Project Note

MULINEX

Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web

Gregor Erbach, Günter Neumann, Hans Uszkoreit
DFKI GmbH
Language Technology Lab
66123 Saarbrücken
Germany
lt-mulinex@dfki.de

Abstract

This paper gives an overview of the project MULINEX, which is a "leading-edge application project" funded in the Telematics Application Programme (Language Engineering Sector) of the European Union. The goal of the project is the development of a set of tools to allow cross-language text retrieval for the WWW, concept-based indexing, navigation tools and webiste management facilities for multilingual WWW sites. The project takes a user-centered approach in which the user needs drive the development activities and set the research agenda.

1 Overview and Objectives

MULINEX is a "leading-edge application project" which addresses the requirements of two kinds of users: web content providers and service operators who wish to provide multilingual information, and the customers of such multilingual information services (henceforth referred to as end users). The objective of the project is to provide multilingual search, retrieval and navigation functionalities for the WWW.

> Leading-edge application projects aim at advanced applications based on existing or emerging IC components and novel Language Engineering technologies. The goal is to meet user requirements dictated by socio-economic changes over the next few years. (from the call for project proposals for the Telematics Application Programme).

The *socio-economic changes* addressed by the MULINEX project are the emergence and widespread acceptance of the WWW, the increasing availability of gigabytes of information in different languages, and the increasing number of people with different mother tongues who need to find information on the web.

Providers of web search engines are already producing localised versions for different countries (e.g., lycos.de for Germany), but so far these provide only the user interface and the advertisements in the local language, but the search and retrieval process itself is not language-aware.

The technologies to be used in the project include a state-of-the-art information retrieval system, advanced linguistic processing tools (morphological analysis, information extraction, lexical semantics), algorithms for alignment of translated texts and terminology extraction, and machine translation systems.

The intended prototype application can run entirely on the server of a content provider or search service operator, so that the end user needs only a standard web browser such as Netscape Navigator, Alis Tango or Microsoft Explorer. The project is committed to supporting open web standards and will avoid dependence on proprietary formats and solutions, in order to make the results applicable to a wider user base.

The application will be realised as a group of interacting tools which improve access to information (search and navigation) in multilingual web document collections, and support the creation and maintenance of multilingual content for the web by information providers. The set of tools will provide the following search, retrieval and navigation functionality for the end user:

1. search by a combination of keywords, phrases, and concepts
2. retrieval of documents in different languages with one monolingual query through multilingual indexing
3. online generation and presentation of navigation maps or menus for supporting interactive refinement of query and search
4. exploitation of context and user profiling information for selecting relevant documents

In addition, it will offer functionalities for the management of multilingual websites. These will only be discussed in this paper insofar as they are relevant to cross-language text retrieval.

2 Approaches

2.1 Application Domains

In the project MULINEX, we consider retrieval and navigation tools for two different kinds of application domains related to the WWW: On the one hand, the project investigates how to provide and improve cross-language retrieval performance for unrestricted search engines such as HotBot, Lycos, or AltaVista that index the entire content of the WWW (or thematically unrestricted portions of it). On the other hand, we consider search services which provide for information retrieval services for single (thematically restricted) web sites. An open, thematically unrestricted, application domain imposes different requirements than a narrow, thematically restricted, one.

In an *open domain*, it is necessary to automatically identify the language of a document. It will be useful to perform an automatic thematic classification of documents in order to present additional information about retrieved documents to the user. Cross-language text retrieval for an open domain will have to rely on general-purpose translation dictionaries and language technologies.

For a *restricted domain,* on the other hand, it is possible to use domain-specific dictionaries, thesauri, and language technologies in order to improve retrieval performance. If translated documents are available in a restricted domain, these corpora can be exploited to learn domain-specific terminology that can be used for the purpose of cross-language retrieval.

Initially, wide and restricted domains will be handled separately and with partially different methods. In the course of the project, we will examine the possibilities for automatically identifying thematically restricted subsets of open domains, and treating these with domain-specific methods.

2.2 Cross-Language Text Retrieval

Cross-language text retrieval is among the core objectives of the project. The prototype will initially use French, English and German, but is designed to be extensible to other languages.

In the following, we consider various options for cross-language retrieval, and discuss how these are applicable to the handling of open and restricted domains.

2.2.1 Translation of Documents

Cross-language text retrieval is reduced to monolingual text retrieval if documents are translated into all potential query languages. A separate index can be built for all target languages produced by a machine translation system, and queried with monolingual queries. Such a *document translation approach* is planned for the document base on sustainable development in the project Twenty-One [Kraaij 1997].

