

Automatic Multilingual Indexing and Natural Language Processing

Bärbel Ripplinger
IAI
Martin-Luther-Straße 14
D-66 111 Saarbrücken
+49 681 3 89 510
babs@iai.uni-sb.de

Paul Schmidt
University Mainz
An der Hochschule 2
D-76 711 Gernersheim
+49 7274 508 35 252
schmidtp@usun2.fask.uni-mainz.de

ABSTRACT

The number of documents being collected by information brokers such as bibliographic database producers, libraries and publishers increases rapidly. The consequence is a huge demand for indexing and classification. So far this has had to be carried out manually. The system AUTINDEX, which is described in this paper offers tools for monolingual as well as for multilingual automatic indexing and classification by taking advantage of sophisticated language processing technologies and already existing special purpose language resources such as thesauri, classification schemes and large lexicons. It will be shown that the use of high quality NLP can achieve appropriate results.

Categories and Subject Descriptors

Text Representation and Indexing, Information Extraction, Lexical Acquisition, Natural Language Processing for IR

General Terms

Algorithms, Performance, Design, Reliability, Experimentation, Languages.

Keywords

Indexing, Classification, Linguistic Processing, Multilinguality

1. INTRODUCTION

The thrust of this paper is to present a method for automatic multilingual indexing based on high quality natural language processing (NLP) which is realised in a system called AUTINDEX. At present there is no commercial system which supports the human indexer in his/her intellectual task, so most publications (journals, conference papers, dissertations, reports, and books) are indexed intellectually by assigning descriptors in the respective language. Manual intellectual indexing is time-consuming, expensive, and often inhomogeneous due to different background knowledge and expertise of the human indexers. The system presented here supports information producers by providing a generic solution to automatic indexing and classification of documents in different languages. The system allows for quicker, cheaper, and more consistent population of information repositories. The paper focuses on the hypothesis that by using sophisticated, reliable, robust high quality NLP-based automatic indexing it is possible to achieve results comparable to manual indexing.

The AUTINDEX system introduced in this paper applies a two level approach for indexing and classification. It takes advantage of sophisticated NLP technologies to produce a set of keywords which are evaluated against existing special purpose language resources such as a thesaurus and a classification scheme. The major achievement is that it performs a form of 'conceptual indexing' for unrestricted text in that frequency statistics are calculated on the basis of semantic information rather than on word forms.

The paper is organised in the following sections:

- The next section sets the scene by presenting some work done so far in intellectual indexing as well as in indexing within NLP-based information retrieval (IR).
- The NLP components that are used to produce the representations for indexing in AUTINDEX are then introduced.
- The fourth section describes the indexing and classification method realised in AUTINDEX.
- A fifth section shows the results that were achieved using AUTINDEX in a realistic scenario.
- Finally a summary shows the impact on IR and the work envisaged in the near future.

2. RELATED WORK

The term *indexing* is used in two different ways: The task of manually assigning keywords that correspond to concepts of a thesaurus in order to describe a document is called 'intellectual indexing'. In information retrieval systems which deal with large text archives, indexing is done automatically, and the queries in most cases are not restricted to thesaurus concepts but are free. A lot of work that has been carried out within IR to increase the quality of the indexing of unrestricted text showed that indexing of concepts broader than the word seems to produce better results. At this point, IR and intellectual indexing converge, an area that only recently has attracted interest.

2.1. Intellectual Indexing

Commercial indexing systems such as CINDEX or MACREX support the human indexer in the index preparation and the processing (editing and formatting) of manually produced indexes. Most of them provide also a spell-checking facility. The time consuming intellectual task, however, - the assigning of descriptors to documents - is only supported by maintaining the actual list of terms used for the indexing.

Intellectual indexing is subject of only a few research projects. As an exemplification two of them are briefly discussed here.

At the University of California at Berkeley [10], an indexing method based on lexical associations and a controlled vocabulary has been developed. These associations which hold between lexical items occurring in the titles and abstracts and the controlled vocabulary which was extracted from abstracts indexed by human indexers were identified on the basis of statistical methods and used to create a dictionary storing these associations. This approach has only been proven on monolingual data. Limitations concerning the identification of descriptors that were reported are due to the lack of sufficient NLP.

