Chapter 9

MULTILINGUAL WWW

Modern Multilingual and Cross-lingual Information Access Technologies

Feiyu Xu
DFKI LT Lab, Saarbrücken, Germany

Abstract: In this chapter, we describe the state of the art cross-lingual and multilingual strategies and their related areas. In particular, we show a WWW-based information system called MIETTA, which allows uniform and multilingual access to heterogeneous data sources in the tourism domain. The design of the search engine is based on a new cross-lingual framework. The framework integrates a cross-lingual retrieval strategy with natural language techniques: information extraction and multilingual generation. The combination of information extraction and multilingual generation enables on the one hand, multilingual presentation of the database content, and on the other hand, free text cross-lingual information retrieval of the structured data entries. We will demonstrate that the framework is useful for domain specific and multilingual applications and provides strategies for the future question answering systems, which should be able to deal with heterogeneous data sources in a multilingual environment.

Key words: multilingual and crosslingual information access technologies, information extraction/retrieval, multilingual generation, internet

1. INTRODUCTION

World Wide Web (WWW) plays more and more an important role as a global knowledge base and an international comfortable communication, information and business network. Above all, huge amount of business-to-business and business-to-consumer transactions will take place in the near future online and internationally. Although the major online transactions are still in USA and North America, it is predicated that the Internet Marketing will soon spread to Asia and European nations.

The growth of Non-English users and content on the web in the recent years is impressive. Various approaches are developed to estimate its multilingual characteristics [Nunberg, 1996; Crystal, 1997; Grefenstette and Nioche, 2000; Grefenstette, 2001; Xu, 2000a].

In 1998, the statistical data provided by an OCLC Web Characterization Project[1] and GVU's WWW user surveys[2] showed that English had a dominant role on both the content and the user sites: GVU's WWW user surveys found out that 92.2% users were primary English-speakers, while OCLC reported that English occupied more than 70% of the web content (Table 9-1):

Table 9-1: content distribution in 1998

Language	%	Language	%
English	71	Portuguese	2
German	7	Dutch	1
Japanse	4	Italian	1
French	3	Chinese	1
Spanish	3	Korean	1

According to the newest data in 2002 provided by Global Internet Statistics[3], the English-speaking users are only 36.5%, while the Non-English users play now a dominant role with 63.5%, in which the European languages contain 35.5% and Chinese-speaking people are 10.8% and Japanese 9.7% (more information can be found in Table 9-2).

Table 9-2: distribution of users

Language	%	Language	%
English	36.5	Italian	3.8
Chinese	10.8	French	3.5
Japanese	9.7	Portuguese	3.0
Spanish	7.2	Russian	2.9
German	6.7	Dutch	2.0

At the same time, the content distribution has changed too: the Asian languages Japanese and Chinese grow at faster space than other languages (see Table 9-3), although English still plays a dominant role.

Table 3: content distribution in 2002

Language	%	Language	%
English	68.4	Spanish	2.4
Japanese	5.9	Russian	1.9
German	5.8	Italian	1.6
Chinese	3.9	Portuguese	3.4
French	3.0	Korean	1.3

[1]http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm
[2] http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/graphs/general/q11.htm
[3] http://global-reach.biz/globstats/index.php3

Parallel to the multilingual development on the web, there are a series of conferences and competitions organized by governments, research institutions and industries in the last few years to push idea exchange and further development of the multilingual and cross-lingual information access technologies.

Text Retrieval Conference[4] (TREC), sponsored by American government institutions NIST and DARPA, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community including government, industry and academia from all over of the world. TREC develops tasks and provides training corpora and performs evaluation. In TREC-5, Multilingual Information Retrieval task in TREC was integrated, focusing on information retrieval in non-English languages Spanish and Chinese. Cross-Lingual Information Retrieval (CIIR) was a new task since TREC-6. In contrast to the multilingual task, CIIR allows users to send a query in a different language than the document language [Schäuble and Sheridan, 1997].

EU is faced with the multilingual problem in her everyday life. Therefore, EU has provided and is going to provide big funding for research and development of multilingual and cross-lingual information access technologies. Human Language Technologies (HLT) RTD[5] is part of the *Information Society Technologies* (IST) program founded by EU, addressing on multilingual communication and cross-lingual information management

http://www.ee.umd.edu/medlab/mlir/conferences.html has collected a list of conferences related to multilingual IR and CLIR. In addition to the main conferences like TREC, SIGIR, ACL, there have been also special workshops for Asian languages like IRAL and for Arabic language within the ACL context active in the past years.

