 5, 17, 58
Multilingual information retrieval (MLIR), 5, 17, 58
Multilingual Informedia project, 181
Multilingual knowledge management, 199
Multilingual legal knowledge base, 197
Multilingual ontologies, 199
Multilingual patent retrieval, 198
Multilingual search interfaces, 115, 121
best practices, 114
Multilingual taxonomy, 110
Multilingual thesauri, 6, 196
Multilingual web search, 189–191
Ajeeb, 189
Baidu.com, 189
Google, 189
multiple language support, 190
SpidersRUs, 189
Webkhoj, 191
MultiMatch, 7, 91, 111, 157, 181
Multimedia retrieval, 171
Multiword expressions, 62, 65

**N**
Named entity recognition (NER), 36, 185
National Institute of Standards and Technology (NIST), 9
National Science Foundation (NSF), 7
Natural Language Processing (NLP), 5
NER. *See* Named entity recognition (NER)
*n*-grams, 27, 29, 30, 34, 59, 73, 178
NII Text Collection for IR (NTCIR), 9, 133–135, 137, 139, 144, 162, 186–187, 198
NIST. *See* National Institute of Standards and Technology
NLP. *See* Natural Language Processing
NSF. *See* National Science Foundation

**O**
Optical character recognition (OCR), 180
Out-of-vocabulary terms, 64, 69, 81

**P**
Page Rank, 51–52
PanImages, 97, 100, 176
Performance measures, 145–154
  average precision, 137
  Bpref, 153
  discounted cumulative gain, 152
  efficiency, 161
  GMAP, 151–152
  graded relevance measures, 152
  interactive searching and judging, 163
  mean average precision, 137, 151
  mean reciprocal rank, 152, 163
  normalized discounted cumulative gain, 152
  P@10, 155, 163
  precision and recall, 22, 131, 151–152
  precision @ n, 149
  ranked half-life, 161
  relative relevance, 161
  R-Precision, 151
  user-oriented measures, 161
  utility, 161
Polyglots, 92, 98
Polysemy, 35
Probability ranking principle (PRP), 21

**Q**
QBVE. *See* Query-By-Visual-Example
Query-biased summaries, 103
Query-By-Visual-Example (QBVE), 95
Query expansion, 69–71
  blind relevance feedback, 48, 69–70, 109
  post-translation query expansion, 70
  pre-translation query expansion, 69
  pseudo relevance feedback (*see* Blind relevance feedback)
  relevance feedback, 46–48, 53, 69, 109, 183
Query formulation, 18, 21, 57, 73, 95–102
Query reformulation, 108–109
  categories of reformulation, 108
Query structuring, 64
Query translation, 95–102, 109
  fully automatic, 97
  user-assisted, 97
Question answering, 172, 183–188
  answer extraction, 185
  answer validation, 185
  cross-language search, 183, 189, 193
  document retrieval, 185
  question analysis, 184
  system architecture, 184

**R**
Relevance, 20, 88, 131, 132, 137, 145
Requirements elicitation, 90
  personas, 117
  scenarios, 117
  task analysis, 91, 117
  use cases, 117
Retrieval status values (RSVs), 23, 44, 73

**S**
SemanticVox project, 181
Shoah Visual History Foundation, 179
Similarity thesaurus, 36, 65, 67, 72, 198
Singular value decomposition (SVD), 67
Speech recognition, 177–178
Speech retrieval, 177–180
  cross-language search, 178
  indexing, 178
St Andrews University Library, 174–176
Statistical significance testing, 154–155
  sign test, 154
  Student's t-test, 154
  Wilcoxon signed rank test, 154
Stemming, 32–33, 37, 62
  Porter stemmer, 32
Stopwords, 27, 30–31
  Smart stopword list, 37
SVD. *See* Singular value decomposition
System architectures
  bilingual with query translation, 74
  multilingual with document translation, 77
  multilingual with query translation, 75
SYSTRAN, 7

**T**
TANGO, 11
Tate Online, 90–91, 111–112, 192
TDIL programme, 7
TEL. *See* The European Library
Term ambiguity, 61–62
Term selection, 63
Test collections, 9, 132–140
  documents or docs, 133–134
  ground truth creation, 137, 145
  interactive searching and judging, 144
  move-to-front pooling, 143
  pooling, 138–140
  queries or queries, 134–137, 139
  relevance assessment, 137–139
  reusability, 140
  speech data, 181

test data availability, 162
topics, 134
user satisfaction measures, 155
Text REtrieval Conference (TREC), 9, 65, 133–135, 137, 144, 178, 186
The European Library (TEL), 8, 193
TIDES programme, 5, 8
TIDES Surprise Language Exercise, 8
Training data, 65, 67, 68, 188
comparable corpora, 65–66
Europarl corpus, 65, 198
Hansards corpora, 65, 198
JRC-Acquis corpora, 198
noisy comparable corpus, 68
parallel corpora, 65–66
UN parallel texts, 198
Translation methods, 58–60
combination approaches, 71–73
document translation, 58, 79
indirect translation, 79–80
lexical triangulation, 79
machine-readable dictionaries, 62–65
machine translation, 71
no translation, 59
pivot language, 58, 79
pseudo-translation, 61
query translation, 59
statistical approaches, 65–69
substring matching, 59
Translation quality, 61
Translation resources, 60–61
TrebleCLEF, 80, 162
TREC. *See* Text REtrieval Conference
Trec_eval, 162

**U**
Unicode, 6, 25, 53, 118, 134
Unified index, 75, 76
Unified Medical Language System (UMLS) Metathesaurus, 196
Universal multiple-octet coded character set (UCS), 6
Usability, 114
User behaviour, 156
User-centered design, 116
UTF–8, 6, 118, 134

**V**
Vector space model, 41–43, 46
Video retrieval, 180–183
cross-language search, 181, 183
indexing, 181
Virtual keyboard, 119, 122–123
Visualisation techniques, 109–113
Voice translation service, 200

**W**
Weighting schemes, 38, 40–52
BM.25, 50, 73
divergence from randomness, 41
language models/langmod, 41, 50–51, 68
Lnu.ltn, 46, 73
ntc.ntc, 45
probabilistic weighting schemes, 41, 48–50
Robertson-Spärck Jones weighting, 49
tf.idf-Cosine, 43–46, 68, 73
Wikipedia, 66, 109, 182
Word error rate, 178–179
WordNet, 197
Word sense disambiguation, 63–64
World Wide Web, 1–3, 18, 26, 51, 66

**X**
XML, 25–26, 53, 134

**Y**
Yahoo, 2, 11, 19, 71, 189–190

**Z**
Zipf's law, 30