Translation of documents may be a feasible approach for a restricted domain with a limited number of documents, but it will not be feasible for general-purpose WWW search engines due to the large number of documents. In addition, there are scalability problems with the addition of new languages, since each new language would require a re-translation of documents that are already indexed, and are probably not stored with their full text.

We will therefore investigate document translation only as a technique for restricted domains, and compare its performance to (and in combination with) other techniques.

2.2.2 Translation of Index Terms and Queries

Translation of index terms is a problematic approach because index terms without a phrasal or sentential context are hard to translate accurately. Better results can be expected from the use of machine translation of documents to derive index terms, even if the results of the machine translation are not used otherwise.

The translation of query terms entered by a user is also very problematic because these terms are often ambiguous, and a short query does not provide enough context to enable an accurate translation.

2.2.3 Relevance Feedback with Parallel Texts

Relevance feedback is a useful technique for improving recall and precision by using (parts of) a document which is considered relevant for expanding a query. Relevance feedback can be used for cross-language retrieval if a document which is considered relevant exists in several different languages. In this case, the words in the translations of a relevant document (or of passages thereof) can be used to construct a new query to find similar (untranslated) documents in other languages. This method is advantageous if there is a significant proportion of translated documents in the search space, so that a query in the user's language is likely to find a relevant document which has translations that can be used for relevance feedback.

Note that using translated documents in this way requires the CLTR system to know which documents are translations of each other. This requirement is addressed by a document management system (cf. section 2.6).

2.2.4 Machine Translation for Relevance Feedback

In the case where no translated versions of a relevant document are available, these can be constructed by means of machine translation. The output of the MT is then used to construct a new query in the target language. We expect this approach to be superior to the translation of queries because a long document will provide more context that helps the MT system arrive at a correct translation.

A problem with using MT for text retrieval is that recall will suffer because MT systems choose only one of several possible translations. If all the possible translations produced by an MT system were added to a query in the target language, the recall could be improved.

2.3 Concept-based retrieval

Cross-language retrieval performance, both recall and precision, suffer from the fact that there is no one-to-one correspondence between words in different languages. Recall suffers because the multilingual thesaurus or MT system used for query translation may choose a wrong translation that does not occur in the target language document. Precision suffers if a translation of a query term is chosen that corresponds to an unintended reading of the query term, and/or if the translation has additional unintended readings.

Since the undesirable consequences of ambiguity in monolingual retrieval are compounded in cross-language retrieval, performance gains can be expected from any system that performs indexing and retrieval according to the concepts expressed in the documents and the queries rather than the words.

Adequate cross-language retrieval is therefore concept-based retrieval.

The two approaches based on relevance feedback (2.2.5 and 2.2.6) are a step in this direction since expanding a query with the translation of an entire document that is judged relevant tends to smooth out the undesirable effects of wrong translations of single query terms.

In the longer run, it will make sense to direct research and development in several directions: on the one hand towards better disambiguation of words, toward index terms that go beyond single words, and towards indexing based on grammatical relations. All of these will be briefly discussed in the following

2.3.1 Disambiguation and Domain Modelling

Disambiguation is an important requirement for cross-language retrieval because it helps to avoid the negative consequences of the ambiguities of the source and target languages combined.

It is an important observation that words carry different meanings (and have different translations) depending on their syntactic context. For example, if something is ***in*** *a table*, one is normally talking about statistical material or word processing, and *table* should be translated into German as *Tabelle.* On the other hand, if something is ***on*** *a table*, one is normally talking about furniture, and *table* should be translated into German as *Tisch.* Likewise the word *key* will usually refer to different concepts in the phrases ***hit*** *a key* and ***turn*** *a key*. Such facts have been used for the acquisition of lexical semantic knowledge from corpora [Johnston et al. 1995].

The project will use and further develop corpus-based techniques for syntactic (part of speech) and semantic disambiguation of words depending on their syntactic context and the words with which they co-occur.

2.3.2 Grammatical Relations and Phrasal Indexing

We assume that grammatical relations such as *subject* or *object* play an important role in document retrieval. People retrieve documents not only to find out about a particular topic, but often because they want to find information about a particular class of events (e.g., events in which a bull kills the torero in a bullfight) or because they want to achieve a particular task (e.g., install an operating system).

With current retrieval systems based on keywords or statistical similarity, a query such as "bull kills torero" would also find documents in which the torero kills the bull, and the query "installing an operating system" would also locate documents in which the operating system installs programs, files or drivers.

