

EMIS

A MULTILINGUAL INFORMATION SYSTEM

Bärbel Ripplinger
IAI
Martin-Luther-Straße 14
D-66111 Saarbrücken
babs@iai.uni-sb.de

Mai 1998

Abstract

The objective of the EMIS project is the conception and realization of a multilingual information system on European media law with the following functionalities:

- search by words, a combination of words, phrases or keywords,
- guided search by using a so-called *thematic structure*,
- retrieval of documents in different languages with one monolingual query by using language processing and MT technology,
- exploitation of additional information for the retrieved documents, which is stored in a database.
- structured representations of the document archive, the so-called *dogmatic structure*,
- multilingual user interface.

The core component of EMIS is the multilingual cross language retrieval which is enhanced by linguistic processing i.e. by a morpho-syntactic analysis applied to the documents and to the queries and the relevant bilingual dictionaries. The languages covered by the project are German, English and French.

System Description/Demo

Introduction

The project EMIS aims at the conception and realization of a multilingual information system in the domain of European media law. The system will provide multilingual search, retrieval and navigation functionalities for an archive of documents. In the current application domain these documents consist of European media law texts.

EMIS is a joint project of the Institute for European Media Law (EMR) and the Institute for Applied Information Sciences (IAI) funded by the German Ministry of Economics. The project started in November 1996, the development of the information system in May 1997. The prototype presented here and the final system runs entirely on a webserver, so the end users need only a standard web browser. No proprietary software or formats are used. The final system will also be offered as Intranet application on a CD-ROM which will be updated at regular intervals.

The Document Base

At present the database consists of 166 media law texts from different European countries and the European Union in the languages German, English and/or French. The texts from the EU exist in all three languages and are used to build up bilingual thesauri on the one hand and for relevance feedback to improve recall and precision on the other. The documents are stored in a relational database (ORACLE), together with additional information about the law. There will also be information about which of the working languages the document is available in. This information will be used during generation of the output, i.e. the user will receive the document in the interface language selected.

Each document is split into its sections and/or paragraphs which form the search space of the retrieval.

Cross Language Retrieval in EMIS

In the EMIS project a cross language retrieval is done by query translation. This query translation is for keywords done by using multilingual thesauri and for free text retrieval by a simple MT tool which is part of the Mpro program package: a development of the IAI [1].

To improve the query translation from the interface language into the other working languages, the translation is done domain dependently. If there are translations in the given context - in the EMIS system the domains are *media law*, *telecommunications*, and *general law* -, these are preferred. If no such translations are available all others found in the bilingual dictionary are used for retrieval. Multi-word units or phrases will be translated compositionally if there are no translations of the whole expression; they will therefore specially marked to inform the user about their 'unauthorized' status.

The Retrieval Functionalities of EMIS

EMIS provides the user with several possibilities to access the documents. The first three represent a kind of retrieval by concepts and the last one a cross language free text retrieval.

Retrieval using the 'dogmatic structure' The dogmatic structure gives an overview of all existing media laws in Europe and the European Union. For each country the relevant docu-

ments are listed sorted into particular law areas such as constitutional, broadcasting, telecommunication, multimedia, copyright, data protection, press, film and competition law. Clicking on a document the user can view the whole text at once or read paragraph by paragraph. The structure is available in all three working languages.

Retrieval using the 'thematic structure' The EMR as the domain expert in the project has developed the so-called *thematic structure*, this hierarchical structure represents the domain by different topics. The upper structure consists, for instance, of the following concepts: *Broadcasting, Multimedia, Youth Protection, Media Concentration, Financing and Telecommunications*. These concepts are refined in the lower levels. By clicking on a concept the user will get documents relevant to these special topic. Compared to the first retrieval functionality the user will get here, for instance, all the documents from all countries relevant to telecommunications (and not only one).

The base thematic structure (in all working languages) and the classification of the documents along the different themes is done manually because expert knowledge is absolutely essential. The thematic structure itself will be automatically generated each time new documents are entered into the database.

