Multilingual Information Service System for the Beijing 2008 Olympics Forum

May 21. 2004

Keynote

Title:
Modern Multilingual and Cross-lingual Information Access Technologies

Authors:
Hans Uszkoreit and Feiyu Xu

Email:
{uszkoreit, feiyu}@dfki.de

Affiliation:
DFKI GmbH
LT-Lab
Stuhlsatzenhausweg 3
66123 Saarbruecken
Germany

Abstract:
The presentation will summarize the state of the art in cross-lingual and multilingual information systems and related areas. We will start with an outline of basic problems and major approaches [Uszkoreit, 1998]. In order to demonstrate the relevant concepts and strategies, we will present several multilingual information systems that we have realized in various projects, e.g., MIETTA, MULINEX, GETESS and MUMIS ([Xu et al., 2000a], [Capstick et al. 1999], [Staab et al., 1999], [Declerck and Wittenburg, 2001]), In conclusion, we will discuss how some of the developed ideas may be exploited by the *Multilingual Information Service System for the Beijing 2008 Olympics*.

Above all, we will present a framework ([Xu, 2003], [Xu et al, 2000a] and [Xu et al., 2000b]), which permits uniform and multilingual access to heterogeneous data sources, e.g., web pages and database contents. Within this framework, it is easy to build up a multilingual information service system for relevant domains such as weather, public traffic, sports information, travel, restaurants, hotels, etc. 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. The realistic integration of crosslingual information retrieval and natural language processing techniques helps to achieve the following goals:

- 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.

We will demonstrate that the framework is useful for domain specific and multilingual applications and can be also adapted to advanced question answering systems. This framework has been developed, tested and validated in the EU-funded project MIETTA. The MIETTA system provides multilingual information access to heterogeneous 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. The covered languages are English, Finnish, French, German and Italian.

Real-world application scenarios such as the Olympic Games require that the information system supports content providers in offering multilingual content-based access to multimedia material (e.g., image/video and textual documents). In the 2008 Olympics context, quick access in several languages to video snippets of certain sport events would constitute an extremely useful service for sports fans, journalists and officials. We will present technologies developed within the EU-funded project MUMIS that support automatic indexing of multimedia recordings and retrieval of indexed multimedia archives [Declerck and Wittenburg, 2001]. The test domains of MUMIS are Soccer Games, e.g., the UEFA 2000 and FIFA 2002 championships.

Information extraction (IE) techniques play a crucial role in building robust and advanced multilingual information systems, such as MIETTA and MUMIS, since IE helps to analyse and extract relevant information from large volumes of textual data in a realistic time. IE also delivers input for further processing and applications. In recent years, IE systems have become commercially viable by supporting diverse information discovery and management tasks. We will present a multilingual information extraction system developed by DFKI LT-Lab, called SProUT [Becker et al., 2002]. SProUT is a platform for the development of multilingual shallow text processing system. SProUT has been adopted as the core information extraction component in several EU-funded and industrial projects, facilitating various tasks such as content extraction and acquisition for text/data mining, dynamic hyperlinking [Busemann et al., 2003], machine translation, text summarization and question answering. At the current stage, SProUT supports about ten languages including: Chinese, Japanese, English, Dutch, French, German, Italian, Polish and Czech.

Furthermore, we will provide a brief description of recent developments in multilingual and crosslingual question answering systems ([Neumann and Xu, 2003] and [Neumann et al., 2003]).

We expect that the presented insights, methods and systems will constitute a valuable constribution to the definition and realization of the envisaged advanced IT solution that supports linguistic diversity and overcomes language barriers in the Olympic Games of 2008. Therefore, we propose to develop a powerful software platform supporting content providers in offering multilingual content-based access to multimedia material (e.g., image/video and textual documents). This is a large-scale distributed and cooperation-oriented software system containing specialized intelligent multilingual information agents. The system processes natural language user queries and large volumes of heterogeneous multimedia data resources. It integrates several existing advanced but tested NLP/AI technologies in a novel way.

Nevertheless it is realistic and robust since it always provides matured state-of-the-art database technology as a safe fallback in cases where the system cannot deal with unexpected language input. Such unexpected input can be due to speech impairments, strong dialectal variation, grammatical errors, noisy signals, or overly complex utterances. Thus the technology will provide its everyday users equipped with the appropriate connecting technology with robust direct interactive access to exactly the multimedia content and information they are interested in, expressed in their own languages.

References

[Uszkoreit, 1998] Uszkoreit, H. (1998): Cross-Lingual Information Retrieval: From Naive Concepts to Realistic Applications. In: Language Technology in Multimedia Information Retrieval, Proceedings of the14th Twente Workshop on Language Technology.

[Xu et al., 2000a] Feiyu Xu , Klaus Netter and Holger Stenzhorn. *A System for Uniform and Multilingual Access to Structured Database and Web Information in a Tourism Domain*. In Proceedings of ACL 2000 Demo Session, Hong Kong.

[Capstick et al. 1999] Capstick J., A. K. Diagne, G. Erbach, H. Uszkoreit , A. Leisenberg, and M. Leisenberg (1999) A System for Supporting Cross-Lingual Information Retrieval, In: Amanda Spink & Jian Qin (eds). Information Processing and Management - Special topic issue "Web Research and Information Retrieval", 1999.

[Staab et al., 1999] Staab, S., C. Braun, A. Düsterhof, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, B. Wrengers (1999) Getess - searching the Web exploiting German texts. In CIA 99 - Proceedings of the 3th workshop on Cooperative Information Agents, LNCS, Berlin, Springer, 1999.

[Xu, 2003] Feiyu Xu. *Multilingual WWW --- Modern Multilingual and Cross-lingual Information AccessTechnologies*. In Knowledge-Based Information Retrieval and Filtering from the Web. Witold Abramowicz (Ed.), Kluwer Academic Publishers, page 165--184, 2003.

[Xu et al., 2000b] Feiyu Xu, Klaus Netter and Holger Stenzhorn. *MIETTA-A Framework for Uniform and Multilingual Access to Structured Database and Web Information*. In Proceedings of IRAL 2000, Hong Kong.

[Declerck and Wittenburg, 2001] Thierry Declerck and Peter Wittenburg. *MUMIS -- A Multimedia Indexing and Searching Environment*. In Proceedings of the First International Workshop on MultiMedia AnnotationP, MMA-2001. Tokyo.

[Becker et al., 2002] M. Becker, W. Drozdzynski, H.U. Krieger, J. Piskorski, U. Schäfer, Feiyu Xu. *SProUT - Shallow Processing with Typed Feature Structures and Unification*. In the Proceedings of ICON 2002 - International Conference on NLP, Mumbai, India, 2002.

[Busemann et al., 2003] S. Busemann, W. Drozdzyski, H.U. Krieger, J. Piskorski, U. Schäfer, H. Uszkoreit, Feiyu Xu. *Integrating Information Extraction and Automatic Hyperlinking* In Proceedings of ACL-Demo Session 2003, SAPPORO, Japan.

[Neumann and Xu, 2003] Günter Neumann and Feiyu Xu. *Mining Answers in German Web Pages*. In Proceedings of IEEE/WIC WI-2003, Halifax, Canada, 2003.

[Neumann et al., 2003] Günter Neumann, Feiyu Xu and B. Sacaleanu. *Strategies for Web-based Cross-Language Question Answering*. In proceedings of 2nd CoLogNET-ElsNET Symposium on Questions and Answers: Theoretical and Applied Perspectives, Amsterdam, 18 December, 2003.