

The Use of NLP Techniques in CLIR

Bärbel Ripplinger

IAI

Martin-Luther-Str. 14
66111 Saarbrücken, Germany
babs@iai.uni-sb.de

Abstract

The application of NLP techniques to improve the results of information retrieval is still considered as a controversial issue, whereas NLP techniques in cross-language information retrieval (CLIR) are already well established, at least for the translation part. In this paper, the CLIR component - MPRO-IR - is represented which has been developed as the core module of a multilingual information system in a legal domain. This component uses not only the lexical base form for indexing but, in addition derivational information and for German information about the decomposition of compounds. This information is provided by a sophisticated morpho-syntactic analyser (MPRO) and is exploited not only for query translation but also for query expansion as well as the search and the document ranking. MPRO-IR is developed for a domain with a focus on recall. The objective of the CLEF evaluation was to assess this linguistic based retrieval approach in an unrestricted domain. The focus of the investigation was on how derivation and decomposition can contribute to improve the recall. Due to time restriction we could only submit one cross-language run with English as query language using the phrase search component of MPRO-IR which is restricted to search only for one translation, and where the queried terms have to occur in the same sentence. These restrictions result in a small number of retrieved documents, and thus in a low performance.

1 Introduction

The MPRO-IR system is a CLIR system based on query translation and focuses rather on a better recall than on a balanced recall and precision. To improve the recall, the system tries to take advantage of a sophisticated linguistic processing component whose results are used in the monolingual retrieval modules. Based on the output of a morpho-syntactic analysis which provides the full range of morphological information, not only inflection which would correspond to the power of a stemmer (Porter stemmer) but also derivational and decomposition of compound nouns is exploited. This information is used for the indexing, query expansion, search and document ranking. The translation component takes advantage of the provided part-of-speech as well as of the syntactic structure of the source query. Section 2 gives a short overview how this information is obtained and exploited in the system.

For the CLEF, as a first time evaluation within the TREC framework, we did one official run mainly to test our system in an unrestricted domain. We carried out the retrieval by querying only the English title section of the topics and using a retrieval component especially developed for phrase search in a legal domain, i.e. the whole phrase has to occur in the same sentence. But the main aim was to investigate whether derivational information, and decomposition of nouns could contribute to a better recall. As discussed in section 3 the restrictions of MPRO-IR's phrase search are too strong to get a satisfying performance. They don't even allow a final conclusion whether the application of the additional linguistic information improves the recall or not.

2 Mpro-IR System Description

The CLIR component MPRO-IR has been developed as the core component of a multilingual web-

based information system on European Media Laws¹. The document basis is multilingual, there are documents in German, English, and French. For these languages, an interface is available that enables the users to enter their queries in the selected language. The design of MPRO-IR is guided according to the requirements an information retrieval system in a legal domain has to satisfy: It has to support the lawyers' work which means finding as much as possible information about a certain subject. In terms of IR, the IR component should provide the best possible recall. The design of the system also had to take into account that the domain is relatively new and neither a thesaurus nor an approved term list is available, thus queries using an uncontrolled vocabulary are necessary. In addition, the type of queries has some impact on the design: The system has to be capable of processing single word queries such as *advertising*, compound terms as *subliminal advertising* as well as complex phrases like *actions leading to competition distortions*, *private broadcasters' obligation to provide information*, ... In the legal domain, such phrases often have to occur within one sentence to be relevant, therefore a special phrase search component has been developed which searches the input query within this restricted space. However, to allow the search of each of the meaning bearing terms within a whole document, a Boolean search facility is also provided to the users.

Independently of the search facility used, the input query as well as the documents are undergone a linguistic processing to take advantage of the information provided.

