MULINEX: Multilingual Web Search and Navigation[♣]

Joanne Capstick, Abdel Kader Diagne, Gregor Erbach, Hans Uszkoreit
German Research Center for Artificial Intelligence, Saarbrücken
Francesco Cagno, Giovanni Gadaleta – Datamat, Rome
Juan Antonio Hernandez – Grolier Interactive Europe, Paris
Rene Korte, Anne Leisenberg, Manfred Leisenberg – Bertelsmann Telemedia, Gütersloh
Oliver Christ – Trados, Stuttgart and Brussels

http://mulinex.dfki.de/
mulinex@dfki.de

Abstract

MULINEX is a multilingual search engine for the WWW. During the phase of document gathering, the system extracts information about documents by making use of language identification, thematic classification and automatic summarisation. In the search phase, the users' query terms are translated in order to enable search in different languages. Search results are presented with a summary and information about the language and thematic categories to which the document belongs. Summaries and documents are translated on demand by making use of the LOGOS machine translation system. The system is to be deployed in the online services of Bertelsmann Telemedia and Grolier Interactive Europe, and supports French, German and English. The current MULINEX prototype is the first system for translingual information access integrating retrieval, summarisation and translation.

Keywords
translingual information retrieval, categorisation, summarisation, language identification, query translation, machine translation

1 Motivation

The Internet is rapidly changing from an English dominated medium to a multilingual information and communication service. At present navigation in this multilingual information space is still far from the ideal scenario – the ability to access, in one's own mother tongue, the mass of multilingual documents over the Internet in a seamless and transparent fashion.

1.1 Social and Economic Factors

The total number of online households in the world is expected to rise from 23.4 million (8.3 % of all households) in 1996 to more than 60 million (near 25 %) in 2000. Increased PC penetration, telecommunications deregulation, indigenous content development, and deployment of integrated services digital network (ISDN) in Europe's and Asia's most advanced online markets will be among the key factors driving this growth.

By the year 2000, countries outside the US (especially in Europe and Asia) will account for 46 per cent of all online households, up from 37 percent today. France, Germany, Japan and the UK will be the next largest online markets in 2000. Scandinavia and Italy will experience growth rates that are at least as large as the major markets.

MULINEX addresses some of the issues of multilinguality by developing a leading-edge application that facilitates multilingual information access with navigation and browsing, enabling effective multilingual searching on the Internet by providing translation of queries, customised summaries, and thematic classification of documents.

MULINEX processes multilingual information from the WWW and other online sources and presents it to the user in a way which facilitates finding and evaluating the desired information

[♣] The work reported here was financially supported by the European Union's Telematics Application Programme, contract LE-4203 in the sector Language Engineering.

quickly and accurately. It does this by combining the newest information retrieval technology with advanced language technologies to improve search and navigation in the WWW.

The current MULINEX prototype is the first system for translingual information access integrating retrieval, summarisation and translation.

1.2 Consortium Objectives

The MULINEX consortium consists of five European companies, who aim to improve their competitiveness in the internet market through the development and application of advanced language technology for providing improved user-friendly web search and navigation services.

The co-ordinating partner DFKI conducts basic and application-oriented research in artificial intelligence and other fields of advanced computer science. DFKI's Language Technology Lab applies cutting-edge language technology to a variety of application areas. It develops large scale resources such as lexicons, morphology and grammars.

The two user partners provide web services. Grolier Interactive Europe operates Club Internet and creates and hosts websites for commercial customers. Bertelsmann Telemedia defines itself as an Internet Solutions Company which provides online services and is active in the area of electronic commerce.

DATAMAT specialises in system integration, has been involved in the development of a second generation text retrieval software called Fulcrum SearchServer, which is used in the project as the basic full-text retrieval system.

TRADOS develops translation tools, terminology database systems and translation memory systems.

2 User Requirements

The analysis of user needs by Grolier Interactive Europe and Bertelsmann Telemedia has established the following requirements for a multilingual web search engine (Hernandez 1997):

- *Average response time:* Two seconds should be the upper limit for the response time (one on average).
- *Restriction of the search:* It is considered very important to allow the user to restrict the search to a certain subject area. A search by internet domain or by website is also useful, especially if the user wants to explore one particular site.
- *Presentation of search results:* The system should be a powerful aid to decision making; that is, at the end of the search process, a user should know what to expect from the URL he is given: a web site with information about the requested subject, a page with links, a web site in one's own or in a foreign language. At the end of the search process, a short abstract or a list of keywords (in the language of the document) should be generated. Then the user may request a translation of the abstract or of the keywords.
- *Level of translation*: An indicative level of translation is generally sufficient for most users, since most users are "zappers": once they know what it is about, they switch to another subject. On the other hand, certain users are specialists, and once again they just want to know what it is about in order to make a decision.
- *Personalisation*: it should be possible for the user to register with the system and to save a set of preferences for future sessions (such as interface language, language preferences, preferred thematic categories, presentation preferences such as frames or Java).

