A System for Uniform and Multilingual Access to Structured Database and Web Information in a Tourism Domain

Feiyu Xu, Klaus Netter, Holger Stenzhorn
DFKI Language Technology Lab
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
{feiyu, netter, holger}@dfki.de

Thematic Session: Application systems using language technology components

Keywords: structured database, web documents, crosslingual information retrieval, machine translation, information extraction and multilingual generation

Abstract

We present an information system, which was 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. MIETTA facilitates multilingual information access in a number of languages (English, Finnish, French, German, Italian) to the tourist information (web documents and database information) provided by three different geographical regions: the German federal state of Saarland, the Finnish region around Turku and the Italian City of Rome.

The challenge of the approach is to merge the technologies of crosslingual information retrieval (Jamie Carbonell et al, 1997) and natural language processing to achieve the following goals:

- Provide full access to all information independent of the language the information was originally encoded in 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.

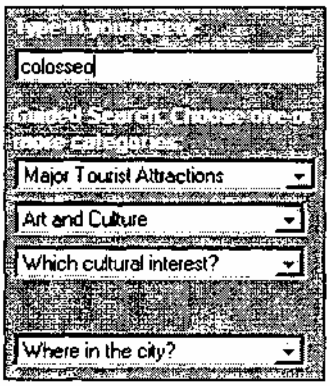
For the purpose of cross-lingual retrieval, we apply two different methods. We use offline automatic document translation to be able to construct indices from web documents in others than the original document language. This allows the user to access the content of a document without knowledge of the 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 (Piskorski and Neumann, 2000) and multilingual generation (Busemann and Horacek, 1998). Information extraction extracts domain-relevant templates from database and normalizes them in a language-independent format, while multilingual generation 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 advanced free text retrieval methods to them.

As for query and navigation options, it can be observed that 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 user can formulate their query in their own language, while the retrieved results are presented in a uniform textual representation in their query language too.

The hybrid search options provided in MIETTA are:

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

MIETTA uses the existing TNO ISM/VSM search engine for free text retrieval (Hiemstra and Kraaij, 1998). The ISM part makes use of a fuzzy matching algorithm based on trigrams. It allows to match 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 which contain 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 the Figure 1.



Type in your query:

colossed

Guided Search: Choose one or more categories:

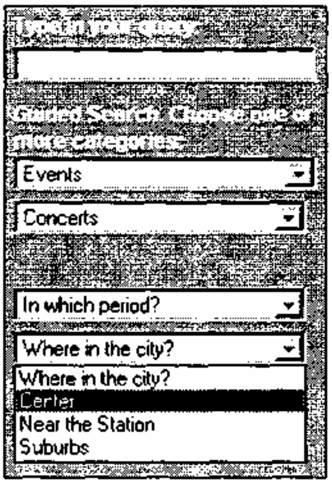
Major Tourist Attractions

Art and Culture

Which cultural interest?

Where in the city?

Figure 1



Type in your query:

Guided Search: Choose one or more categories:

Events

Concerts

In which period?

Where in the city?

Where in the city?
Center
Near the Station
Suburbs

Figure 2

More restricted and goal-directed 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 the following example, the user has formulated a query corresponding to the constraint *"give me all information about concerts in the city center today"* (see Figure 2).

All queries are processed by the query processing component and 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 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 from a template.

To summarize, the MIETTA search engine represents a flexible way of combining crosslingual 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.

Acknowledgements

This system is developed within the MIETTA consortium. We are grateful to Paul Buitelaar, Olga Goldmann and to our colleagues at the MIETTA Partner institutions: Luca Dini, Vittorio Di Tomaso and Giampaolo Mazzini in CELI (Centro per l'Elaborazione del Linguaggio e dell'Informazione), Alessandro Giarante, Marcello Vispi in Unidata S.p.A., Kimmo Koskenniemi, Jyrki Niemi in the University of Helsinki, Elena Baralis and Rosa Meo in Politecnico di Torin, our project coordinators Francesco Bellini and Audrey Boss in Comune di Roma and our user partners.

References

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

Jaime Carbonell, Yimying 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.

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

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