In the CONDORCET project [1] carried out at the University of Twente, a so-called *controlled-term approach* was used to index scientific documents. Based on a *structured ontology*, concepts and relations rather than lexical items are used as indexes. This means that after the tagging of a document each lexical item has to be mapped onto the proper concept and/or relation. This is done by determining the syntactic structure and the deep structure of a sentence to get information about semantic roles. By means of a knowledge-based module, this deep structure is then mapped onto index terms. The approach increases the precision because concepts are mostly language-independent and non-ambiguous. The system also takes relations between the concepts into account which means that thesauri must provide such relations. This is not the case for most available classical thesauri. Only a few of them such as UMLS in the medical field or AGROVOC can provide such information. So, the extension and generalisation potential is rather limited. Another drawback of CONDORCET is the knowledge-based module which performs deductive matching. Both constraints limit down the applicability of the system to a particular domain. Multilinguality is not addressed in this project either.

The approaches for intellectual indexing show that linguistic processing is essential to identify terms as well as to map them onto the correct thesaurus concept. Unfortunately, using a parser for term recognition does not necessarily achieve better results. On one hand, parsers developed for NLP often lack robustness and thus applicability to unrestricted text. On the other hand they are often developed for a specific domain, and/or to experiment with a certain type of linguistic phenomena.

So, major drawbacks of current automatic indexing tools are that they are too much dependent on specialised resources such as thesauri of particular domains or that an intensive preprocessing is necessary to get the necessary information such as the mentioned lexical associations.

2.2. Automatic, Unrestricted Indexing

Traditional approaches to IR as well as to CLIR are based on surface word forms and their statistical distribution such as Latent Semantic Indexing (LSI) [5]. These approaches in practice are still the most successful ones, though they ignore the fact that terms may be ambiguous. Neglecting lexical ambiguity reduces the performance of IR as non-relevant documents are retrieved on the basis of the 'irrelevant' readings of terms. Separating relevant documents from non-relevant ones requires the determination of the reading of the queried terms occurring in the documents. Krovetz [4] showed that word sense disambiguation can contribute to a better performance of IR by disambiguating word senses for indexing

Lexical ambiguity has several sources. It may be due to syntactic category. For instance the words *duck* and *call* can both be verbs or nouns. But ambiguity can also arise from different semantic readings a word form may have, which may either be unrelated (homonymy), or related (polysemy). An example of homonymy is *bark* (of a dog) vs. *bark* (of a tree). An example of polysemy is *opening* (a window) vs. *opening* (a book) vs. *opening* (a shop). Different types of knowledge are necessary to disambiguate word senses, for instance knowledge about word co-occurrences, word collocations, (syntactic) knowledge about subcategorisation and in some cases world knowledge. Currently, the state-of-the-art in NLP is such that it does not allow for the general disambiguation of words in unlimited text.

In the following three main methods of using NLP (in monolingual as well as in cross-language) IR, resulting in performance gains are presented.

Stemming

The most prominent NLP technique used in IR is stemming. It is used as a basis for indexing (applied to documents as well as to the queries). Stemmers developed by Lovins [6] and Porter [11] provided the means to reduce word forms to their root based on simple suffix stripping. As these stemmers ignore word meanings and do not even use a lexicon, they produce a lot of errors. For instance *generalisation*, *generalise*, *generate*, *generation*, *generic* are conflated to *gener*, *car*, *care*, *career*, *caress* are conflated to *car*. Irregular forms represent a general problem to these stemmers. Words such as *medium*, *media* or *go*, *went* are not conflated at all. Root forms such as *gener* are not words at all, thus cannot be disambiguated, as the root form will not be found in the lexicon.

Numerous other attempts have been made with similar results. Nevertheless stemming with the Porter algorithm at least was considered useful for languages with a poor morphology such as English, less so for languages such as French, German, Finnish or Greek with a rich morphology. German exhibits an additional problem, namely the productive processes related to compounding which cannot be solved by stemming.

The next generation of stemmers were based on a morphological analysis comprising inflection and derivation to overcome some of the mentioned deficiencies.