The following functionalities were listed as desirable additions, which add value to the MULINEX system, compared with other search engines.

- *Query construction tools:* When using search engines it is quite common that the first answer is not the right one, and hence a tool for refining the search would be welcome.
- *Subscribing to a certain field:* This functionality scans online news and web sites according to the user's interest profiles, and notifies the user if new information becomes available. Example applications include staying up-to-date with certain events, e.g. in sports, or monitoring the competition in a certain area of business.

2.1 Questionnaires and Interviews

Bertelsmann Telemedia have conducted a survey with about 70 internal and external users, the results of which are summarised in the following:

- The majority of participants (86%) is interested in WWW documents written in a known foreign language.
- Only 22% of our participants are interested in search results in unknown foreign languages.
- Automatic translation of retrieved WWW documents is required by the majority (67%) of the end users.
- Preferred queries are AND/OR-combinations of words or strings.
- The majority of end users wants to be able to restrict the search to particular subject areas or languages.
- 65% of end users want a search engine that translates the query and performs the search in other languages.

2.2 Psychological Experiments

For evaluating various design options of the MULINEX User Interface, a psychological experiment was carried out jointly by DFKI and MEFIS, an institute which specialises in media psychology. The purpose of the experiments was to find out how the system can best support the user in the choice of relevant documents on the results page. The following questions were addressed in the experiment:

1. *How should the search results be ordered?* According to relevance **or** according to subject area.
2. *What type of document summary should be used?* The first x characters of the document **or** an automatically generated query-independent summary **or** an automatically generated summary tailored to the query.
3. *In which language should the summary be presented to the user?* The language of the document **or** a translation into a language of the users choice.

A total of 84 German subjects were tested with a mock-up system which presented subgroups with various design alternatives. The subjects were given the task to submit a predetermined query to a search engine, and to select documents which were relevant to a given information need by looking through lists of potentially relevant documents. The information needs were formulated as follows:

- What is good and bad for the heart?
- What are the effects of ozone on human health?

We gathered data about the background of the users (about their language skills, computer skills etc.), their impression of the system (by making use of a semantic differential designed for software evaluation and open-ended questions), and their performance (by analysis of the log files).

The users reacted very positively to the (manually constructed) thematic classification of documents. Document summaries were criticised as being too short and uninformative, and different kinds of summaries had no significant quantitative effect on the subjects' performance. There was a subgroup of subjects which made extensive use of automatic translations of summaries. Translations were considered useful although their quality was criticised.

3 Functionality

MULINEX is a multilingual Internet search engine that supports selective information access, navigation and browsing in a multilingual environment. During the phase of document gathering by the web spider, documents are analysed in order to obtain useful information about documents in addition to the traditional keyword-based indices. The project emphasises a user-friendly interface, which supports the user by presenting search results along with information about language, thematic category, automatically generated summaries, and allows the user to sort results by multiple criteria. Translingual search (Greffenstette 98) is supported by interactive translation of user queries. Commercial machine translation technology (LOGOS) is used to provide translations of foreign-language documents on demand.

The demonstrator includes the following functionalities:

- A document gatherer (web spider) which performs language identification, thematic document classification and document summarisation.
- Translation of the user's query
- Simultaneous search in English, German and French document collections
- Automatically generated document summaries

- On-demand translation of summaries and search results.
- Registration of users and user preferences for a personalised search environment

4 System Architecture

The underlying architecture is **object-oriented** and **manager-based**. The benefits of an object-oriented architecture are increased modularity, flexible systems and ease of reusability through inheritance of structures and components. The main characteristic of a manager-based approach is encapsulation of the interaction of components, leading to increased independence between components and objects.

