

MproIR – A Cross-language Information Retrieval Component Enhanced by Linguistic Knowledge

Bärbel Ripplinger

IAI

Martin-Luther-Str. 14

66111 Saarbrücken, Germany

babs@iai.uni-sb.de

Abstract

With the globalization of the world markets, the need for multilingual information processing increases, because the users are increasingly forced to deal with information available in multiple languages. Due to the fact that translation services are quite expensive, the multilingual society needs smart technologies to access foreign language documents at least to a degree which allows the user to make use of the information provided by them. These technologies should take into account that a user has often only some comprehension ability for a given foreign language. Recent developments in Information Retrieval, therefore, try to support users by providing so-called cross-language retrieval techniques. These systems allow the user to query information about a particular subject in their own language and retrieving relevant documents written in foreign languages.

In this paper, the focus is on a cross-language information retrieval component integrated into the multilingual information system EMIS which is dedicated to the domain of *European Media Law*. This component, MproIR, makes use of linguistic knowledge provided by a morpho-syntactic analysis to increase its effectiveness. Based on stem information as well as compositional and derivational knowledge, automatic query expansion and translation as well as a classification of the retrieved documents is carried out. The languages currently covered by the system are German, English and French.

1. Introduction

Text retrieval is defined as a process by which the user seeks for documents which contain information about a certain subject. In monolingual systems, the user has to formulate his query in the language of the document repository in which he is looking for information, ideally in his/her own language. With growing globalization, the need for a facility which allows the user to access multilingual information increases. Such a tool should take into account that the user's knowledge of a foreign language is often not sufficient enough to formulate a query. As the full translation of documents is quite expensive even if machine translation (MT) systems are used, and because these systems cover only certain languages and are of lower quality (cf. Kay, 1995) *cross-language retrieval* (CLIR) are under development to overcome the language barrier. These CLIR systems start out realising that the user is not able to formulate a query in a foreign language but has enough knowledge of this language to roughly assess the contents of the retrieved documents. Therefore the query can be formulated in the user's own language and documents written in foreign languages are retrieved.

One ambiguous application area for multilingual information systems is the legal system of the European Community. Companies doing business in this economic area have to deal with the national law, the European Law (and its current transposition into national law) but also with the national law of the other European states. Although, the directives of the European Community are available in all languages of the member states, the national legislations, especially those of European countries not member of the EU, exist in most cases only in their

national language. Compared to other application areas, for information systems in the legal domain the provision of exact and complete information is a must (Krüger, 1997). This is a more challenging task if the information has to be extracted from foreign language documents.

The project EMIS, currently carried out at the IAI, targets at the specification and implementation of a multilingual information system on *European Media Law*. The languages covered are German, English and French. The multilinguality in this system is introduced, on the one hand, by a multilingual interface and, on the other hand, by a multilingual information and document repository. The system provides the user with fast and reliable information about this domain, through pre-defined access methods:

- a so-called *Systematic Structure* lists all relevant laws classified by countries and areas,
- a so-called *Thematic Structure* which is comparable with a classical thesaurus, describes the domain by different further structured topics, and
- a keyword search.

Legal work, however, often demands for seeking as much as possible information about a special subject to get a kind of global overview. To provide the users with a facility which allows them to carry out an *unstructured* search, i.e. retrieving as much information as possible about a special topic in the underlying document archive, a CLIR component is integrated into EMIS. This component, MPROIR, allows searches for single words, phrases and words combined by Boolean operator. In the following MPROIR will be further presented and the methodologies which are based on linguistic processing techniques will be discussed. A full description of the multilingual information system EMIS can be found in (Ripplinger, 1998; Ripplinger, 1999b).

1. Introduction

Text retrieval is defined as a process by which the user seeks for documents which contain information about a certain subject. In monolingual systems, the user has to formulate his query in the language of the document repository in which he is looking for information, ideally in his/her own language. With growing globalization, the need for a facility which allows the user to access multilingual information increases. Such a tool should take into account that the user's knowledge of a foreign language is often not sufficient enough to formulate a query. As the full translation of documents is quite expensive even if machine translation (MT) systems are used, and because these systems cover only certain languages and are of lower quality (Kay, 1995) *cross-language retrieval* (CLIR) are under development to overcome the language barrier. These CLIR systems start out realising that the user is not able to formulate a query in a foreign language but has enough knowledge of this language to roughly assess the contents of the retrieved documents. Therefore the query can be formulated in the user's own language and documents written in foreign languages are retrieved.