The Linguistic Processing

Stemming is the NLP technique which is frequently used and successfully applied in IR systems. A standard tool is the Porter stemmer [7] which achieves a normalisation by simply chopping off suffixes. To overcome some of the serious deficiencies of such stemmers, for instance *general* is mapped to *gener*, and *distribute* to *distribut*, both no lexical base forms, and thus lead to improper connotations, advanced stemmers are developed and combined with a lexicon [4] to verify the identified stem. This approach produces far better results, it avoids error as shown above but others such as the mapping of *distributed* to *distribut* still occur. In this case, the word *distributed* cannot be found in the dictionary. Also irregular plural (*media/medium*) or declination forms (*went/go*) cause errors. The main drawback of this approach lies thus in the coverage of the lexicon.

For languages with a rich declensional morphology such as French or German the results of such a stemming are rather unsatisfying because considering only inflection (or even suffix reduction) is not enough (cf. [6], [12]). For instance, the stemming of the German past participle *gegangen* (gone) to *gang* results in a wrong form (the correct one is *gehen/to go*). German verbs as well as French verbs such as *aller* (to go) or *recevoir* (to get) have numerous forms which makes it almost impossible to stem them by using suffix algorithms. For German, the compound formation leads often to failures because of the underlying highly productive morphological process (cf. [3]).

In MPRO-IR, the MPRO programme package [5] developed at the IAI is used for the linguistic processing, and its major features will be described in the following. MPRO has been primarily developed to process German language but is now available for different languages (including Eastern European languages). However, the same level of functionality as the German module is not available for all language modules. MPRO performs a morpho-syntactic analysis consisting of a lemmatisation, a part-of-speech tagging, and for German, a compound analysis as well as optionally, an additional syntactic and semantic disambiguation evaluating mainly context information. For the reduction of syntactic ambiguities there is also a shallow parsing component available for each language.

The morpho-syntactic analysis is combined with a look-up in a word-form dictionary. In a first step, the word-forms are looked up in a special tagging dictionary, for which an entry looks as follows:

```
{string=Word-form,c=w,sc=CAT,lu=Citation-form,...}
```

where CAT is the category. Nouns, verbs, adjectives, and derived adverbs are looked up in a morpheme lexicon. This morphological dictionary contains allomorphs but also some irregular word-forms which cannot be identified in another way as well as variety of toponyms and other names. Each entry shows how the associated stems behave morphologically, as shown in the

¹ A demonstrator is available at <http://www.iai.uni-sb.de/EMIS/emis.html>.

examples below:

```
{string=corrupt,c=a,n={ness=quality}}
{string=corrupt,c=v,n={ion=massnahme},a={ible=able},t={c=v,double=no,end=s,funct=no}}
```

To reduce overgeneration we can also prohibit prefixes or certain nonsensical compounds. For each word-form the morphological analyser produces at least one description which is represented as attribute-value pairs. In the following, the analyses of the English noun *corruption*, the verb *corrupt*, and adjective *corrupting* are given (only the features of interest are shown):

```
{string=corruption,lu=corruption,ds=corrupt~ion,ts=corruption,ls=corrupt,
t=corruption,c=noun,s=massnahme,...}

{string=corrupt,lu=corrupt,ds=corrupt,ts=corrupt,ls=corrupt,t=corrupt,c=adj,...}
{string=corrupt,lu=corrupt,ds=corrupt,ts=corrupt,ls=corrupt,t=corrupt,c=verb,...}

{string=corrupting,lu=corrupting,ds=corrupt~ing,ts=corrupting,ls=corrupt,
t=corrupting,c=noun,s=vn,...}
{ori=corrupting,lu=corrupting,ds=corrupt~ing,ts=corrupting,ls=corrupt,
t=corrupting,c=adj,...}
{string=corrupting,lu=corrupt,ds=corrupt,ts=corrupt,ls=corrupt,t=corrupt,c=verb,...}
```

The feature *ds* contains the morphological derivation, and *ls* the respective normalised form. The features *s* and *ss* (for compounds) contain semantic information. In the example above, all three words have the same derivation. For German words, a compound analysis is performed additionally (cf. Example below), and the result is given in the feature *ts* and its normalised form² in feature *t*. These features are also assigned for English analyses but correspond always to the *lu* feature.