For languages with a rich morphology substantial improvement compared to Porter-like stemmers is reported for the work done by Kraaij and Pohlmann [3], Tzoukermann, Klavans and Jacquemin [13] who use inflectional and derivational morphology, as well as some compound analysis. Common to these approaches is that they generate word lists by creating all possible variants occurring in the documents processed. These lists are then used as additional lexical resources. There are limits for languages which are highly productive, especially for languages which have compound formation by concatenating the elements into one single word such as German. So these methods even if they result in some performance gain are still prone to high error rates.

Phrase Indexing

To overcome some of the lexical ambiguities, the trend goes towards phrase indexing assuming that a phrase (multiword unit) is a better discriminator than a single word. For instance, if in addition to *television* and *advertising* also *television advertising* is indexed, documents can be retrieved in which the compound term occurs. Depending on the search algorithm this does not avoid documents also being retrieved in which only one or both terms

occur independently which means that these are not necessarily documents about *television advertising*. Some systems therefore offer a special operator which allow for determining the longest distance between two or more queried terms. The problem with this approach is that the distance factor has to be big enough to cover structures with the terms ordered in a different way (for instance *advertising in television*), or with an additional modifier (*advertising in UK television*). A similar problem is created by syntactic variants. For instance the German compound *Frequenzübertragung* (*frequency transmission*) may be expressed as '*mit hoher Frequenz übertragen*' (*transmit via high frequency*), i.e. as a complex syntagma. Another example also shows the richness of structures available *Hochfrequenzübertragung* (*high frequency transmission*) may be realised as *Übertragung mit Hochfrequenz*, *Übertragung mit hoher Frequenz*, *Übertragen von Signalen mit höherer Frequenz*, (*transmission via high frequency, transmission of signals using a higher frequency*). In an ideal situation all these syntagmas are mapped onto the same representation to be available for retrieval. This requires, though, syntactic parsing, especially a sophisticated parsing of NPs, and moreover information about derivation. Tzoukermann, Klavans and Jacquemin used a fully-fledged derivational morphology, together with a shallow parsing to identify French multiword terms and their syntactic variants. Their algorithm allows the expansion and conflation of morpho-syntactically related terms, e.g. *publicité télévisée* (*television advertising*), *publicité en télévision* (*advertising in television*), *publicité émit par télévision* (*advertising broadcasted in television*). Nevertheless they started with a term list, and generated all morphologically derived forms of each constituent of the multiword terms. Taking into consideration only syntactic, morphological and morphosyntactic variants, they reported an increasing indexing coverage up to 30% with a precision of 90% for correct identification.

The use of NLP raises the serious problem of robustness, i.e. linguistic processing must deliver a result in an appropriate time. Furthermore, syntactic information is sometimes not sufficient. In some cases a semantic analysis has to be carried out, for instance, semantic information about agent, object, modifier, etc are necessary to produce an unambiguous representation. This raises more problems of feasibility.

Concept-Based Indexing

The attempts to use semantic information in form of semantic networks for the monolingual retrieval process such as WordNet [9] or EuroWordNet [14] did not meet expectations concerning performance gains. Nevertheless, some recent CLIR systems such as ITEM [2] use EuroWordnet's synsets together with a stopword list for word sense disambiguation as well as for translation purposes. They reported an improvement of performance (29% compared to SMART) due to a more precise indexing.

Using concepts for indexing allows documents to be ranked as relevant to a query, even if the query term itself does not occur in the text, but only a related term which denotes the same concept. For instance, the synset *Berlin* denotes the concept *capital of Germany*. If a document is indexed as relevant for the concept *Berlin*, and the query contains the expression *capital of Germany*, all documents are found. But this approach requires full syntactic and semantic processing. For instance if the query contains the expression *former capital of Germany* then this phrase must be related to *Bonn* and not to *Berlin*.

A drawback of concept indexing by synsets is that only those words can be assigned to a particular synset and thus to a concept which are described as a surface realisation of such a synset. This

requires a 'normalised' representation of the syntactic structure, i.e. all possible variants have to be mapped onto the same representation (of a concept), or have to be listed as realisations of a particular synset. Although this approach seems promising in principle, EuroWordNet as well as WordNet cover only a few domains and are not available for general language. Thus, the conceptual indexing performed over unrestricted text must be coupled with statistics on word usage.