The following examples show that information about installing an operating system can be expressed in a number of different syntactic forms (compounds, complex noun phrases, finite and infinitive verb phrases, gerunds etc.):

How to install the operating system
Installation of the operating system
Operating system installation
Procedures for installing the operating system
The operating system is installed by ...

The techniques for discovering such relationships in a text have been developed in the area of information extraction (also called "message understanding"), where the task is to extract predefined pieces of information from a text. The performance of information extraction systems is evaluated regularly in the MUC (Message Understanding Conference) competitions, in which the systems have to find information about terrorist attacks or joint ventures from newspaper articles.

The techniques (shallow parsing and template filling, see section 3.7) developed for information extraction can also form a basis for information retrieval based on grammatical relations.

The project will develop special data structures for providing efficient storage of and access to indices based on grammatical relations.

2.4 Extraction of terminology from multilingual corpora

The availability of parallel corpora of translated documents for thematically restricted domains enables the extraction of multilingual terminology . The correspondences between these corpora can be determined at the sentence, phrase and word level by automatic alignment algorithms, which are to be provided and further developed in the project by TRADOS (see section 3.5).

2.5 Navigation tools

2.5.1 Interactive Search

It is clear that search for information is not a one-step process in which the user gives one query and is presented with a list of solutions. Rather it involves an iterative refinement of the query until the desired pieces of information are found. This is in contrast with applications such as information filtering (e.g., personalised newspaper) or information routing, which are performed without human interaction.

The project strives to provide methods and tools for helping this kind of navigation process in an information space. Among the methods provided are established methods such as relevance feedback, thesaurus-based query expansion, but also new approaches such as partitioning the space of found documents according to criteria such as language, thematic classification, physical location etc., and letting the user choose among these subclasses.

Further opportunities for interaction with the user are in the area of the selection of word senses of ambiguous query terms, and interactive thesaurus-based query expansion and translation.

2.5.2 Filtering Options

Current WWW search engines already give the user the option to filter out unwanted documents as part of the query. In existing systems, users can limit the search space

- by protocol (http, ftp, nntp, ...),
- by location (top-level domain),
- by document type (text, images, video, sound),
- by date of creation of last modification,
- by popularity (number of accesses), or
- by rating/recommendation.

Filtering according to language is not yet implemented in existing search engines, but can easily be done by using algorithms for language identification (see section 3.2) at indexing time to detect the document language, and using the language negotiation features of the HTTP 1.1 protocol to retrieve only documents in the language(s) preferred by the user.

Future system will have to go beyond these more or less superficial criteria to offer more options for iteratively constraining the search to find the desired documents. It appears unreasonable to expect the user to fix thematic categories in advance of the search since there is a vast range of such categories which would be hard to learn.. We intend to perform a keyword-based search first, and then group the found documents into thematic categories for selection by the user.

2.6 Multilingual Document and Website Management

The issue of document management is relevant for cross-language retrieval for two reasons:

1. If several translations of one document are retrieved by a cross-language query, they should only be shown as one "hit".
2. In order to derive multilingual terminology from translated documents, it is necessary to store alignment information for translated documents.

In addition to these requirements motivated by cross-language retrieval, there should also be tools to support the consistency of the information across different languages, and perhaps the integration of the document management systems with translators' workbenches to support the creation of multilingual web sites.

3 Technologies

This section discusses the technologies and algorithms chosen or considered for the project.

3.1 Information Retrieval Engine

Fulcrum SearchServer is a second generation information retrieval product, based on inverted files to perform fast searches. It includes features such as the intuitive and Fuzzy Boolean search strategies. This software is used with a variety of document formats and operates in heterogeneous computing environments that include multiple operating systems, networks and graphical user interfaces. It conforms with open system standards and is well suited for use in client/server computing.

Fulcrum's software has been adopted by the Commission of the European Community and has been selected as standard information retrieval product by the European Space Agency.

At the core of Fulcrum's product family is Fulcrum SearchServer, a multi-platform indexing and retrieval server engine, which makes use of an SQL-based query language and complies with Open Database Connectivity (ODBC).

Fulcrum SurfBoard combines the SearchServer indexing engine with Internet access protocols to allow information providers to search-enable their Internet sites. World Wide Web browsers and other common Internet clients can be used to search and navigate effectively through corporate publications. Automatic conversion to HTML means that information providers do not have to invest significant resources in converting extensive document collections.

SearchServer and SurfBoard consitute the basic full-text retrieval system to which multilingual and concept-based search facilities will be added in the MULINEX project.