One ambiguous application area for multilingual information systems is the legal system of the European Community. Companies doing business in this economic area have to deal with the national law, the European Law (and its current transposition into national law) but also with the national law of the other European states. Although, the directives of the European Community are available in all languages of the member states, the national legislations, especially those of European countries not member of the EU, exist in most cases only in their national language. Compared to other application areas, for information systems in the legal domain the provision of exact and complete information is a must (Krüger, 1995). This is a more challenging task if the information has to be extracted from foreign language documents.

2. Related Work

Currently, there are different approaches to CLIR which can be classified as follows: systems which translate the query, systems which operate on translated documents and those systems which try to get advantage from both approaches. The last ones can be further categorised by the type of resources they use, i.e. machine translation systems (Gachot et al, 1998), transfer dictionaries (Gey et al, 1999) or corpus-based resources deployed, for instance, by using latent semantic indexing (Landauer & Littman, 1990) or similarity thesauri (Sheridan et al, 1997).

The use of linguistic knowledge provided by techniques from natural language processing is proven to be most successful to increase the effectiveness of a retrieval system (Liddy, 1998). If only stemming information for indexing and query processing is used, the resulting recall and precision figures are most promising. Also the use of derivational information can contribute to a better recall but often reduces the precision (Kraaij & Pohlmann, 1996) at the same time. Nevertheless, as shown in (Gaussier, 1998) derivational information can be successfully used for query expansion and translation. These techniques are not only deployed in research system, today search engines such as the German Alta Vista¹ and Infoseek experiment to improve their search process by adding morpho-syntactic information. Infoseek, for instance, linked their search engine with LinguistX of Xerox Corp., (cf. Linguistic Inside) which consists of a morphology analyser, part-of-speech disambiguator and a noun-phrase identifier where mostly stem information is used:

When you search for a plural word or a word ending in "ed," Search by Infoseek automatically searches for the root word. (This is known as stemming.) For example, if you search for recycled, Infoseek automatically searches for it recycle too.
Infoseek Help Documentation.

This process is only successful, if the index over which the search is performed undergoes the same linguistic processing. This means the main drawback of the search engines, their documents are primarily manually indexed.

To improve the recall, one proven method is *query expansion*. This is often done by making use of a thesaurus or an ontology to extract synonyms, generic terms and subconcepts. Such knowledge sources do not exist for all domains, and ontologies for general language such as EuroWordNet are often not specific enough to be deployed in a certain domain. Taking into account that the development of such a resource is very expensive, the CLIR method, described in this paper, performs a query expansion based on information gained from a morpho-syntactic analysis. To get most advantage of the linguistic information, this analysis is applied to the documents as well as to the query. The outcome is then used, on the one hand for the indexing, and on the other hand, for the query expansion and translation. Due to the exploitation of this knowledge within the retrieval algorithm, an inherent classification of the retrieved documents is provided.

The next section describes the linguistic processing techniques and the extracted information which is used to enhance the indexing described in section 4 and the retrieval approach presented in section 5. The final section discusses the CLIR method underlying MPROIR and describes the work envisaged for the next development stage.

3. Linguistic Information

Full text retrieval as realised in the MPROIR system is enriched with techniques for linguistic

¹As the German company ELEXIR reported during System'99, AltaVista.de uses their retrieval system. Unfortunately no further information how the search is improved is publicly available.

processing, i.e. a morpho-syntactic analysis. The most prominent type of such knowledge is stem information. As several systems have shown, this *normalisation* of the input, as well as of the indexed terms achieves more precise results. As a side-effect, the user can input entire words (also inflected ones) and does not have to use wildcards, as for example *transmi**, as a query term to find all occurrences related to *transmission*. However, using only information about stemming, hits such as *retransmission* are in general not found.

In order to improve precision and recall in MPROIR, we differentiate between retrieving exact hits and other relevant hits which will then be further categorised. To find the exact hits, the algorithm uses mostly stem information, but for German, due to its special compound building process, compositional information is also considered. This allows to find, for instance, defective nouns (4) or verbs with a separated prefix (5). To find the other types of relevant documents, derivational information and compositional information is exploited.