[Weikum, 2002] assumed that the current Web contains about 1 Billion (10^9) documents, while the so-called "deep Web" can potentially have 500 Billion documents containing databases and other digital libraries and archives behind Intranet and other portals. Most information retrieval approaches deal either with unstructured textual information like web documents [Braschler and Schäuble, 1999; Davis and Ogden, 1997; Erbach et al., 1997; Oard, 1999a; Oard, 1999b; Busemann, 1998] or with structured information like relational database information [Weikum, 2002]. Systems capable of handling both kinds of information are rare. However, in real world applications, information providers in many domains, often have to provide access to heterogeneous data sources. Systems suitable of dealing with them are very important for future Web [Burger et al., 2001; Lin, 2002].

[4] http://trec.nist.gov/
[5] http://www.hltcentral.org/page-842.0.shtml

The framework described here has been developed within the project MIETTA (Multilingual Information Extraction for Tourism and Travel Assistance), a project in the Language Engineering Sector of the Telematics Application Program of the European Commission[6] [Xu et al., 2000b].

The tourism domain is by its very nature multilingual [Tschanz and Klein, 1996], and its information is typically maintained as web documents or as database information by institutions like national or regional tourism offices.

The main objective of MIETTA is to facilitate multilingual information access in a number of languages (English, Finnish, French, German, Italian) to the tourist information provided by three different geographical regions: the German federal state of Saarland, the Finnish region around Turku and the Italian city of Rome.

In many applications, structured database information is accessed by means of forms, unstructured information through free text retrieval. In our approach, we attempt to overcome such correlations by making it completely transparent to the user whether they are searching in a database or a document collection, leaving it open to them what kind of query they formulate. Free text queries, form-based queries and their combination can yield documents and structured database information. The retrieved results are presented in a uniform textual representation in the user language.

We use automatic document translation to handle web documents, because it allows the user to access the content without knowledge of the document language and provides good retrieval performance within our limited domain. At the same time, multilingual access to the database information is supported by the combination of information extraction (IE) and multilingual generation (MG). IE identifies domain-relevant templates from unstructured texts stored in databases and normalizes them in a language-independent format, while MG produces natural language descriptions from templates. As a result, the database content becomes multilingually available for the result presentation, and natural language descriptions can be handled in the same way as web documents, namely, we can apply the advanced free text retrieval methods to them. The challenge of the approach is to merge the technologies of CLIR and natural language processing to achieve the following goals:

[6] The MIETTA consortium consists of following institutions: The technical partners are DFKI (Deutsches Forschungszentrum für künstliche Intelligenz), CELI (Centro per l'Elaborazione del Linguaggio e dell'Informazione), Unidata S.p.A., the University of Helsinki and Politecnico di Torin. The user partners are the city of Rome, Staatskanzlei des Saarlandes and city of Turku and Turku TourRing.

- Provide full access to all information independent of the language in which the information was originally encoded and independent of the query language;
- Provide transparent natural language access to structured database information;
- Provide hybrid and flexible query options to enable users to obtain maximally precise information.

In the following sections we describe how these goals can be achieved in the MIETTA framework. The chapter is organized as follows: Section 2 discusses various CLIR approaches. Section 3 and section 4 show how MIETTA combines different cross-lingual strategies to achieve its ambitious goal and describe the overall MIETTA system. Section 5 discusses how much effort has to be paid to adapt the framework to new domains and new languages. Section 6 describes some evaluation issues for the MIETTA System. Section 7 summarizes the whole approach and discusses future work.

2. STATE OF THE ART

2.1 Monolingual Information Retrieval

In the traditional and classical information retrieval praxis, a user of a search engine sends a query in the same language as the document language [Salton and McGill, 1983]. The primary data structure of IR is the so-called inverted-index. Each index entry encodes the word and a list of texts, possibly with locations within the text, where the word occurs. The match between index words and the query helps to find out which documents are relevant to the query. In a monolingual environment, the best match is that the index and the query share the most common words. Vector Space Model (VSM) and Latent Semantic Indexing (LSI) [DeerWest et al., 1990] are two widely used IR techniques.

2.2 Cross-lingual Information Retrieval

In contrast to monolingual information retrieval, CLIR aims to find documents written in a different language than the query language. A useful application of CLIR is as follows: a user is capable of reading in a foreign language, but prefers to write a query in the native language. The search

engine can find documents in the foreign language in spite of the query in another language.

CLIR is an intersected area between classic information retrieval and machine translation. Therefore, natural language processing tools are needed, e.g., language identification, multilingual dictionaries, morphological stemming, part of speech tagging, and terminology translation.