3.2 Language Identification

Identifying the language used of documents is of crucial importance for document retrieval for the web. On the web, one cannot always expect that the authors or site creators make use of the existing standards for specifying the language of a document. Therefore it is necessary to perform automatic language identification. We will use a technique based on trigrammes (sequences of three consecutive letters), which has been shown to be superior to the alternative method based on frequent words [Grefenstette 1995].

Two benefits will be gained from automatic language identification:

1. Once the language has been identified, it is possible to use the appropriate linguistic processing components for that language, for example to avoid classifying the German noun *Wetter* as the comparative form of the adjective *wet*.
2. It becomes possible to inform the user in which language the retrieved documents are written, and to filter out undesired languages according to the user's preferences.

3.3 Machine Translation

No machine translation systems will be developed in the project. Commercial MT systems will be evaluated and selected. It is an imporant requirement that MT systems can be customised in order to improve performance for restricted domains.

3.4 Morphological Analysis and Part-of-Speech Tagging

Morphological analysis is a crucial component for normalisation of terms in richly inflected languages such as German, Finnish or Georgian. For languages such as German or Swedish in which compounds appear as one orthographic word, the analysis of compounds is an important requirement, especially since these compounds are often translated by noun phrases in other languages (e.g., German *Waschmaschine*, English *washing machine*, and French *machine a laver*).

For German, we will use the morphological analyser MONA, developed by DFKI, which has a broad coverage (more than 120.000 stem entries) and an excellent speed of 2800 words/sec on a SUN SparcStation 20. For other languages, existing commercial morphological analysers will be used.

Part-of-Speech tagging (disambiguation) is an important step in the identification of phrases. We will use an unsupervised tagger described in [Brill 1995].

3.5 Alignment

For alignment, the program TAlign from TRADOS will be used. TAlign is a program for synchronizing (aligning) two texts that are translations of each other. TAlign creates a translation memory from corresponding source and target language texts.

TAlign combines statistical and heuristic methods to achieve optimum results. Numerous parameters adjust TAlign for specific input texts, allowing creation of as many reliably-aligned sentence pairs as possible.

Depending on the quality of the texts, TAlign can handle switched or missing paragraphs.

Even if a sentence was omitted or translated by multiple sentences, TAlign can in most cases make the correct alignment decision. To support this decision process, the user has the option of specifying bilingual word lists and so-called "priority lists."

Multilingual terminology that can be used for cross-language retrieval will be extracted from the aligned texts.

3.6 Document Classification

For the thematic classification of documents, we will use statistical (vector space) models, obtained from sample corpora which contain representative texts for given thematic categories. For example, the documents found in different YAHOO! categories could be used to construct the vector space models for these categories, to which new documents can be compared.

3.7 Shallow Parsing and Information Extraction

For phrasal and relational indexing, it is necessary to get a structural analysis of sentences that reveals the phrase boundaries and the grammatical relations. For this purpose, we will make use of the Saarbrücken Message Extraction System (SMES) [Neumann et al. 1997], which provides a set of basic powerful, robust and efficient natural language components: the morphological component MORPHIX, a declarative tool for expressing finite-state grammars, an efficient and robust bidirectional lexically-driven parser, and an interface to a typed feature-based language and inference.

3.8 Multilingual Document and Website Management

No independent development of document management and website management tools is foreseen in the project. Instead, the project will build on existing systems for document and website management and add functionality for the storage of alignment information.

Hyperwave [Maurer 1996] (formerly Hyper-G), a second-generation hypermedia system, from IICM at the Technical University of Graz makes use of multilingual *document clusters,* in which versions of a document in different languages are treated as one unit, and one version is delivered to the user according to his/her language preference. It is not yet clear whether Hyperwave will be used in the project, because its mechanism for language preferences has been superseded by the content negotiation of HTTP 1.1, and one of its most attractive features, a separate database for hyperlinks, is not (yet) available for HTML documents.

The multilingual HTML extension MLHTML, developed in the EU project RELATOR, which contains the different language versions of a document in one source file has been rejected as too inflexible.

4 User-Driven Approach

4.1 Who are the Users?

The users are publishers, web content creators and information providers who want to create and manage multilingual document collections, and/or provide multilingual search capabilities for their site. They are also providers of Internet access and search services, who want to add value through multilingual search and navigation facilities, as well as their customers or end users. Another group is companies and institutions whose business requires searching in large multilingual document collections.

Beyond the users represented in the project consortium (see section 6), a wider project *user group* has been established that contains a number of public institutions and commercial companies interested in the technologies and results of the project. The user group will review intermediate deliverables and prototypes, and ensure that the project produces results will be re-usable beyond the immediate requirements of the project's partners.