The necessary linguistic information is determined by means of a morpho-syntactic analysis carried out by **Mpro** (Maas, 1996), a tool developed at the IAI and available for different languages. The analysis is based on morpheme lexicons (which cover, for instance, 99% of the German language). During the morpho-syntactic analysis, Mpro performs a part-of-speech tagging, a lemmatisation, and an analysis of homographs (which is optional, i.e. only useful for documents and phrases). For German input, a compound analysis is done. As output, each word is assigned with information about its morphology, grammatical and semantic attributes. The following examples show the analysis outputs for the English multiword *youth protection* (1) and its German correspondent *Jugendschutz*, (2):

- (1) {ori=youth,wnra=1,wnrr=1,snr=1,c=noun,s=coll;time,lu=youth,ds=youth,ls=youth,ehead={nb=sg,case=nom;acc},}
 {ori=protection,wnra=2,wnrr=2,snr=1,c=noun,s=ation,lu=protection,ds=protect~ion,ls=protect, ehead={nb=sg,case=nom;acc}}

- (2) {ori=Jugendschutz,wnra=1,wnrr=1,snr=1,c=noun,lu=jugendschutz, s=massnahme,t=jugend#schutz,cs=n#n,ts=jugend#schutz, ds=jugend#schuetzen~IRREG,ls=jugend#schuetzen,ss=coll;time#massnahme,lngs=germ#germ, ehead={case=nom;dat;acc,nb=sg,g=m}}

Mpro usually categorises unknown words as nouns but also proper nouns and abbreviations are recognised.

Based on information such as that given above, the current system exploits for the different tasks (indexing, query processing and retrieval) only the following features types:

- **ori** corresponds to the input term (original wording),
- **lu** indicates the lexical basic form (stem), and
- **ls** the morphological derivation
- **t** marks the parts of a compound.

Currently the t-feature is only used in the processing of German documents and queries.

4. Indexing

On the basis of the analysis of the documents, different indices are constructed, depending on the language and the feature which serves as key. Function words are excluded from indexing. Together with the key (i.e. the value of the lu-, ls- or t-feature) the document identifier, the sentence *snr* and the word number *wnrr* as well as the original wording (*ori*) in the document are stored. The sentence and word number are used for input which consists of multiword units to decide whether a multiword unit can be seen as compound or whether all parts of a phrase

are within the same sentence (only then a hit is scored). The *ori*-value is used to highlight the hit in the corresponding retrieved document by displaying the text.

For English and French, two different indices are constructed using the features *lu* and *ls*. For German, an additional index is created, based on the values of the *t*-feature. The construction of the German *lu*-index consists additionally of a special treatment to identify German verbs with a separable prefix.

5. The Cross-Language Information Retrieval Component MPROIR

The retrieval starts with a linguistic analysis of the query and the determination of the set of search patterns, i.e. the query expansion, followed by the query translation. The translations found are then analysed and expanded in the same way as the queried term. For each language, the corresponding search patterns are then used for the retrieval which consists inherently of a classification method based on the same linguistic information.

5.1 Query Expansion

Many retrieval systems try to expand the query by adding synonyms or hypernyms to the set of search patterns to improve the recall. These additional terms are mostly obtained from thesauri or ontologies which are developed for a particular application or publicly available. However, there are not always thesauri or ontologies available for all subject fields, for instance for the domain of the EMIS system. Ontologies for general language are often not precise enough, or a certain overhead has to be accepted, if the system is dedicated to a special domain. Because developing a thesaurus or an ontology is very time consuming, the MPROIR system exploits instead compositional and derivational information provided by the *t*- and the *ls*-feature by the morpho-syntactic analysis.

To expand the query, from the outcome of the linguistic analysis, function words are eliminated and the values of the features *lu*, *ls* and *t* are used to determine a set of search patterns. For the example, the set of search patterns for *youth protection* (1) is: *youth (lu/ls)*, *protection (lu)*, *protect (ls)*.

For German compounds, the values of the features *t* and *ls* mark their several parts. In order to find syntactic variants, these parts are added to the set of search patterns. For the example, the set of search patterns generated for *Jugendschutz* (2) consists of: *jugendschutz (lu)*, *jugendschutz (t)*, *jugendschuetzen (ls)*, *jugend schutz schuetzen* (i.e. parts extracted from *t* and *ls*).

5.2 Query Translation

The input to the translation component is the complete morphological analysis of the queried term whereas for multiword units, shallow parsing is carried after the morphological processing to disambiguate the syntactic structure. MPROIR uses a shallow MT tool which performs a lexical transfer using huge bilingual transfer lexicons (for instance, the German-English lexicon contains about 488.000 entries, the German-French about 62.000 entries, and the English-French about 39.000) to translate the query. These lexicons contain not only single words and compounds but also phrases.

The translation itself is performed domain specifically due to the fact that the EMIS system is dedicated to a special domain. The necessary information is also encoded in the bilingual dictionaries.