Due to a special treatment some defective noun constructions in German - such as these occurring in coordinations like *Informations- und Kommunikationsdienst* (Information and Communications services) - are recognised. MPRO assigns the missing head information by using a lookahead algorithm:

```
{string=Informations-,lu=informationsdienst,ts=informations#dienst,
t=information#dienst,ds=informieren~ation#dienst,ls=informieren#dienst,c=noun,...}
{string=und,lu=und,c=w,...}
{string=Kommunikationsdienst,lu=kommunikationsdienst,ts=kommunikations#dienst
t=kommunikation#dienst,c=noun,...}
```

Although MPRO is very complete, a strategy for handling unknown words is provided. Three cases can be differentiated:

- The word-form can not be analysed at all:
MPRO marks this word with the feature **state=unknown** and classifies the word as 'noun', for instance
{string=settlor,lu=settlor,ds=settlor,state=unknown,c=noun,s=n,}
- The word-form can partly be analysed:
MPRO tries in each case to assign the most appropriate information. For instance:
If a string consists only of numbers such as *1864* the word get as category *cardinal number* (*c=z*), and **morpho** provides an analysis whereas the value of the lexical unit is identical with the string: {string=1894,ds=1894,ls=1894,c=z,lu=1894,s=year,...}
- The word form is analysed but not found in the lexicon:
Strings which consist only of capital letters such as *CNN* are marked as acronyms, and have as the part-of-speech *c=noun*: {string=CNN,lu=CNN,ds=CNN,ls=CNN,c=noun,s=acronym,...}

The analyser recognises lexicalised multiword units such as *look up*, *United States*, German

²Hyphens and German 'fuge' elements are removed.

prefix verbs *mitteilen*, fixed expressions such as *in Bezug auf*, *de facto*, abbreviations like *etc.*, *i.e.* as well as proper names such as *Bill*, *Berlin*.

After this analysis, for German the output can be further disambiguated by evaluating context information, i.e. if the first letter of word-form is capitalised, and the word is not the first in a sentence, it must be a noun. In a final step, a shallow parsing can be applied to reduce other syntactical ambiguities such as verb/noun readings. This parsing process can also be performed on English and French output of the morphological analysis to get an almost unambiguous representation. MPRO does not reduce ambiguity where the correctness of the decision is doubtful.

In the remainder of the section, it is described how these results of the morpho-syntactic analysis is applied for various stages of the IR process.

The Retrieval

For all three, indexing, query expansion, the search together with a document ranking the information provided by the features *lu*, *ls* as well as *t* (currently for German only) are exploited.

Based on the analyses of the documents, several indices are built up: One using the information about the lexical unit (i.e. the *normalised form*), one using the derivational information, and for German a third index is constructed with the decomposition information. Though English and French nouns have a *t*-feature we have not exploited this kind of information because this information is subject of an ongoing revision of the English and French morpheme lexicon (see above). With each key the document identification number, the sentence number (*snr*), the word number (*wnr*), as well as the word-form (the form of the word as occurring in the text) are stored. Function words (entries with *c=w*) are discarded from the indexing. This process is done within a preparation phase.

At search time, the queries are processed by the same morpho-syntactic analysis as the documents. For the monolingual search, the function words are removed from the analysis output, and for the meaning bearing words the values of the *lu*-, *ls*- and, for German queries, the *t*-feature are extracted to construct a set of search patterns. For the input query *Competitiveness of European industry* the set of search terms consists of *competitiveness*, *compete*, *european*, *europa*, *industry*.