4.1 Architecture Overview

The MULINEX system consists of weakly coupled subsystems. Subsystems communicate with each other through their managers by sending requests and receiving corresponding responses. Requests and responses are encapsulated in objects. According to the current architecture we identify the following subsystems:

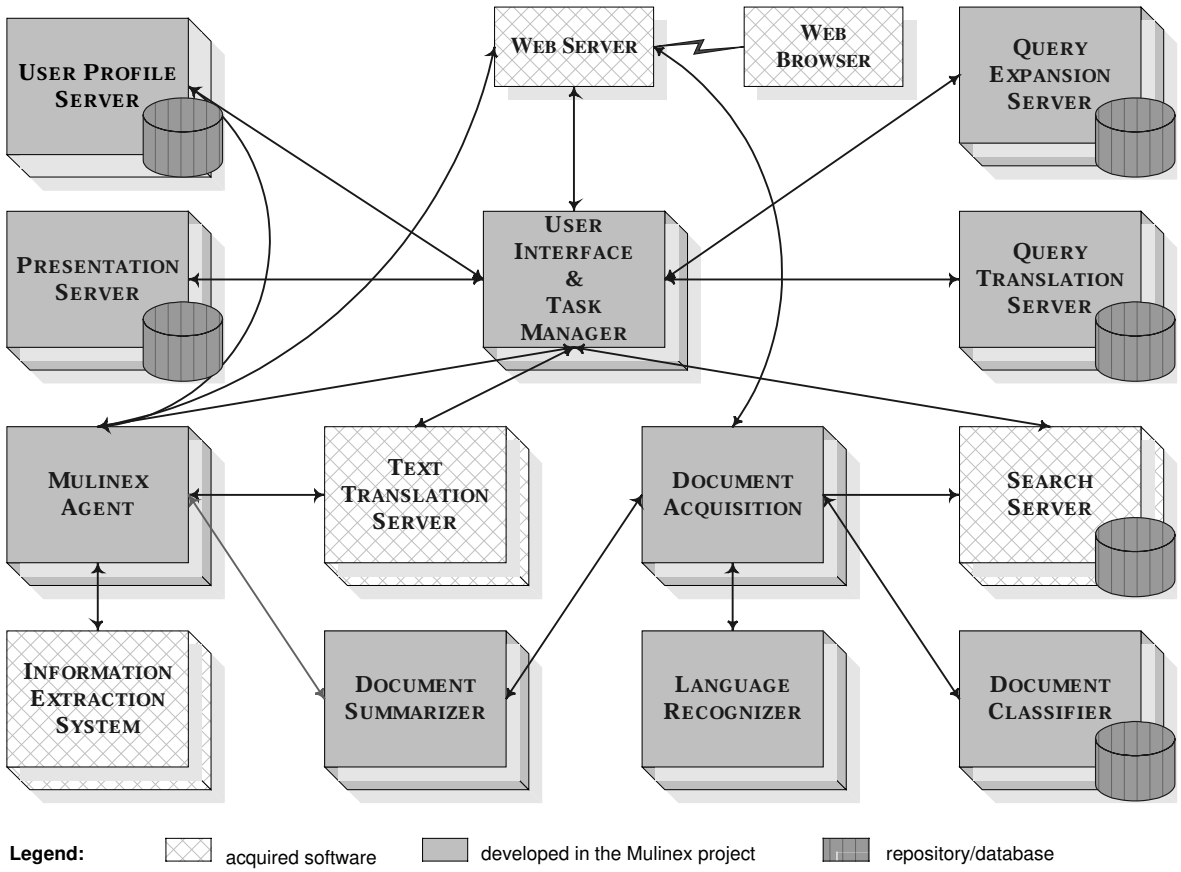

Figure 1 - A subsystem-oriented view of the architecture of Mulinex.

- **User Interface:** makes the system's functionality available to the user by enabling the communication between the user and the system.
- **User Profile Server**: provides information about a registered user and encapsulates the user profile repository which contains relevant information about registered users (user profiles, agent configuration, etc.)
- **Presentation Server**: encapsulates the presentation repository which contains all presentations (HTML files, templates for forms and e-mails). The presentation repository is encapsulated by a presentation manager which

generates or customises specific pages (e.g. result pages). The presentation manager handles all aspects of presentations which determine the concrete user interface.