In the current version, if no specialised translation is found, all translations available are used for the search. This method has no serious impact on the precision, because the EMIS

system deals within a special domain, and therefore only a certain subset of the possible translations occurs at all. Otherwise, the precision will rapidly decrease depending on the number of meanings and hence translations found of the queried term.

For multiword units, the MT component first looks up whether the dictionary contains a translation for the whole phrase. If no translation exists, the phrase is translated compositionally. This means that, if there exist domain specific translations for certain parts of the phrase, these will be used. For instance, the German translation for *data protection working party* is *Arbeitsgemeinschaft für den Datenschutz*, which is not in the transfer lexicon. However, the MT system identifies the partial multiword units *data protection* and *working party* and finds translations for them *Datenschutz*, and *Arbeitsgemeinschaft*. The missing function words can be neglected, because they would be deleted anyway by the following query expansion of the translation. If the translation for the phrase has to be done word by word, domain-specific translations are preferred and shallow parsing is carried out afterwards to get only one possible translation. The same methodology is used to translate German compounds, if there is no entry in the transfer lexicon. For each part of the compound all possible translations are determined and from the parsing out the first translation found, is used in the search. Selecting only the first translation found is done due to time restrictions: If a phrase is translated word by word, the set of possible translations (i.e. the permutation of all translations of all words of the phrase) can explode and, in the most cases, only a few results are obtained combined with a response time which is not acceptable. Therefore `mpro1R` uses only the first translation found but which consists of as many domain specific word translations as possible.

For the same reason, for German as the target language, the syntactic variants are sorted out. For example, there are two entries in the English-German dictionary for *human dignity*, *Menschenwürde* and *Würde des Menschen*. In these cases, the compound is preferred, because due to the query expansion all occurrences of the syntactic variant *Würde des Menschen* are equally found but the search for a compound is much faster than that for a phrase.

Each of the translations found undergoes a linguistic analysis to find the information necessary to expand also the translated query.

5.3 The Retrieval Algorithm

By means of the search patterns determined by the query expansion, different look-ups in the corresponding indices are performed, depending on the language of the patterns. The pattern has to match the index keys exactly, to avoid hits such as *entsenden* by looking up *senden*, for instance.

Searching German Terms For the monolingual German retrieval, the following look-ups are done:

1. Looking up the lu-index with the value of the lu-feature i.e. using stem information:
This will retrieve documents in which the term occur exactly (3), in an inflected form (4), and due to the morphological analysis `Mpro` performs, also documents in which the search terms occur as defective noun² (5) as well as a verb with a separated prefix (6):

(3) Query: Jugend
lu-feature: jugend
hit: ...die nach dem Gesetz zum Schutz der *Jugend* ...index: jugend

²Only valid for German compounds.

- (4) Query: senden
 lu-feature: senden
 hit: ...darf nicht am gleichen Tag *gesendet* werden ...
 index: senden
- (5) Query: Informationsdienst
 lu-feature: informationsdienst
 hit: ...geregelten elektronischen *Informations-* und Kommunikations*dienste* ...
 index: informationsdienst (ori: Informations-)
- (6) Query: mitteilen
 lu-feature: mitteilen
 hit: ...und *teilt* dem Hauptprogrammveranstalter die zulassungsfähigen Anträge *mit* ...
 index: mitteilen (ori: teilt)

2. Looking up the t-index with the value of the t-feature, i.e. compositional information:
 This results in a list of documents containing compounds built with the search term as an element. Not only elements at the beginning (a) or at the end (b) of a compound are found but also insertions (c):

- (7) Query: senden
 t-feature: senden
 hits:
 (a) ...und weiteren *Sendelizenzen* ...
 index: senden#lizenz
- (b) ...keine *Werbesendungen* ausgestrahlt...
 index:werben#sendung
- (c) ...eine unbefugt errichtete oder betriebene Funk*sendeanlage* ...
 index: funk#senden#anlage

3. Looking up the ls-index with the value of the ls-feature, i.e. derivational information:
 This results in documents containing terms with the same derivation:

- (8) Query: Werbung
 ls: werben
 hit: ...Aussagen *werbenden* Charakters ...
 index: werben
- (9) Query: Personenschutz
 ls: person#schuetzen
 hit: ...Frages des *Persönlichkeitsschutzes* ...
 index: person#schuetzen

For compounds, the provided compositional and derivational information is further evaluated and exploited by the following additional three look-ups:

4. Looking up the lu-index with the values of the t- and ls-features of the compound elements:
 This retrieves documents containing the syntactic variants of the input compound. Because

no information about syntactic structure is used, a so-called *environment* is determined, in which the compound parts have to occur in order to be identified as a syntactic variant. The environment is calculated as follows:

$$(\mathbf{n} - 1) * \mathbf{3}$$

whereas

- 'n' is the number of the elements a compound consists of
- '3' is a fixed factor which represents the distance in which the elements may occur

$$-3 \leq wn_i - wn_{i+1} \leq 3,$$

whereas wn_i is the value of the *w_{nrr}-feature* of word i .