Most experiments described in this section use linguistic tools which are developed for a context often not relevant to indexing purposes or information retrieval. The promising results these 'handicapped' tools delivered leads to the hope that NLP techniques adapted to the specificities of retrieval systems can further improve them. Our paper tries to support this claim.

3. THE AUTINDEX NLP-COMPONENTS

In this section the NLP components which AUTINDEX is based upon will be introduced. There are four processing modules :

- Word form identification
- Lemmatisation and tagging
- Homograph disambiguation
- Robust shallow parsing

These NLP components have properties that are considered essential for NLP-based indexing. They are high quality, highly robust and reliable.

3.1. Word Form Identification

This module identifies sentence boundaries, single words and fixed expressions (including lexicalised multiword units). A sequence of up to ten units are looked up in the word form dictionary as well as in the lexicon for fixed expressions. If the look-up is successful, or if the concatenation can be identified as a number or date or acronym it is output as one word together with the information that goes with it. The output is a flat feature bundle as the following schematic (incomplete) example shows:

```
{string=W-FORM, c=WD, sc=CAT, lu=CIT-FORM ... }
```

It represents information about the wordstring (string), the syntactic category (c), about a syntactic subcategory (sc), about the 'normalised' string (lu).

If the dictionary look-up produces more than one result, all of them are output, represented as a sequence of attribute value pairs as in the following example which represents an NP 'die Beauftragten' (the Commissioners) where the German 'die' exhibits an ambiguity. 'die' may be an article or a relative pronoun:

```
{string=Die, lu=d_art, c=w, sc=art, spec=def, pctr=no, last=no, gra=cap, wnrr=1, snr=1}
```

```
{string=Die, c=w, lu=d_rel, sc=rel, ref={c=n, nb=sg, g=f};
```

```
{c=noun, nb=plu}, last=no, pctr=no, pctl=no, gra=cap, wnrr=1, snr=1}
```

```
{string=beauftragten, lu=beauftragter, c=w, sc=art, pctr=no, pctl=no, gra=small, wnrr=2, snr=1}
```

In addition to what was already described the entries exhibit information about 'graphical' representation, capital or small (gra), word number (wnrr) etc.. The relative pronoun in addition contains information about the antecedent in form of an alternation (ref).

3.2. Lemmatisation and Tagging

Lemmatisation and tagging is based on a morphological dictionary that contains tens of thousands of morphemes for each language and a morphotactic module that makes a segmentation of the words into morphemes by looking up the segments in the morpheme dictionary which contains all necessary information for morpheme combination and their inherent properties. To reduce overgeneration and avoid non-sensical morpheme combinations, there is a stop list that contains prohibited words. For instance, 'Mitgliedschaft' (membership) may be segmented into 'Mitglied' and 'schaft' (schaft being ambiguous, (-ship, shaft)). The feature 'rn' contains the information about prohibited rightmost intraword, and 'ln' about leftmost elements. So, the nonsensical 'Schaft' ('shaft') interpretation is discarded.

The lexical entries also have semantic information such as 's=act' (which means that a verb like 'exchange' is an action) or 's=agent&ano' which means that a noun like 'Mitglied' is an unanimate agent.

After the segmentation a morphotactic module concatenates the morphemes.

For German, MPRO (the morphological system) recognises more than 98% of the words. For the remaining small percentage a treatment is foreseen that consists of the following:

- If the word form cannot be analysed at all the word is marked 'state=unknown' and interpreted as a 'noun'.
- If the first part of a hyphenated compound, the non-head part, (e.g. in 'Blair-effect') cannot be analysed, the information of the second noun (the head part) is represented.
- If the word form can be analysed but not found in the dictionary, such as a sequence of capitals (e.g. CNN), it is interpreted as an acronym and a noun.

The whole of the analysis results in two types of information associated with the word form:

- Morpho-syntactic information for each category. For nouns: Gender, number, case, in addition degree for adjectives, for verbs tense, number, person, mood etc..
- Information on word structure and semantics: (E.g. for the German word 'Jugendschutz' (youth protection):

```
{string=Jugendschutz; t=jugend#schutz, cs=n#n, ts=jugend#schutz, s=coll, time#measure}
```

t, ts expose the elements of a compound, cs is the sequence of categories in a compound, s describes to which semantic class a word belongs to: 'Jugend (youth) is 'coll' collective or a time period and 'schutz' is a measure.