Typical CLIR strategies are based on query translation, document translation or a combination of both. We will discuss their advantages and disadvantages in the following.

2.2.1 Query Translation

The main goal of query translation is to help users to formulate their query in another language, such that the translated query can then be used as a search term (see Figure 9-1).

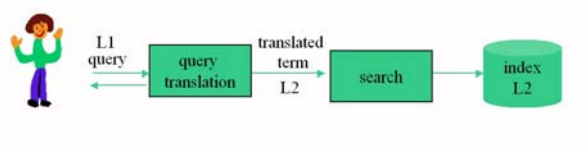

Figure 9-1: query translation in IR

Early experiments found in [Salton, 1971] showed that CLIR could work as well as monolingual information retrieval if a good transfer dictionary is available.

[Grefenstette, 1998] has summarized three problems concerning the query translation strategy:

- Finding translations
- Pruning translation alternatives
- Weighting the document relevance given translation alternatives

If a dictionary is available, the biggest problem is the poor coverage of the dictionary because of “missing word forms”, “spelling mismatches”, “compound”, etc. Some systems used the parallel corpora to construct dictionaries automatically [Chen and Nie, 2000]. To achieve reliably results, the parallel corpora should be large enough. In comparison to machine translation, in CLIR, the exact translation of a given word in a given context is not always needed in order to find the relevant documents. Some experiments have been done with LSI [Littman et al., 1998].

The primary problem of query translation is that short queries provide less context for word sense disambiguation, and inaccurate translations can lead to bad recall and precision of the search results [Carbonell et al., 1997].

The MULINEX system [Erbach et al., 1997] provides a kind of user interaction for the sense disambiguation of translated terms. However, this kind of approach is only feasible in a specific scenario, namely, if the user has enough knowledge of the target language in order to select the right sense. Retranslating possible translations back into the original query language can solve this problem to a limited degree.

The third problem is concerned with weighting the document relevance given the alternative translations of the query, e.g., if a query contains two terms with each term owning different translations, whether a document contains one translation of each term should be more relevant than a document contains many translations of a term.

2.3 Document Translation

In a document translation approach, the search index is built from automatically translated documents, such that the search becomes similar to a monolingual search, i.e., the user query can be used directly as the search term (see Figure 9-2).

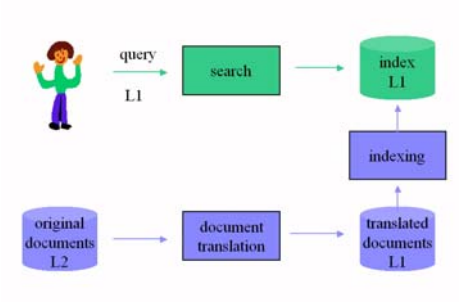

Figure 9-2: document translation in IR

The strategy would result in a higher translation and retrieval accuracy, since the full original document provides more contexts for disambiguation. Although retrieval performance still heavily depends on the quality of the underlying MT system, the word sense disambiguation problem is less severe. Therefore, this option is often preferred compared to query translation [Carbonell et al., 1997; Dumais et al., 1996]. The main limitation is, of course, that under this approach at least the indices have to be duplicated, and in the offline translation the translated documents also need to be stored. Thus, for a global search engine, this approach is practically not viable due to massive cost of computation and storage. However, the approach is quite suitable in a restricted domain where the number of documents is limited.

3. A UNIFORM FRAME FOR MULTILINGUAL ACCESS TO HETEROGENEOUS INFORMATION

3.1 Document Translation in MIETTA

In MIETTA, document translation was preferred, as it allows for direct access to the content, and provides better performance within a restricted domain. MIETTA web documents are limited to regional servers for the tourist domain, such that we do not face big storage problems. To translate the documents, the LOGOS system was employed, which covers the following directions:

German ⇒ English, French, Italian
English ⇒ French, German, Italian, Spanish

The situation to start from in MIETTA was as follows: Rome could provide documents manually translated into English, French, German, Italian and Spanish; Turku had Finnish documents, most of which were also translated into English, while the documents in the Saarland were mainly in German. The final document collection in MIETTA after the document translation yielded an almost fully covered multilingual setup.

3.2 Information Extraction and Multilingual Generation

The database information offered by the MIETTA information providers was mostly semi-structured and encoded in different languages (Italian, German and Finnish). Hence, most of the relevant information could only be found in the comment fields, mixed with other information. In order to make the database content more structured and multilingually accessible, we pursued an approach that combines IE and MG [Busemann, 1998]. The objective of IE in MIETTA is thus twofold:

- extraction of the domain relevant information (templates) from the unstructured data so that the user can access large amount of facts more accurately;
- normalization of the extracted data in a language-independent format to facilitate the MG.