The factor **3** is selected to catch possible occurrences of function words and/or adjectives between the parts.

(10) Query: Personenschutz

t/ls-features: person#schutz person#schuetzen,
environment: 3

hits: (a) ...zum Schutz von Personen ...

index: schutz person

environment: 2

(b) ...Schutz der betroffenen Person ...

index: schutz person

environment: 3

5. Looking up the lu-index with the value of the t/ls-features whereas the parts of the compounds occur outside the environment:

(11) Query: Personenschutz

t/ls-features: person#schutz, person#schuetzen,
environment: 3

hit: ...zum Schutz lebenswichtiger Interessen der betroffenen Person

(to protect the vital interests of the data subject or of another person)

index: schutz, person

environment: -5

A further evaluation of the underlying syntactic structure is certainly more precise but needs an efficient representation in order so that the performance does not decline.

6. Looking up the ls-index with the values of the t- and ls-features of the compound parts. This produces a list of documents containing *semantically similar* terms. These are terms which point to a common concept in a virtual hierarchy (i.e. all elements of the 'transitive closure' of the particular concept denoted by the compound).

(12) Query: Personenschutz

t/ls-Features: person, schutz schuetzen

hit: ...zum *Schutz personenbezogener* Daten ...

(protection of personal data)

index: schutz person#beziehen

For phrases, the topmost result list consists of documents which contain the elements of the phrase exactly (excluding function words). For example, to search for *gemeinsamer Markt* (common market), the hit *innerhalb des gemeinsamen europäischen Marktes* (within the common European market) is an exact one. The next list contains documents in which one phrase element occurs only as part of a compound, the others occur exactly (ex.: *gemeinsamer Binnenmarkt* (common internal market)). All further results lists are analogously calculated.

Searching English or French Terms Due to the less complex morphology of English or French, only two features are used for the query expansion. Therefore, maximally, three different result lists can arise from the following look-ups:

1. Looking up the lu-index with the value of the lu-feature:

This results in a list of documents which contain terms with the same stem:

(13) Query: advertising

lu-feature: advertising

hit: ...from revenue from *advertising* on TV 2 ...

index: advertising

(14) Query: publicité

lu-feature: publicité

hit: ...de restreindre la *publicité* en faveur du tabac ...

index: publicité

2. Looking up the ls-index with the value of the ls-feature:

This results in documents mentioning terms with the same derivation:

(15) Query: advertising

ls-feature: advertise

hits: (a)...it shall be prohibited to *advertise* tobacco products, ...

index: advertise

(b) ...The contents of *advertisements* ...

index: advertise

(16) Query: émission

ls-feature: émettre

hit: ...le service privé de radiodiffusion sonore *émettant* le même programme, ...

index: émettre

The formation of compounds in English and French does not take place via concatenation as in German, but English or French compounds usually consist of several words. In the current version of the MPROIR system, these compounds are not yet indexed (but will be in the future), so the same thumb rule already described above to identify syntactic variants of a German compound (cf. 4.) is applied to detect English and French multiword units (compounds). Thus, the result for a search of a multiword unit can produce the following result lists:

3. Looking up the lu-index with the value of the lu-features:

This results in a list of documents which contain terms with the same lu-value as the parts of the compound, within the defined environment:

(17) Query: television advertising
lu-features: television advertising
environment=3
hits: (a)...concerning Radio and *Television Advertising*...
index: television advertising
environment: 1

(b)...that religious *advertising* on UK *television* ...
index: television advertising
environment: 3

(18) Query: publicité télévisée
lu-features: publicité télévisée
environment=3
hits: (a) ...collaborer à la *publicité télévisée* ...
index: publicité télévisée
environment=1

(b) ...la commission de la *publicité* radiophonique et *télévisée*, ...
index: publicité télévisée
environment=3

4. Looking up the ls-index with the value of the ls-features:

This produces a list of documents which contain terms with the same derivation as the parts of the compound, within the defined environment:

(19) Query: television advertising
ls-features: television advertise
environment=3
hit: ...radio and *television advertisements* shall be lawful, ...
index: television advertise
environment: 1