For the cross-language retrieval, we decided to translated the queries, and to carry out a monolingual search afterwards. This approach seems more appropriate because legal information is highly related to the original wording, and machine translation systems provides only a poor quality [2]. The input to the translation component is the complete morphological analysis of the query. MPRO-IR uses a shallow translation tool which performs a lexical transfer based on huge transfer lexicons (coverage of the English-German lexicon is about 400.000 entries) comprising single words, abbreviations, compound terms but also fixed phrases. 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 whereas the translation is guided by the part-of-speech, i.e. for verbs only the translations for verbs are assigned. The translation output is undergone by a shallow parsing based on a phrase grammar to get only one possible translation whereas the syntactic representation of the source is taken into account. For German as target language, the syntactic variants of a term are additionally 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 then that for a phrase.

The search itself consists of several look-ups in the different indices, for each content bearing term the following look-ups are done:

1. Looking up the index built over the lexical base forms (*lu*-index) with the value of the

lu-feature

2. For German only: Looking up the index built over the t-feature (t-index) with the value of the t-feature to find compounds with the queried term as element
3. Looking up the index built over the derivations (ls-index) with the value of the ls-feature

For compounds, the different formation in English and French compared to German leads to a different search strategy: Having in mind that open compound terms in English and French has almost a fixed word order, we defined a *distance factor* to decide whether the occurrence of the two or more words represent an open compound or not. Based on statistical data the longest distance between each meaning bearing word of a phrase is fixed to 3. This allows to classify occurrences of *advertising in UK's television* as exact hit of *television advertising*. For English as well as for French compounds, the occurrences of each word within a phrase is evaluated against this distance factor using the word number provided by the index, and sorted into the following three lists:

1. The lu-values looked up in the lu-index of each element occur within the determined distance.
2. At least for one element only the derivation occurs within this distance.
3. All other occurrences.

We apply this distance measure also to German to find syntactic variants of compound terms:

1. Looking up the lu-index with the values of the t- and ls-features of the single compound elements. This retrieves documents containing the syntactic variants of the input compound, for instance searching for *Verbraucherschutz* (Consumer protection) hits *zum Schutz der Verbraucher* as well as *um die Verbraucher zu schützen*.
2. Looking up the lu-index with the value of the t- and ls-features whereas the parts of the compounds occur outside the environment.
3. 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). For instance, the search for *Verbraucherschutz* found hits such as *Schutzbestimmungen bezüglich der Verbraucherdaten* (regulation to protect consumer data).

For phrases, the topmost result list consists of documents which contain the elements of the phrase exactly (excluding function words). The next list contains documents in which at least one phrase element occurs only as part of a compound. All further results lists are analogously calculated.

Usually the rank of a retrieved document is computed by the *tf*idf*. Using a weight based on frequency seems not to be adequate in this environment of a legal domain in which some terms occurs only once in a document which is much more relevant than a document in which the term occurs several times. Thus, in MPRO-IR, the documents are ranked by the information used to retrieve them, in the order of the lists described above. This ranking mirrors the relevance related to the reliability of the 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 of the retrieval. The results of the first list have a higher precision than those of the lower lists because the probability that mismatched documents are retrieved increases.

3 Mpro-IR in CLEF

We participated the first time in a CLEF/TREC evaluation to investigate how MPRO-IR developed for a special domain fares with unrestricted documents related to recall and precision.

Setting up the Experiment

Currently the MPRO-IR system covers only the languages German, English, and French. To perform the CLIR task which additionally comprises the search in Italian documents, we integrated a small Italian component into MPRO-IR. To provide a sufficient coverage for this module, we analysed the complete Italian topics (titles, description, and narratives), and added unknown words (morphemes) to our monolingual lexicon. For the translation component we added only translations for the words occurring in the title sections of the topics. Thus the Italian morpheme lexicon has now 27.800 entries compared to the English morpheme lexicon with about 48.300 entries. We used English topics and retrieved documents in English, French, German, and Italian, therefore we added missing translations for the terms of the topic titles to the respective transfer dictionaries.