MG takes a template provided by the IE component as input and generates a natural language description from it. The interaction of IE and MG is depicted in the Figure 9-3.

Figure 9-3: combination of IE and MG

A desired side effect of this strategy is that we can apply the same free text retrieval methods to generated descriptions as to web documents.

3.2.1 Information Extraction

The main task of IE is to analyze unstructured text and identify the relevant pieces of information [Cowie and Lehnert, 1996; Grishman and Sundheim, 1996; Appelt and Israel, 1997; Xu and Krieger, 2003]. One of its application areas is to detect the domain-relevant information from unstructured texts and convert it into database entries. The relevance of the information is determined by templates, which are predefined for the domain. We describe our usage of IE with the help of some examples.

The following German example text comes from the description field in an event calendar from the Saarland:

(1)
St. Ingbert: -Sanfte Gymnastik- für Seniorinnen und Senioren, montags von 10 bis 11 Uhr im Clubraum, Kirchengasse 11.

St. Ingbert: -Gentle Gymnastics for seniors, every Monday from 10:00 to 11:00 am, in Club room, Kirchengasse 11

The above text contains three pieces of information about the event, namely, event name, its location and its temporal duration:

<event> <city> St. Ingbert </city>: <name> -Sanfte Gymnastik- für Seniorinnen und Senioren </name>, <time> montags von 10 bis 11 Uhr: </time> <location> im Clubraum, Kirchengasse 11 </location>. </event>

To extract the relevant pieces of information contained in texts like the above example, we designed three steps:

- **NL shallow processing**:
  Identifying the relevant chunks of the text; for example, noun phrases and named entities (date, time, location, geographic names, phone no. and addresses).

- **Normalization**
  Converting information into a language-independent format; for example, date, time, location, addresses and phone no.
- **Template Filling**
  Mapping the extracted information into the database fields by employing specific template filler rules.

We applied SPPC [Piskorski and Neumann, 2000] for German shallow processing and IUTA[7] for Italian text analysis. For the "normalization", uniform formats are defined for date and time expressions, phone numbers, addresses, etc. E.g., "montags von 10 bis 11 Uhr" (*Every Monday from 10:00 to 11:00*) is normalized as follows:

Start time:	10:00
End time:	11:00
Weekday:	1
Weekly:	Yes

For our tourist domain, a specific set of templates was defined, corresponding to concepts like "event", "accommodation", "tours" etc. These concepts were organized in a three level concept hierarchy, drawing on the expertise of the MIETTA user partners in the tourism sector. Even if the hierarchy was designed for the MIETTA users, it can be easily adapted and generalized to other regions of tourism interest. The underlying format of the concept hierarchy is language independent. Each general concept can have several daughters, e.g., "event" has "theatre", "exhibition", "cinema" and "sports" as its sub-events. Templates in the same concept hierarchy inherit all attributes from their parents. For example, all the event templates have location, time and title as their attributes.

3.2.2 Multilingual Generation

In recent years, shallow natural language generation approaches have been shown to be quite useful for real-world applications within limited domains [Busemann and Horacek, 1998]. In particular, the combination of IE and MG provides a useful approach to a multilingual information presentation of structured information into a textual format. The basis for shallow text generation applied in MIETTA is the system TG/2. This system was developed in the TEMSIS project [Busemann, 1998], whose objective was to generate summaries of environmental data in German and French from database information. We use a JAVA implementation of TG/2, called

[7] http://celi.sns.it/~celi/projects/Iuta/iuta-top.html

JTG/2[8]. JTG/2 takes some language-independent input, applies language specific grammar rules and morphological lexicon, and returns some language-specific description. In MIETTA, five language-specific grammars were developed for the template generation. Because the JTG/2 rule formalism supports shallow grammar rules, construction new language-specific rule sets requires comparatively little effort. We illustrate the approach through a simple example from a MIETTA template. An event, such as a theater play, is encoded in a corresponding template as follows:

(2)

Level1:	Event
Level2:	Theater
Level3:	_
Event-Name:	Faust
StartDate:	21.10.99
PlaceName:	Staatstheater
Address:	Schillerplatz 1, 66111 Saarbrücken
Phone:	06 81-32204

The above template is used as the input for generation into five languages, resulting for example in the following texts:

(3)

English:
The theater show Faust will take place at the Staatstheater in Schillerplatz 1, 66111 Saarbrücken (in the downtown area). The scheduled date is Thursday, October 21, 1999. Phone: 06 81-32204

(4)

Finnish:
Teatteriesitys Faust järjestetään Staatstheaterissa, osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Tapahtuman päivämäärä on 21. lokakuuta 1999. Puhelin: 06 81-32204.