(20) Query: Publicité télévisée
ls-features: publicité téléviser
environment=3
hit: ...à la *télévision*, la *publicité* ne peut être diffusée ...
index: publicité téléviser
environment=3

5. List of documents in which terms with the same lu-value or the same derivation of one of the compound elements occur outside the environment:

(21) Query: Television advertising
lu/ls-features: television advertising advertise
environment=3
hit: ...to *advertise* pharmaceutical and health products on *television* ...
index: advertise television
environment: 6

- (22) Query: Publicité télévisée
 lu/l_s-feature: publicité télévisée téléviser
 environment=3
 hit: La *publicité* diffusée à la radio et à la *télévision* doit ...
 index: publicité téléviser
 environment=7

Also, this methodology can certainly be improved by using a term recognition component for indexing purposes (i.e. identification of English and French compounds or syntactic variants) as well as for the retrieval.

5.4 Document Classification

In most retrieval systems, the retrieved documents are output with a ranking which gives the user an idea how relevant the document is according to the queried term. The weight is often based on the frequency of the term in the retrieved document compared to its frequency in all documents. In MPROIR, the information provided by the features lu, t, ls and for compounds additionally the extracted parts of the t- and the ls-feature are used for a classification of the retrieved documents as shown in the figures 1 and 2. For all input types, single word or phrase, and languages the topmost result list contains documents found by using stem information (lu-value). Within the search in German documents, this list is followed by the documents containing the search term as a part of a bigger compound (compositional information). The next result list consists of documents containing terms with the same derivation, whereas the derivational information gives no hint for the meaning. This means this list could contain incorrect hits due to homonymy. For instance, searching for the German word *Richter* (judge) with the ls-value *richten*, this list contains hits such as

(23) ... *Werbung, die sich auch an Kinder oder Jugendliche richtet ...* (*Advertisements also addressed to or using children or adolescents*) or

(24) ... *der verschiedenen religiösen und weltanschaulichen Richtungen ...* (*the different religious and ideological views*).

Searching for the English word *means* with the ls-value *mean*, this list will contain also hits such as

(25) ... *any trading stock, within the meaning of section 100 of the Taxes Act ...* or

(26) *a fairness complaint means a complaint to the BSC ...*

For compounds, there are three further categories for German and two for English and French: the first contains the syntactic variants of the searched compound. This list can include incorrect hits such as

(27) Query: Fernsehwerbung

hit: ...oder Rundfunksendern (Hörfunk und *Fernsehen*), aus *Werbung*, ...
 (...or broadcasting (and Television), from advertising...)

(28) Query: television advertising

hit: The time for news, sports broadcasts, radio and *television games, advertising*, ...

(29) Query: émission de radiodiffusion (broadcasting transmission)

hit: La *radiodiffusion* est l'émission et la *transmission*,...
 (Broadcasting is the provision and transmission ...)

The next contains documents in which the elements of a phrase (or of a compound) occur outside the defined environment

- (30) Query: Fernsehwerbung
 hit: ...Sendezeit im *Fernsehen*, die nicht aus Nachrichten, Sportberichten, Spielshows, *Werbung* ...
 (...programme time on television which do not consist of news, sport broadcasts, game shows, advertising...)
- (31) Query: gemeinsames Programm
 hit: ...Übergang von den nationalen Märkten zu einem *gemeinsamen* Markt für die Herstellung und Verbreitung von *Programmen* sichern ...
 (...the transition from national markets to a common programme production and distribution market ...)
- (32) Query: television advertsing
 hit: ...a natural person operating a radio or *television* broadcasting organization, persons working in the area of *advertising* ...
- (33) Query: publicité télévisée
 hit: ...En télévision, la *publicité* ne peut être insérée dans les journaux *télévisés* ...
 (On television, it is not permitted to insert advertising in news programmes ...)

The result list, only produced for German input, contains documents containing semantically similar terms such as:

- (33) Query: Fernsehwerbung
 hit: Das Landes- und Regional*fernsehen* - mit Ausnahme des sich nicht auf Film spezialisierten Programmanbieters - hat sechs Prozent der *Werbeeinnahmen*...
 (National and regional televisions, with the exception of broadcasters specializing in programmes other than films, shall appropriate six percent of their advertising revenues ...)

The ranking of the documents mirrors the relevance related to precision of linguistic information used to retrieve a document: a document retrieved by stem information is more relevant to the query than a document retrieved by derivational information. It expresses at the time the degree of precision, the lower a result is ranked the lower is the precision, i.e. the higher is the probability that mismatched documents are retrieved.