3.3. Homograph Disambiguation

Lemmatisation and tagging does not take into account word context within sentences. Thus, ambiguities remain. A tagging result with word ambiguities is useless and may spoil semantic evaluation. A homograph reduction module partially solves the problem. The algorithm used for this consists in a small set of rules that evaluate the context of words on the basis of word order regularities and lexical information available from the dictionary. Ambiguities that cannot be resolved reliably remain.

3.4. Shallow Syntactic Parsing

A shallow parsing component without recursive analysis resolves the remaining ambiguities. Above all, it reliably identifies noun

phrases. It determines the subject and the finite verb of a sentence and determines agreement. It consists of a number of phrase structure rules that are split into subgrammars which are successively applied. A subgrammar is applied as long as no rules fire. Then the next subgrammar starts working. So, the parsing is a procedural sequential processing.

There is no deep analysis, i.e. no subcategorisation information is systematically used, apart from some information about transitivity. Thus, some ambiguities still remain, e.g. case ambiguities of nominal elements. However, these are irrelevant in an indexing context. Usually, the shallow analysis is sufficient to completely disambiguate all words and there is all information available to be exploited in indexing on the basis of semantic frequencies.

To summarise the benefits of the NLP-components: In the first place they are complete, robust and reliable and all the linguistic structures necessary for indexing are delivered. The most important point is that all information concerning NPs is available.

4. AUTINDEX: INDEXING AND CLASSIFICATION

AUTINDEX combines intellectual indexing with an automatic free vocabulary indexing and classifies documents. Classification means to determine the domain the document belongs to. This is a certain type of conceptual indexing. The domain is fixed by an external classification scheme. The AUTINDEX approach performs two major steps: The first one is indexing with uncontrolled vocabulary using NLP techniques to produce an unambiguous representation, and to identify multiword terms. The result is a list of keywords, a list of most relevant discriminators. This list is refined in a second step by using a controlled vocabulary in form of thesauri and a classification schema. AUTINDEX works with general language. No knowledge base or extensive preprocesses are needed.

In its current version, AUTINDEX works for German and English and also allows for bilingual indexing and classification. Bilingual indexing means, that a German document is indexed and classified using English descriptors, or vice versa.

Step 1: Identification of an initial set of keywords

The first step the system performs is the *linguistic analysis* which consists of a segmentation, a part of speech tagging and a homograph analysis for German texts as described in the previous section. In a second step, multiword terms and their respective syntactic variants (*measurement procedure vs. procedure of measurement vs. procedure to measure*) are identified using the shallow parsing described in section 3.4. The grammar also determines which of these phrases are term candidates.

To calculate the keywords AUTINDEX uses a statistical function based on frequency. However, in contrast to other systems, the weighting is not based on word frequency but on the frequency of semantic classes which are calculated from each word during the linguistic processing: For each major content bearing word (words of category noun, verb or adjective) the semantic features are collected, then. The weighting does not only take complete words into account for this statistics, but also the components of German compounds. For English, compounds which are mostly multiword units the semantic features of each word of the multiword unit are considered.

The most frequent semantic features are then calculated, and all words/phrases of the text which have such a feature are collected, in the so-called *keyword set*. Functional words are excluded from this process.

Step 2: Refinement

The set of keywords is then evaluated against the thesaurus by performing an indexing similar to the concept-based indexing in IR. Thereby thesaurus-specific hierarchical information, e.g. more general or more specific terms and synonyms are used to generate a set of *descriptors*. At this step, the classification is also done, using either the classification scheme provided by the user or the NACE code which is assigned to terms in the general lexicon. For each descriptor, the classification information is collected, and the most frequent descriptors are assigned to the document. This information can then be used for further refinement of the descriptors by discarding those which do not belong to the classification (In Fig.1 these terms are marked in italics under *Descriptors*). This method allows also the filtering of lexical ambiguities in that only those readings of terms are taken that belong to the classification areas determined before.