Texts like above are employed in two forms: as result presentation of the template content, and as input for free text indexing to allow advanced free text retrieval.

[8] JTG/2 is a Java implementation of DFKI's TG/2, developed by CELI, the Centro per l'Elaborazione del Linguaggio e dell'Informazione, for more information see http://www.celi.it.

To summarize, the combination of information extraction and multilingual generation is useful w.r.t. the following issues: It can make the translation of textual information in a database unnecessary, thus saving the duplication of the same piece of information in different languages. It greatly facilitates the maintenance of such data and ensures a higher degree of consistency across different languages. At the same time, it allows for an integrated or hybrid free text retrieval to both structured database and document information with the added dimension of multilinguality.

4. THE MIETTA SYSTEM

The user requirements of a tourism information system are quite varied with respect to the content. The needs for information range, on the one hand, from fairly structured information and precise facts (such as the opening times of a museum or the price of hotel accommodation), to some more general background information, as it is typically described on web pages concerning certain regions, towns, or vacation facilities. While the former type of information will be typically stored in a structured format (as a relational database), the latter is mostly available only in an unstructured format (as text documents). In order to allow the user to access these two different sorts of information in a uniform way, we provided hybrid search options:

- **Free text retrieval**
  Users can enter several words or phrases to find both web documents and descriptions generated from templates.
- **Concept based navigation**
  Users can navigate through web documents and templates according to the MIETTA concept hierarchy.
- **Form-based search**
  Users can select fields in a search form to access templates.

Our motivation is to make it completely transparent to users which source of information they are searching in, and to allow them to formulate their query as precise as they desired. In order to realize our goal, we developed our framework which integrated intelligently the existing techniques occurred in the cross-lingual free text retrieval and natural language processing communities. The implementation of the framework is our MIETTA system, which contains following three main components:

- **Data Capturing**
  collecting web documents and recording document information such as the title, URL, manual and automatic classifications, etc.

- **Data Profiling**
  document translation, IE, offline MG of templates, free text indexing.
- **Search Engine**
  search and visualization of the search result.

The interaction of the three components is illustrated in Figure 9-4.

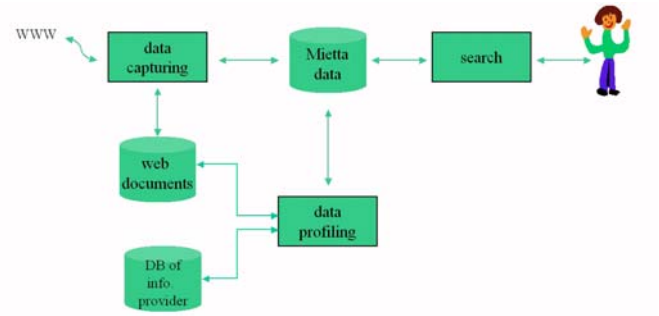

Figure 9-4: MIETTA overall system.

4.1 Data Capturing

A large part of the basic material provided by the MIETTA user partners comes in the form of web documents. To make this information accessible through the MIETTA search tools, they must be registered, gathered and indexed. The data-capturing tool provides a convenient user interface, which allows the information providers to register their web pages. The registration user interface provides a broad range of facilities: The user can enter the URL of the page or the site, the depth to be disclosed, and to enter the address of a potential document translation. Furthermore, he can also classify the page content based on the MIETTA concept hierarchy. Besides the URL registration and manual classification, the data capturing tool also integrates the JumboScan Web crawler package provided by UNIDATA[9] for downloading and indexing the document information like the URL name, title etc.

[9] See http://www.unidata.it/.

4.2 Data Profiling

The aim of the data-profiling component in the MIETTA system is to disclose the data sources in such a way that access through different search options becomes possible. It contains

- Document translation, based on the LOGOS machine translation system;
- IE from database entries for template construction;
- MG from templates to obtain natural language descriptions;
- Free text indexing

The first three aspects have been discussed in the last section. The result of applying these processes is that both web documents and template information becomes available in the different languages covered by MIETTA. We will focus on the free indexing work here. We employ an existing indexing and search tool developed by TNO[10] [Hiemstra and Kraaij, 1998] for the free text retrieval task. The TNO indexing tool generates two kinds of indexes:

- A lemma-based fuzzy index based on tri-grams (ISM);
- A Vector Space Model (VSM) index based on lemmatas.