6. Conclusion and Future Work

6.1 Effectiveness

A major drawback of the use of linguistic information within the retrieval process are the costs, mainly caused by an increasing response time. In order to develop an effective system a good balance between cost and effectiveness has to be found. Within the EMIS system, the user has the choice between a fast search based only on stem information and a search which makes additional use of compositional and derivational information and which could be therefore more time intensive. For example, the search for exact hits of *digital television* in all documents lasts 9 seconds (20 hits), and a search for all documents about 38 seconds (50 hits). The differences in response times depend apart from the extended linguistic processing and the evaluation of these information within the retrieval algorithm on several other factors such as the type of input (single word vs. compound vs. phrase), the frequency of the term in the document repository and the number of translations. For German also the number of elements of a compound, and

how frequent these parts are in turn part of other compounds plays an important role.

Ignoring the time aspect, the most common measures to evaluate the effectiveness of a retrieval system are *precision* which gives the percentage of the retrieved documents that are relevant, and *recall*, the percentage of relevant documents that were retrieved. Incorporating linguistic knowledge will certainly increase the effectiveness. Using only stem information achieves a better recall because the system will find a lot more relevant documents with an higher precision as the examples above show.

In our system, we improve the precision also by applying a shallow parsing to the documents after their morpho-syntactic analysis in order to disambiguate to syntactic homonyms, which reduces the number of wrong hits. On the other hand, the query is expanded by compositional and derivational information and at least the use of derivational information can reduce the precision of the system due to homonyms. Because EMIS operates in a 'closed world', there occur only a few of such homonyms, and the precision reduction is caught more or less by the better recall. Nevertheless, applying this type of retrieval on unrestricted document archives would result in serious drawbacks. Measures to overcome them are described in the next section.

For cross-language retrieval, the recall depends on the quality of the translation. In the EMIS system, the quality is improved by taking advantage of the documents which exist in all languages covered. During the 'normal' working, the system guarantees that a document is only scored once whether it is available in one or in several languages (generally the document in the query language is then preferred). To improve the recall, the parallel documents were manually aligned and the translations found used to increase and/or to add domain-specific information to the bilingual dictionaries. Afterwards a new run was performed again over all documents in order to evaluate whether the result lists all versions of a document. This is very time consuming undertaking, and we consider using an alignment tool in the future.

6.2 Future Work

To improve MPROIR competence for CLIR in documents which do not belong to a certain domain, future work will mainly focus on the improvement of the indexing methodology, the translation component and the involvement of further linguistic information such as part-of-speech and semantics.

Indexing Using a controlled vocabulary provided by an existing thesaurus, a term recognition component could be applied to identify multiword terms which can then be indexed. Incorrect hits such as in (27) – (29) can be avoided. Instead of the thesaurus, the bilingual dictionaries could also act as such a controlled vocabulary. This will make the use of the thumb rule unnecessary. At the IAI, the tool AUTINDEX (Ripplinger & Maas, 1999a) which performs an intelligent indexing based on the linguistic processing as described above, is currently under development. The German component of the AUTINDEX system is already in use at a German producer of bibliographic databases.

For a set of parallel multilingual documents as in EMIS, a term recognition component together with an alignment tool could also help to construct or to improve such a controlled vocabulary and therefore the translation.

Translation Component There are several ideas to improve the translation component. One is to make more implications of the domain specific knowledge and the underlying classification scheme. For instance, if no domain specific translation is found, the translations

belonging to related domains could be preferred. Another possibility would be to develop the translation component towards a more interactive device which allows the user to select one or more translations from a list generated by the system. We also envisage, that a special domain can be selected by the user. This will have the advantage, that the user does not have to select special translations from the set provided by the system, he has only to determine a particular subject area which is often much easier for a user, who has only a passive knowledge of the target language. The system will then automatically use only the translations for the selected domain(s). Spurious translations and the resulting reduce of precision are thus avoided.

Part-of-speech Information For query expansion and retrieval, further information already provided by MPROis currently under consideration, for instance the part-of-speech. On the basis of the following examples, the usefulness of part-of-speech information is briefly discussed. The search for *Recht/right* in an arbitrary set of documents³ results among others in the following hits:

- (1) ... ohne rechten Widerhall ...
- (2) ... erhebliche rechtliche Probleme ...
- (3) ... gegen die Normen internationalen Rechts ...
- (4) ... to bend her right knee ...
- (5) ... he zeroes right in on the mood ...
- (6) ... to realize the right to self-determination ...