Retrieval Performance

Due to time and space restriction we could perform and submit only one run. Therefore we decided to perform a phrase search only over the titles sections of the topics, although we noticed that the type of queries was not always adequate for this kind of search. To build up the indices, texts were undergone a normalisation, i.e. we discarded all formatting information (including the title sections) which lead in some cases to a lower performance. This process was mainly done due to space limitations, the MPRO tool is able to indentify SGML tag but the analyses are unnecessarily blown up.

The overall result of the CLEF evaluation shows a low retrieval performance of MPRO-IR compared to the other systems. Taking into account that a very restricted retrieval component has been used

- all meaning bearing words have to occur in the same sentence, and
- only one translation is used

the outcome is not to bad. The results show more or less what we expected: For topics which are more or less incomplete sentences such as *French conscientous objector*, *supermarket ceiling in Nice collapses*, etc. we got none or only a few results (cf. Figure below).

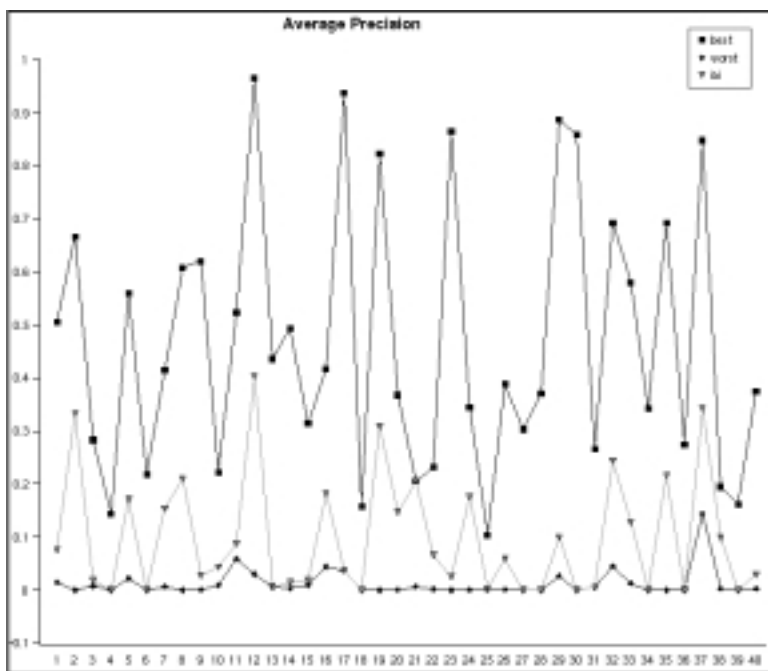


Fig.1 CLEF Results

For topics such as *European Economic Area*, *World Trade Organisation* etc. the results are better though not satisfying.

To get an impression to which degree the restriction to a sentence as search space is too strong, we performed a Boolean search for some sample topics (T6, T9, T10, T14, T21). For queries such as *French conscientious objectors (T6)*, *Methane deposits (T9)*, *Tourism in the US (T14)* a five times better recall is achieved, and a 30% improvement for the query *War and radio (T10)* compared to the run submitted to CLEF. We got the same result for the query *European Economic Area (T21)* because here we got no results in the monolingual retrieval.

Our main objective was to evaluate the use of derivational and decompositional information to improve the recall. Thus, we could conclude that most of the documents are retrieved by using the information of the lexical base form. Only a few others are retrieved on the basis of derivational information. Decomposition information which is only used for retrieving German documents depends on the type of compounds, and in a few cases also on the type of the single words forming a compound. No relevant occurrences of syntactic variants are found in the corpus. We got also only a few results by on the basis of the productive use of decomposition information, i.e. documents containing semantically similar terms. The main reason is certainly the restricted search space, furthermore the German compounds occurring in the queries (such as *Kriegsdienstverweigerer*, *Krebsgenetik*, *Golfskriegssyndrom*, *Nobelpreis*, *Alkoholkonsum*, ...) consist of words which are not frequently used in compound formation within the context of the respective query. Another reason is that only one translation is used (ex: *Methane deposit* is translated into German as *Methanlagerstätte* where in the documents often the synonym *Methanlager* is used).