These free-text indexing components are applied to web documents, translations and descriptions generated from templates, see the following Figure 9-5.

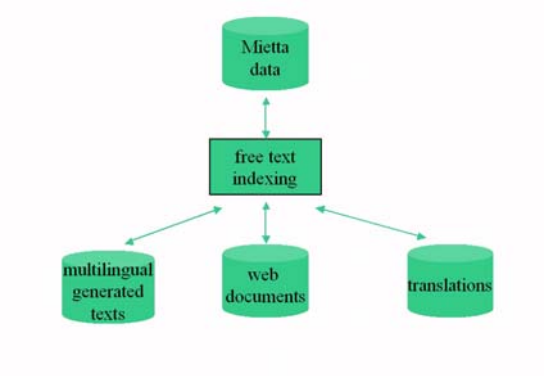

Figure 9-5: free text indexing

The indexing of the automatically generated descriptions adds a specific functionality to the MIETTA search engine, as the template content becomes accessible for fuzzy search and the vector space model search, which are normally not supported by relational database. As a result, both web

[10] See http://twentyone.tpd.tno.nl/.

documents and database information becomes available in a textual format in different languages, disclosed through classifications and a free text indexes.

4.3 Search Engine

We describe here the hybrid search options provided by MIETTA and how they are realized. With the help of data capturing and data profiling, the MIETTA search engine allows for standard free text retrieval as well as the following advanced search capabilities: concept-based navigation, and form-based query (template search).

The architecture of the search engine is described in Figure 9-6.

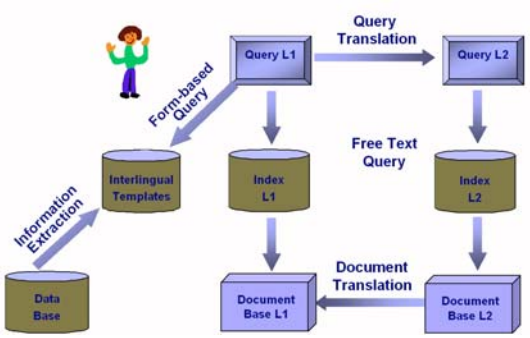

Figure 9-6: search architecture

As mentioned above, MIETTA uses the existing TNO ISM/VSM search engine for free text retrieval. The ISM part makes use of a kind of a fuzzy matching algorithm based on tri-grams. It allows the match of index terms with query words or phrases containing spelling errors or morphological variants. For example, the user can enter "baroque palaces" and find documents and template descriptions containing the phrase "baroque styled palace". In addition to the free text retrieval, the user can also navigate through the concept hierarchy to search for information in a certain category. In contrast to many other search engines, the MIETTA user can also combine the free text retrieval with the concept-based navigation by formulating a query with constrains such as "find all documents containing the word *colosseo* belonging to the category *Art and Culture*", see Figure 9-7.

Figure 9-7: search menu

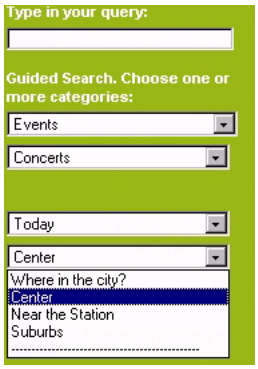

Figure 9-8: form based query

A more restricted and goal-directed query is the form-based query, where the user can select fields in a template form. For example, the user can select the "time" and the "location" fields of a "concert" event template by using a query form. In Figure 9-8, the user has formulated a query corresponding to the constraint "give me all information about concerts in the city center today".

All queries are processed by the query- processing component and are converted either into a standard SQL query or an ISM/VSM query. The result of the retrieval is presented as a uniform list of links to textual descriptions (generated from templates) and to web documents. Both types of information are presented, on the one hand, in an absolute ranking order, where only the relevance of the document plays a role, and on the other hand, sorted according to the different categories. If the user clicks on a link, they receive either a web document or a generated text as in the examples (3) and (4).

To summarize, the MIETTA search engine represents a flexible way of combining cross-lingual free text retrieval with standard database access. The hybrid query options and their interaction provide the user with a highly versatile range of options to express their different search requirements, which is also reflected in the presentation of the results and the further navigation options.

4.4 Scalability of the Framework

Our framework can be easily adapted to other domains. The domain modeling consists mainly of the definition of domain specific templates and the concept hierarchy. With respect to the IE task, the major part of the effort needs to be spent on the new definition of the template filler rules, since the

natural processing and the normalization steps are domain independent. The MIETTA MG tool has already been proven to be reusable, as it has been applied in both the TEMSIS environmental and the MIETTA tourist domains.