The **bilingual indexing** is done by applying the steps above to source language documents and using the thesaurus together with a transfer dictionary which contains general language as well as application specific translations to generate the appropriate descriptors and free terms (cf. Fig. 1) for the target language.

AUTINDEX outputs a structured list as shown in the figure below consisting of different sorts of data: the set of descriptors (concepts from the thesaurus), free terms (keywords which are not descriptors), the classification information, and optionally super concepts, countries, and unknown words together with the document itself.

Title:	Spyware-Trojaner rücken zur Hauptgefahr
Text:	Dieser Beitrag erläutert die Geschichte des Computerspiels und geht auf neue Gefahren und Trends ein. Computerviren gibt es bereits seit 1981. Trojanische Pferde sind erst seit der Verbreitung Internetfähiger E-Mail-Systeme weit verbreitet. Doch Office 2000 zum Beispiel ist die Webverbreitung eines Spyware-Trojaner, der Ende 1999 verbreitet wurde und seitdem weltweit nachverbreitet ist. Defensiv ist Office 2000 ist, das sich selbst geschütztes Applikations die Plug-in-Modulen besitzt. Diese interessieren sich auch ein weiteres Tool mit dem Namen Heiken 1.0. Mit diesem Tool kann in Form von Daten auf Daten zugreifen werden, die ursprünglich vor ungewolltem Zugriff geschützt sind. Bei vorzeitigem Implementieren der Antivirenkonzepte verhindert die Effektivität des hochgradig isolierten Spyware-Wörterbuch oder täglich selbst das Betriebssystem ein Arbeitsplatz aktualisiert werden. Der Heiken werden Virenschutzsysteme verstärkt auch auf die Systeme verwendet, um ein Eindringen in die Internet des Unternehmens gar nicht zuzulassen.
Thesaurus:	FTL
Descriptors:	Computerspiele[14] (Computer Spiele), Internet[1] (Internet), Arbeitsschritt[7] (work place), Struktur[14] (structure), Geschichte[1] (History), Gefahr[7] (risk), Verwendung[7] (use), Begriff[7] (concept)
Free Terms:	Antivirenkonzept, Antivirenschutz, Antivirus, E-Mail-System, Fileschuttschutz, Mail-Server, Schutzkonzept, Schutzsystem, Spyware-Trojaner, Virenschutzkonzept, Virenschutzsystem, Virenschutz
Classification:	Forschung, Entwicklung (Research, Development) Software und -werkzeug, virtuelle Datenverarbeitung (Computer software, distributed information processing) Systemarchitektur, Datenbanksysteme, Programmiersprachen Systemwerkzeuge, Operating system, Programming Language)
Super Concepts:	Netz, Platz, Programm, Computer-Komponente, Datensatz, Hochleistungsrechner, Rechenprogramm, Virenschutzsystem, Webanwendungsmöglichkeit, Internet, Handlung, Internet-Hin
Countries:	
Unknown Words:	Heiken, Office

Figure 1: Result of AUTINDEX

The human indexer can now change the list of descriptors if necessary by adding entries from the *Free Terms* set (identifying gaps in the thesaurus at the same time) or delete terms. Within the list of *Unknown Words* not only spelling errors but also new terms (including abbreviations) are listed for which the thesaurus

developer also has to decide whether they should be included in the thesaurus.

Due to its modularity, the system can easily be extended to other languages. MPRO, the basic linguistic processing tool is currently available for 12 languages. For a multilingual processing with another language, either a corresponding bilingual thesaurus is required, or if such a thesaurus is not available, a conversion dictionary which pairs terms of the new language to either German or English terms has to be made. If none of these resources are available a bilingual thesaurus can be generated by using transfer lexicons for general language together with alignment tools applied to parallel or similar documents.

5. A PROTOTYPICAL APPLICATION

The proposed approach has already been put to practice in a small bilateral project between IAI and FIZ Technik, Frankfurt, a bibliographic database producer. 500 German abstracts were automatically indexed using the fiz thesaurus. The results were manually evaluated by comparing the manually indexed terms with the automatically computed indexing terms. The test results showed a good indexing quality, i.e. the ratio of automatically assigned terms to intellectually assigned terms was better than 70%. A document of an average size (1,5K) can be indexed and classified in approx. 25 seconds, compared to 10 to 15 minutes a human indexer needs. This means a substantial cost reduction can be expected by using AUTINDEX. The prototype we have built shows that such a tool can contribute to a better and more effective production cycle and therefore to a better market position for users. In addition, it supports consistency and reduces the translation work for the indexing of foreign language documents.

