

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

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 TREC¹⁵ (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 (CLEF¹⁶) for European languages, the NII Text Collection for IR (NTCIR¹⁷) for Asian languages and the Forum for Information Retrieval Evaluation (FIRE¹⁸) 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! Babelfish²⁰ 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, IDC²¹ 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 Evans²² 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, DAICT, 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

Multilingual Information Retrieval

From Research To Practice

Peters, C.; Braschler, M.; Clough, P.

2012, XVIII, 218 p., Hardcover

ISBN: 978-3-642-23007-3