When it comes to adding a new language, three language dependent components would be involved in the MIETTA system: natural language generation (JTG/2), document translation and natural language processing. As mentioned in section 2, our natural language generation tool requires less effort for the development of a grammar rule set in a new language. It supports also easy integration of a morphological component. In addition to the five generation-grammars developed within MIETTA, we also carried out some successful experiments with a Chinese grammar. Thus, integrating a new language outside the Western language family into JTG/2 also appears to be quite easy. Document translation is clearly dependent on the language pairs supported by the machine translation system employed, i.e., essentially independently from the MIETTA system itself. However, to deal with the problem of unavailable translation pairs, a statistical translation model, as proposed in [Chen and Nie, 2000] could be employed. Such translation models are much easier to establish than MT systems, and it would be sensitive to the domain of the training corpus. Similar to the scalability of the machine translation, the shallow language processing components are dependent on the developments for a certain language. In the recent years, the natural language processing society in Asia is very active, in particular, as far as Chinese, Japanese and Korean are concerned. Hence, the integration of one of these languages is realistic.

4.5 Evaluating MIETTA

Because of the broad variety of search strategies and the heterogeneous data sources, the standard relevance assessment model used in the ad hoc and routing forums of TREC is difficult to apply to the complete MIETTA system. The evaluation of the individual components such as the TNO free text retrieval engine, the natural generation system and the IE tools SPPC and IUTA can be found in their corresponding literature mentioned in the previous sections. We consider that an end user centered evaluation should be suitable for such complex systems, including following criteria:

- transparency and ergonomics of the search user interface and the result presentation,
- quality of machine translation and natural language generation,
- interaction of the hybrid search options.

5. CONCLUSION

We have presented a novel framework for the uniform and multilingual access to web documents and the structured data and have implemented a practical application that successfully realized this framework. The MIETTA system allows the user to carry out a cross-lingual search in different sources of information at different levels of content granularity. This framework is highly suitable as a domain-specific information system and internet-portal. It can be easily transferred to other domains and is extensible to other languages.

Future work will be directed towards extending the framework to IE from web documents and to fully automatic document classification. In the current MIETTA system, template extraction from web documents combined with MG has not been considered due to limited resources. Such a combination, however, would make the system even more effective as it could provide both summaries of web documents and multilingual access to such summaries.

In the most recent search engine research, question-answering systems are regarded as the future generation of information system. In comparison to the classical information retrieval systems, question-answering systems allow users to send questions as queries to receive answers instead of just a list of documents. In the roadmap of the future question answering systems [Burger et al., 2001], the multilingual and cross-lingual systems and heterogeneous data sources belong to the important issues. The MIETTA strategy, namely, the combination of IE and MG can be adopted for answer generation in a cross-lingual and multilingual environment. In addition, the uniform framework for dealing with heterogeneous data sources is a good experiment for the future question answering systems [Lin, 2002].

ACKNOWLEDGEMENTS

The framework described here is grounded on the cooperation in the MIETTA consortium. I am grateful to the MIETTA colleagues for the useful discussions and contributions.

REFERENCES

Appelt and Israel. 1997. Building Information Extraction Systems. *ANLP-97 Tutorial,* 1997.

C. P. Braschler, and P. Schäuble.1999. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the eighth Text REtrieval Conference (TREC-8*), held in Gaithersburg, Maryland, November 17-19, 1999.

J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, S.Harabagiu, D. Israel, C. Jacquemin, C. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees and R. Weischedel. 2001. Structures to Roadmap Research in Question and Answering (QandA). In *NIST DUC Vision and Roadmap Document*, 2001.

Stephan Busemann. 1998. Language Technology for Transnational Web Services. In *Proceedings of European Telematics: Advancing the Information Society Telematics Applications Programme Annual Concertation Meeting*, Barcelona, 1998, 101-105.

Stephan Busemann and Helmut Horacek. 1998. A Flexible Shallow Approach to Text Generation. In Eduard Hovy (ed.): In *Proceedings of the Nineth International Natural Language Generation Workshop (INLG '98),* Niagara-on-the-Lake, Canada, August 1998, 238-247.

Jaime Carbonell, Yiming Yang, Robert Frederking, Ralf D. Brown, Yibing Geng and Danny Lee. 1997. Translingual Information Retrieval: A comparative evaluation. In *Proceedings of the Fifteeth International Joint Conference on Artificial Intelligence*, August 1997.