The feedback from the users has already led to a modification of the project's workplan: Initially, it was planned to make heavy use of a translator's workbench, and use the terminology databases and translation memories that are created during the (human) translation process for multilingual indexing. However, the users have indicated that human translation of documents does not play a major role in their development process of multilingual websites.

The project will thoroughly assess the user needs and requirements in order to develop technologies and system that satisfy real user needs. The techniques used for the assessment of user needs are described in the following:

4.2 User Interviews

User interviews via questionnaires will be performed among the users of *Club Internet,* a leading French Internet access and content provider for the general public. Among the questions asked will be the native language, proficiency in foreign languages, interest in foreign language documents, usage of foreign language sites, demand for cross-language retrieval and machine translation services, among other demographic data.

4.3 User Monitoring

In order to get a better understanding of the users' search and navigation behaviour, the project will examine the log files of existing search engines. Resources permitting, the project will also perform think-aloud experiments, in which users explain their cross-language retrieval strategies for a given task.

4.4 User Evaluation of Prototypes

In order to evaulate the usefulness and user-friendliness of the prototype systems developed in the project, these will be presented to end-users for evaluation and feedback at an early stage of the development.

5 Evaluation

The prototype will be tested on *Club Internet* which is regarded as the most prominent French access site for the general public. This will provide an open multilingual domain. It will also be tested on a thematically restricted domain, which contains a substantial proportion of translated documents. This will be either a marketing service for professionals, traders and marketing experts, or a specialist service catering for health or construction issues.

The MULINEX prototype will be compared in performance to the best existing systems such as Fulcrum SearchServer, AltaVista and Lycos. Test suites and methodology for quantitative evaluation will be constructed as part of the quality assurance efforts of the project.

6 The Consortium

The partners in the project are:

DFKI GmbH, Saarbrücken, Germany
TRADOS Germany GmbH, Stuttgart, Germany,
Bertelsmann Telemedia GmbH, Gütersloh, Germany
Grolier Interactive Europe, Paris, France
DATAMAT - Ingegneria dei Sistemi S.p.A., Rome, Italy

The consortium combines users (Telemedia, Grolier) and technology providers (Datamat, DFKI, Trados). The five partners bring in expertise from various areas:

The Language Technology Lab of DFKI (German Research Center for Artificial Intelligence) brings in experience in Natural-Language Processing, especially morphological analysis, syntactic analysis and message extraction.

DATAMAT develops state-of-the-art information retrieval technology through its subsidiary Fulcrum. The Fulcrum SearchServer is described in section 3.1.

TRADOS develops software products in the field of translation tools. TRADOS specialises in terminology database systems and translation memory systems, and tools for alignment of text and terminology extraction.

The role of users in the project is played by Grolier Interactive Europe and Bertelsmann Telemedia. Grolier is dedicated to the development of multimedia and interactive communications online, especially focussed on the World Wide Web. Telemedia is the Internet solutions company of Bertelsmann AG, and offers professional service in all areas connected with the WWW, especially electronic commerce.

Acknowledgments

The work presented here was financially supported by the Commission of the European Communities (Language Engineering Sector of the Telematics Application Programme, LE-4203). The ideas presented here are the result of intensive discussions about the project goals and approaches with Gianfranco Abbrescia and Giovanni Gadaleta (DATAMAT), Juan Antonio Hernandez

(Grolier), Joanne Capstick, Abdel Kader Diagne (DFKI), Iain Urquhart and Giovanni B. Varile (CEC).

References

[Brill 1995] Eric Brill. Unsupervised learning of disambiguation rules for part-of-speech tagging. *Proceedings of the Workshop on Very Large Corpora,* 1995.

[Grefenstette 1995] Gregory Grefenstette. Comparing two language identification schemes. *Proceedings of the 3r*d *International Conference on the Statistical Analysis of Textual Data (JADT'95)*, Rome, Italy, Dec. 1995

[Johnston et al. 1995] Michael Johnston, Branimir Boguraev, James Pustejovsky. The acquisition and interpretation of complex nominals *Working Notes of the AAAI spring symposium on the representation and acquisition of lexical knowledge* AAAI, 1995

[Kraaij 1997] Wessel Kraaij. Multilingual functionality in the Twenty-One project. *this volume*, 1997.

[Maurer 1997] Hermann Maurer (ed.). *Hyper-G, now Hyperwave.* Addison-Wesley, 1997.

[Neumann et al. 1997] Günter Neumann, Rolf Backofen, Judith Baur, Markus Becker, Christian Braun. An Information Extraction Core System for Real-World German Text Processing. *Proceedings of ANLP,* 1997.