6. FINDINGS FOR IR

The indexing method is currently integrated into the cross-language information retrieval system MPRO-IR [12] to improve its performance. Indexing using an uncontrolled vocabulary (AUTINDEX's step 1) is already applied. Current experiments investigate how useful the classification as type of conceptual information is. Using the classification information allows the retrieval of documents in which the queried terms do not have to occur. Applying this approach, however, requires that the query can be classified, that at least one term must have classification information. The experiments, using the German documents of the CLEF 2000 corpus, have shown very promising results: The use of the classification information leads to a higher recall and at the same time to a better precision. The results have also shown that the approach works better for long queries than for short ones or single word queries.

7. FUTURE WORK

The indexing and classification approach described so far heavily depends on the quality of the semantic information provided by the lexicon. This refers to completeness but also to the granularity of the semantic classes. The semantic classes so far available are quite coarse-grained and need refinement. Evaluations done within BINDEX have shown that more fine-grained classes allow for a better indexing (more irrelevant terms are discarded) as well as a more precise classification. This is an area where substantial improvements can be achieved.

A second point is that multilingual indexing requires that the monolingual as well as the transfer lexicons have to have the

same quality (concerning semantic information). This has yet to be achieved.

Another direction of ongoing work is the consideration of other information the thesaurus contains such as *Related Terms*. This knowledge can be used to increase the precision of the determination of the classification.

Another step in the ongoing project BINDEX, this paper is based upon, funded by EC (IST-1999-20028) will be the extension and improvement of an existing bilingual English-German component.

8. ACKNOWLEDGEMENTS

We would like to thank our colleague Dieter Maas who contributes to the development of the AUTINDEX system by providing the linguistic processing modules.

9. REFERENCES

- [1] Bakel van, B. Modern Classical Document Indexing. Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, Melbourne.
- [2] Gonzalo, J., Verdejo, F., and Chugur, I. Using EuroWordNet - A Concept-Based Approach to Cross-Language Retrieval. Journal of Applied Intelligence, Vol. 13, Issue 7, 1999.
- [3] Kraaij, W. and Pohlmann, R. UPLIFT - Using Linguistic Knowledge Information Retrieval. Technical Report, OTS-WP-CL-96-001, University of Utrecht, 1996.
- [4] Krovetz, R. Homonymy and Polysemy in Information Retrieval. Proceedings of the 35th Annual meeting of the Association of Computational Linguistics, 1997.
- [5] Landauer, Th. K. and Littman, M. L. A statistical method for language-independent representation of the topical content of text segments. Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, 1990.
- [6] Lovins, J. Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics, Vol 11, 1968.
- [7] Maas, H. D. Multilinguale Textverarbeitung mit Mpro. Lobin, G. et al. (eds.): Europäische Kommunikationskybernetik heute und morgen. KoPäd, München, 1998.
- [8] Maas, H. D. Thesaurus als Wissensbasis für Begriffszerlegungen. Information. Proceedings, Friedrich-Schiller-Universität Jena, 28. bis 30. September 1993.
- [9] Miller, G. A., Beckwith, R., Fellbaum, Ch., Gross, D., and Miller, K. Introduction to WordNet: An On-line Lexical Database. <http://www.cogsci.princeton.edu/wn/papers>, 1993.
- [10] Plaunt, Ch., Norgard, B. A. An Association Based Method for Automatic Indexing with a Controlled Vocabulary, Technical Paper, University of California at Berkeley
- [11] Porter, M. F. An Algorithm for Suffix Stripping. Program 14 (3), 1980.
- [12] Ripplinger, B. MPRO-IR A Cross-Language Information Retrieval Component Enhanced by Linguistic Knowledge. Proceedings of the RIAO 2000, Paris.

[13] Tzoukermann, E., Klavans, J. L., and Jacquemin, Ch. 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. Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, 1997.

[14] Vossen, P. (ed) EuroWordNet - General Document. University of Amsterdam, 1999.