Jiang Chen and Jian-Yun Nie. 2000. Parallel Web Text Mining for Cross-Language IR. In *Proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIAO 2000)*, Paris, 2000.

Cowie and Lehnert. 1996. Information Extraction. In *Communications of ACM*, 39(1):51-78, 1996.

D. Crystal. 1997. *English as Global Language*. Cambridge University Press. 1997.

Mark W. Davis and William C. Ogden. 1997. Implementing cross-language text retrieval systems for large-scale text collections and the world wide web. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, March 1997.

Scott Deerwest, Susan T. Dumals, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391-407, 1990.

S. Dumais, T. Landauer and M. Littman. 1996. Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. In *Proceedings of SIGIR-96*, Zurich, August 1996.

G. Erbach, G. Neumann and H. Uszkoreit. 1997. MULINEX - Multilingual Indexing, Editing and Navigation Extensions for the World Wide Web. In David Hull and Doug Oard (eds.) *Cross-Language Text and Speech Retrieval — Papers from the 1997 AAAI Spring Symposium*, AAAI Press, Menlo Park, 1997.

Djoerd Hiemstra and Wessel Kraaij. 1998. Twenty-One in ad-hoc and CLIR. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7),* NIST special publication, 500-240.

Gregory Grefenstette. 1998. The problems of cross-language information retrieval. In *G. Grefenstette (ed.), Cross-language Information Retrieval*. *Chapter 1*. Kluwer Academic Publishers, Boston, 1998.

Gregory Grefenstette. 2001. Multilinguality on the Web. http://www.infonortics.com/searchengines/sh01/slides-01/grefen.pdf. 2001.

G. Grefenstette and Julien Nioche. 2000. Estimation of English and non-English Language Use on the WWW. In *Proceedings of RIAO'2000*. Paris, April. 2000.

R. Grishman and B. Sundheim. 1996. Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING),* pages 466-471, Kopenhagen, Denmark, Europe, 1996.

Jimmy Lin. 2002. The Web as a Resource for Question Answering: Perspective and Challenges. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands, Spain. May, 2002.

Michael L. Littman, Susan T. Dumais, and Thomas Laudauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefensttte (ed.), *Cross-language Information Retrieva*l, chapter 5. Kluwer Academic Publishers, Boston, 1998.

G. Nunberg. 1996. The Whole World Wired. Commentary broadcast on 'Fresh Air', National Public Radio, Sept. 1996.

Douglas W. Oard. 1999a. Global Access to Multilingual Information. Presented at *IRAL99*, Taipei, 1999.

Douglas W. Oard. 1999b. Cross-Language Text Retrieval Research in USA. In *3rd ERCIM DELOS Workshop*, Zurich, Switzerland, 1999.

J. Piskorski and G. Neumann. 2000. An Intelligent Text Extraction and Navigation System. In *Proceedings of 6th International Conference on Computer-Assisted Information Retrieval* (RIAO-2000), Paris, 2000.

G. Salton 1971. *Automatic Processing of Foreign Language Documents*. Preentice-Hall, Englewood Cliffs, NJ, 1971.

Gerard Salton and M. McGill. 1983. *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

Peter Schäuble and Páraic Sheridan. 1997. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of TREC-6*, 1997.

V. Trau. 1999. Characterisations of Web Server Contents. FH Wiesbaden, University of Applied Sciences, Web Seminar, January, 1999.

Tschanz and S. Klein. 1996. Web-enabled Cooperation in Tourism. In *Proceedings of EMOT Work shop*, Modena, Italian 1996.

Gerhard Weikum. 2002. The Web in Ten Years: Challenges and Opportunities for Database Research. Invited Keynote, In *Proceedings of 10th Italian Database Conference (SEBD: Sistemi Evoluti per Basi di Dati*), Portoferraio, Isola d'Elba, Italy, 2002.

J. L Xu. 2000a. Multilingual Search on the World Wide Web. In *Proceedings of the Hawaii International Conference on System Sciences HICSS-33*, Maui, Hawaii, January, 2000.

Feiyu Xu, Klaus Netter and Holger Stenzhorn. 2000b. MIETTA-A Framework for Uniform and Multilingual Access to Structured Database and Web Information. In *Proceedings of Information Retrieval for Asian Language (IRAL 2000),* Hong Kong. 2000.

Feiyu Xu and Hans-Ulrich Krieger. 2003. Extraction of Domain-Specific Events and Relations via a Combination of Shallow and Deep NLP. DFKI research report. 2003.