- **Task Manager:** The task manager is the main control and co-ordination unit of the system. It receives a user request from the user interface, performs request dispatching, elaborates a work plan, executes it, and constructs the top-level response which is returned to the user.
- **Search Server:** the search subsystem consists of a search manager and a search server. The search manager receives search requests from the task manager, adapts them to the interface of the underlying search server and invokes corresponding operations.
- **Query Translation Server:** provides a translation of a query or query terms in such a way that the translated query can be used to retrieve documents in the corresponding target language. It encapsulates dictionaries and term translation subcomponents.
- **Query Expansion Server**: provides mechanisms and methods for concept-based query expansion.
- **Text Translation Server:** consists of a translation manager and a translation module. The translation module is a machine translation system with a specified interface. The translation manager receives requests from the task manager, adapts it to the interface of the underlying translation module, invokes the corresponding operations, and returns the computed result.
- **Document acquisition:** provides functionality for gathering documents from the web and preparing them in such a way that they can be easily used to populate the search server's database.
- **Language Recognition**: provides functionality for identifying the language of a given document.
- **Document Summariser**: produces a summary of a given text.
- **Document Classifier:** provides functionality for classifying acquired documents with respect to pre-defined categories which are maintained in a database.
- **Mulinex Agent**: A server-side software agent which provides advanced functionality to assist the user in the process of finding the right information. It autonomously performs tasks like initiating search requests on user-selected topics, informing the user when new documents are available, etc.
- **Information Extraction System**: extracts relevant information from documents that belong to specific categories.

4.2 Software Engineering Methodology

We develop the system in an iterative and incremental process through analysis, design, implementation, and testing. We started by developing a use case model (Jacobson et al. 1992) and defining the corresponding scenario diagrams. This way we could easily identify the main classes that carry out the system's functionality. The use case model specifies the functionality of the system. It controls the formation of all other models; i.e. the functionality specified by the use cases is structured by the analysis model, realised by the design model, implemented by the implementation model and tested in the testing model.

UML, the Unified Modelling Language (UML Specification 1997), is used as the modelling language, and RationalRose as the object-oriented modelling and software development tool.

5 Technologies and Resources

In this section, we describe the technologies and resources that are used in the components of the MULINEX system.

5.1 Document Acquisition

The gathering of documents is performed by a modified version of the Harvest gatherer (Bowman et al. 1994). Harvest has been augmented to call routines for language identification, document classification, and automatic summarisation, and to work in conjunction with the Fulcrum SearchServer, which is used as the information retrieval core engine in our system.

5.2 Language Identification

Language identification is performed by making use of an algorithm which compares the relative frequencies of the most frequent n-grams (from 1 to 5 characters) in a document to 40 stored language profiles (Cavnar and Trenkle 1994).

5.3 Categorisation

Document classification is performed by the k-nearest-neighbour algorithm (Yang 1994), a statistical algorithm which classifies a new document by combining the category assignments of the k most similar training documents, weighted by the statistical distance (tf.idf) between the new document and each of the k best matching training document. The categoriser is trained with documents from newsgroups in French, German and English.

In addition, there is a keyword-based categorisation algorithm for narrow, specialised categories.

5.4 Summarisation

Summarisation is performed by selecting the sentences which best characterise a document. During document gathering, it operates in query-independent mode by selecting sentences on the basis of structural and layout HTML markup, and by position in the document or paragraph.

5.5 Query Formulation / Translation

The MULINEX system translates and expands the users' queries. Since the retrieval performance of automatically translated queries is inferior to monolingual information retrieval (Oard 1997), there is an (optional) step of user interaction, where the user can select terms from the translated query and add his own translation. Queries are morphologically analysed by making use of Morphix (Finkler and Neumann 1988) and MMORPH (Petitpierre and Russell 1995), and then translated by making use of multilingual dictionaries. We make use of the terminology database MultiTerm from TRADOS. The translated queries are the input to the search in the document collection.

5.6 Information Retrieval

The search is being performed by the Fulcrum SearchServer, a state-of-the-art information retrieval system, which incorporates linguistic technologies for morphological normalisation of documents and queries. The results are presented along with information about their language, thematic categories and automatically generated summaries. If the result pages are accessible to the system without large delays (e.g., if they reside on the same intranet), a summary which is tailored to the user's query can be produced. Results can be ordered by relevance or by thematic categories.

5.7 Database

Two SQL-based database management systems are used in the MULINEX system: Fulcrum Search Server for all information retrieval tasks and for storing category profiles, and a standard SQL database (MSQL) for storing user profiles and the multilingual lexicon.

5.8 Multilingual Lexical Resources

The MULINEX system uses six bilingual lexicon databases with 100.000 to 200.000 entries each for all six language pairs supported by the system (German-English, German-French, French-English and the converse pairs).

5.9 Machine Translation

Summaries and result documents can be translated on demand by making use of the LOGOS machine translation system.

6 Deployment and Validation

The system will be made publicly available by the user partnters in the consortium, who will obtain feedback from the end users of the system in order to evaluate the usability of the system.

6.1 Validation Sites

In May 1998, the system has been installed in the online services of Grolier Interactive Europe and Bertelsmann Telemedia, two large internet service and content providers in France and Germany. They will use it to provide multilingual search facilities for their sites, and to enhance the functionality of their existing search engines. These services will become publicly available in the 3rd quarter of 1998.