The search for *Menschenrechte/human rights* in:

- (7) ... eine recht grosse Zahl von Menschen ...
- (8) ... ihr Immunsystem ist dem menschlichen recht ähnlich ...
- (9) ... nach menschlichem Recht ...

Part-of-speech information, i.e. the syntactic category of queried term and the indexed word have to be the same, is not very useful for our purposes. For instance, for the German query and its English translation *Recht/Right* the hits (1), (4) and (5) can certainly be avoided by using the category knowledge, because in these cases *Recht* is not a noun (we suppose the German query obeys the German spelling rules). But this would also eliminate (2) as a relevant hit.

Using part-of-speech information provided with a compound (Mpro provides the syntactic category for each element), can be used to avoid incorrect hits i.e. identify occurrences which are not syntactic variants (7), (8), but also suppress relevant hits such as (9). As the few examples above have shown, using only part-of-speech can reduce the effectiveness of the system by inclining the recall. Therefore further tests to identify the scenarios in which part-of-speech can be usefully exploited have to be carried out.

Semantics Mpro provides also semantic information by means of the features *s* for single words and *ss* for compounds. Using this information, for the input terms *Recht* and *Menschenrechte*, the values of the *s/ss-feature* are *Gesetz* respectively *agent#gesetz* the hits (1), (2), (4), (5), (7) and (8) can be avoided. Only occurrences of the terms *mensch* and *recht* with

³For demonstration purposes a subset of the TREC-8 corpus is used

the corresponding semantic values, i.e. *agent* respectively *gesetz* are considered as hits (3), (6) and (9). This will also avoid incorrect hits given in the examples (23)-(26).

As the examples show, semantic knowledge can contribute to a better precision. Due to the fact that this kind of information can easily be integrated into and exploited by the system, i.e. adding the values of the semantic features to the index and evaluate the values during the retrieval by adding a further condition, we will focus on this knowledge in the near future.

References

- Gachot, D.A. et al (1998). The SYSTRAN NLP Browser: An application of Machine Translation Technology in Cross-Language Information Retrieval. In G. Grefenstette, *Cross Language Information Retrieval* (pp. 105ff).
- Gaussier, E. et al (1998). Xerox TREC-6 Site Report: Cross Language Text Retrieval. In *Proceedings of the Sixth Text Retrieval Conference (TREC6)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD.
- Gay, F.C. et. al. (1999). Manual Queries and Machine translation in Cross Language Retrieval at TREC-7. In *Proceedings of the Seventh Text Retrieval Conference (TREC7)*, National Institute of Standards and Technology (NIST), Gaithersburg, MD.
- Grefenstette, G. (1998). Cross Language Information Retrieval. Kluwer Academic Publishers.
- Kay, M. (1995).] Multilinguality. In *Survey of the State of the Art in Human Language Technology*, Elsnets Publication.
- Kraaij, W.& Pohlmann, R. (1996). UPLIFT – Using Linguistic Knowledge in Information Retrieval. Technical Report, OTS Working Papers, Utrecht University.
- Krüger, F. (1997). Nicht-lineares Information Retrieval in der juristischen Informationssuche. Elwert Verlag, Marburg.
- Landauer, T.K.& Littman, M.L. (1990). A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*.
- Liddy, E.D. (1998). Enhanced Text Retrieval Using Natural Language Processing. In *ASIS Bulletin*, April/Mai.
- Maas, D. (1996). MPRO – Ein System zur Analyse und Synthese deutscher Wörter. In: Roland Hauser (ed), *Linguistische Verifikation*, Max Niemeyer Verlag, Tübingen.
- Ripplinger, B. (1998). EMIS – A Multilingual Information System. In *Machine Translation and the Information Soup, Proceedings of the AMTA '98* (pp. 506ff), Springer.
- Ripplinger, B. & Maas, D. (1999).] AUTINDEX - Automatic Multilingual Indexing and Classification. IAI Memo.
- Ripplinger, B. (1999). EMIS – An Multilingual Information System on European Media Law. IAI

Working

Paper, forthcoming.

Sheridan P. et al (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain.
In

Proceedings of the First Conference on Research and Advanced Technology for Digital Libraries, Pisa.

Infoseek: User Guide: How to search. <http://www.infoseek.com>

Linguistic Inside: Search Engines Get Grammatical, *Language International* 9.6, (pp. 10,29).

Preissuchen – WWW-Suchmaschinen, Kataloge und Metasucher im Vergleich. *ct* 23/99 (pp. 162ff).