Conclusion

The results of the CLEF evaluation are coincident with those we got from the evaluation of the retrieval algorithm within the EMIS system [10]. Also here most hits could be retrieved by using precise lexical units, and derivational information. Compositional information was also valuable to detect syntactic variants of German compounds. The improvement of the recall by so-called semantically similar terms is very poor. Because this approach is also very time consuming, we will concede this in favour of a better morpho-syntactic analysis. This will then provide the grounds for a better indexing by using a term recognition component, and a better translation component. For the query expansion on the monolingual side, we currently experiment with a method to add synonyms which will be automatically computed by translating the translations back to the source language. Whilst the search itself could be improved by taking advantage of the part-of-speech together with the semantic information already provided by the morpho-syntactic analyzer [9].

As the results here show the phrase search as implemented in MPRO-IR is useful in retrieval systems developed for a special type of domain where the search of complex phrases is necessary as in the legal domain. In retrieval systems dealing with unrestricted texts, a Boolean search achieves much better recall. With a Boolean search we could certainly get a better insight in the usefulness of derivational and compositional information in the retrieval process due to the higher recall.

The approach we pursue in MPRO-IR using a sophisticated morpho-syntactic analysis has shown that the recall can be improved by more precise identification of the lexical base units and the almost unambiguous representation of the documents and the queries. The possible impact of derivational and decompositional information has to be further evaluated. Results from the clef experiment have no significance so far. However, part-of-speech, currently exploited only for translation purpose together with semantic information can be expected to contribute to a better retrieval performance which still has to be proven.

References

- [1] Brill, E. *A simple rule-based part-of-speech tagger*. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992.

- [2] Kay, Martin. *Multilinguality*. In Varile, G. and A. Zampolli (Eds). **Survey of the State of the Art in Human Language Technology**, Elsnet Publication 1995.
- [3] Kraaij, W., R. Pohlmann. *UPLIFT - Using Linguistic Knowledge in Information Retrieval*, Technical Report, University of Utrecht.
- [4] Krovetz, R. *Viewing Morphology as an Inference Process*. In *Proceedings of the 16th International Conference on Research and Development in Information Retrieval*, (SIGIR'93), Pittsburg, 1993.
- [5] Maas, D. *Multilinguale Textverarbeitung mit MPRO*. In G. Lobin et al.(eds): **Europäische Kommunikationskybernetik heute und morgen**, KoPäd, München, 1999. <http://www.iai.uni-sb.de/global/memos.html>
- [6] Popovič, M., and P. Willet. *The effectiveness of stemming for natural-language access to Slovene textual data*. In *Journal of the American Society for Information Science*, 43(5):384-390, 1992.
- [7] Porter, E. *An algorithm for suffix stripping*. In *Programm*, 14, 1980.
- [8] Ripplinger, B. *EMIS: A Multilingual Information System*. In Farwell, D., L. Gerber, E. Hovy (eds) **Machine Translation and the Information Soup**, Third Conference of the AMTA, Springer, 1998.
- [9] Ripplinger, B. *MPRO-IR - A Cross-language Information Retrieval Component Enhanced by Linguistic Knowledge*. In *Proceedings of the RIAO 2000*, Paris, 2000.
- [10] Ripplinger, B. *Linguistic Knowledge in a Multilingual Information System*, IAI Working Paper, forthcoming.
- [11] Strzalkowski, Tomek, F. Ling, J. Wang, J. Perez-Carballo. *Evaluating Natural Language Processing Techniques in Information Retrieval*. In Strzalkowski, Tomek (Ed.). **Natural Language Information Retrieval**. Kluwer Academic Publishers, 1999.
- [12] Tzoukermann, E., J. L. Klavans and Ch. Jacquemin. *Effective use of Natural language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing*. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval* (SIGIR'97), Philadelphia, 1997.