Keyword retrieval For the text retrieval by keywords traditional multilingual thesauri are used for the document indexing. In the current state of the system the keywords are assigned by the domain experts only in German. The multilingual thesauri are used to translate the particular query and retrieve the relevant documents. For the future it is foreseen that the linguistic analysis described below will be used together with the thesauri to extract the keywords automatically. First attempts to do automatic indexing have been completed successfully within another project of IAI.

Free text retrieval The retrieval approach used in this project takes advantage of linguistic processing technologies. Using the **Mpro** tool every document is morpho-syntactically analyzed. The tool performs a part-of-speech tagging, a lemmatization, an analysis of homographs (optional) and for German texts also a compound analysis.

Each word is assigned with information about its morphology and grammatical attributes. In the following examples for a German (1), an English (2) and a French (3) word analysis are given. Only the features used in the retrieval system are shown:

- (1) {ori=Rundfunkanstalt,wnra=1,wnrr=1,snr=1,c=noun,lu=rundfunkanstalt,s=loc&ag,t=rundfunk#anstalt,cs=n#n,ts=rundfunk#anstalt,ds=rundfunk#anstalt,ls=rundfunk#anstalt,ss=medium#loc&ag,w=2,lngs=germ#germ}
- (2) {ori=Broadcasting,wnra=1,wnrr=1,snr=1,c=noun,s=ation,lu=broadcasting,ds=broadcast~ing}
- (3) {ori=Radiodiffusion,wnra=1,wnrr=1,snr=1,c=noun,s=process,lu=radiodiffusion,ds=radiodiffuser~ion,nb=sg,g=f}

Indexing

The results of the document analyses are used to generate different indexes. For German texts, three indexes are used: the first is an index generated by using the 'lu-feature' (i.e. the lexical unit) – *lu-index* –, the second is generated from the 'ls-feature' which indicates the

derivation of the lexical unit – *ls/ds-index* – and the third is built up using the 'ts-feature' – *ts-index* – which marks the possible word parts of a compound (in cases of simple words, the value of the ts-feature is the same as the value of the lu-feature).

For English and French two indexes are generated at a time, the first takes the lu-feature and the second the ds-feature (the derivation).

Function words are excluded from the indexing process.

The Algorithm

Each query which can be a simple word, a multi-word unit or a phrase undergoes the same morpho-syntactic analysis as the documents. From the output of this analysis the values of the lu-, ts-, and the ls-features or the lu- and the ds-features are used to search the indexes. The result of the access with the lexical unit (lu) of the lu-index represents the amount of exact matches. The access of the ls-index with the value of the ls-feature results in a list of documents which contain words with the same derivation as that of the query. Here, some wrong results can occur if the derivation denotes a homonym. For example, the derivation of *publicity* is *public*, then all occurrences of 'public' are listed as result. At present we are investigating whether the integration of semantic information (s, ss-features) into the search process can solve this problem. As the result of the access of the ts-index we will get a list of all compounds containing the input word as part.

If the query consists of a compound, for all parts of the compound extracted from the ts-feature and the corresponding derivations (extracted from the ls-feature) an access to the ls/ds-index and the ts-index is performed at a time. The first results in a list of documents where the compound occurs with its parts. For instance, for the query *Jugendschutz* (*youth protection*) the list contains hits like ...*Schutz der Jugend* .../ ...*protecting of the youth* ... or ...*die Jugend schützen* .../ ...*protect the youth* The search space in these cases are one single sentence (i.e. the unit Mpro marks as sentence). By accessing the ts-index for all parts, the result is a list of documents containing semantically similar terms. This case is only relevant for German compounds. For the example of *Jugendschutz*, hits are *Kinderschutzbund*, *Jugendarbeit*.

The different result lists represent equally the different relevance of the hits which will also be clearly represented in the output.

If the input is a multi-word unit or a phrase all function words are removed and the remaining meaning-bearing words undergo the procedure described above. Additionally the units must be within a particular environment within one sentence to be an exact hit. This environment is computed using a rule of thumb, i.e. the length of the environment corresponds to three times the number of units. All other hits are collected in a second list which represents the minor relevance of those documents.

References

- [1] D. Maas, 1996: MPRO – Ein System zur Analyse und Synthese deutscher Wörter, in: **Roland Hauser** (ed): **Linguistische Verifikation**, Max Niemeyer Verlag, Tübingen 1996