6.2 Validation Methodology

The end users of the system will be invited to provide feedback on the usability of the system via questionnaires, in which they evaluate the system, suggest improvements and can provide personal details. Users can also use a mailto-link to give feedback in free form. In addition, there will be in-depth interviews with a selected group of end-users.

7 The Road Ahead

In the next months until the end of the project, a number of enhancements will be made to the baisc functionality of the system described above.

7.1 Mulinex Agent

The agent system performs information search tasks periodically on behalf of the user. Registered users can specify search queries and define interest profiles. The agent system runs these queries periodically and informs the user of new information which matches his interests.

For selected domains, the agent will extract the important facts from documents for the user by making use of the Saarbrücken Message Extraction System SMES (Neumann et al. 1997). The user will be notified by e-mail, a personalised web page or a push channel.

7.2 Query Expansion

We plan to add a query expansion module based on the results of the EuroWordNet project (Vossen 1997).

7.3 Result Clustering and Visualisation

Presently, search results can be grouped into a set of pre-defined categories. In future versions, MULINEX will use clustering methods for automatically grouping search results.

The clusters will be presented to the users with a graphical user interface based on VRML.

7.4 Query Disambiguation

Presently, translations of queries are disambiguated by the user who selects among alternative translations. This presupposes knowledge of the target language by the user.

In the next version, the user will be presented with terms in his own language which correspond to the translation alternatives, so that he can perform disambiguation in his own language.

References

(Bowman et al. 1994) C. Mic Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber, Michael F. Schwartz, and Duane P. Wessels. *Harvest: A Scalable, Customizable Discovery and Access System*. Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado, Boulder, August 1994

(Capstick et al. 1998) Joanne Capstick, Gregor Erbach and Hans Uszkoreit. *Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents.* Working Notes of the AAAI Spring symposium ‘‘Intelligent Text Summarisation”. Stanford, CA, 1998.

(Cavnar and Trenkle 1994) William B. Cavnar and John M. Trenkle. *N-Gram-Based Text Categorization.* Symposium on Document Analysis and Information Retrieval, Las Vegas, 1994.

(Finkler and Neumann 1988) Wolfgang Finkler and Günter Neumann. *MORPHIX: A fast realization of a classification-based approach to morphology*. In: H. Trost (ed.): Proceedings der 4. Österreichischen Artificial--Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung, Springer, Berlin, 1988.

(Fowler and Scott 1997) Martin Fowler and Kendall Scott. *UML Distilled: Applying the Standard Object Modelling Language.* Addison-Wesley Longman, 1997.

(Grefenstette 1998) Gregory Grefenstette (ed). Cross-Language Information Retrieval. Kluwer, Boston, 1998.

(Hernandez 1997) Juan Antonio Hernandez 1997. *MULINEX User Requirements: Synthesis Report.* MULINEX deliverable report 2.3, Grolier Interactive Europe, Paris, 1997.

(Jacobson et al. 1992) I. Jacobson, M. Christerson, P. Jonsson and G. Övergaard. *Object-Oriented Software Engineering – A Use Case Driven Approach.* Addison-Wesley, Reading, MA; ACM Press, New York, 1992.

(Neumann et al. 1997) Günter Neumann, Rolf Backofen, Judith Baur, Markus Becker and Christian Braun. *An Information Extraction Core System for Real World German Text Processing.* 5th Conference on Applied Natural Language Processing, ANLP-97, Washington DC, 1997, pages 209 – 216.

(Oard 1997) Doug Oard. *Alternative Approaches for Cross-Language Text Retrieval.* AAAI Spring Symposium on Cross Language Text and Speech Retrieval, Stanford, CA, 1997.

(Petitpierre and Russell 1995) D. Petitpierre and G. Russell. *MMORPH – The Multext Morphology Program.* Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva, February 1995.

(UML Specification 1997) OMG (Object Management Group). *UML Specification.* http://www.omg.org/, November 1997.

(Vossen 1997) Piek Vossen. EuroWordNet: a multilingual database for information retrieval. Third DELOS workshop – Cross-Language Information Retrieval. European Research Consortium for Informatics and Mathematics, Zurich, 1997, pages 85 – 94.

(Yang 1994) Yiming Yang. *Expert Network: Effective and efficient learning from human decisions in text categorization and retrieval.* 17th ACM SIGIR Conference on Research and Development in Information Retrieval. pages 13 – 22.