Pergamon

Information Processing & Management Vol. 34, No. 6, pp. 693–707, 1998
© 1998 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0306-4573/98/$19.00

PII: S0306-4573(98)00030-2

PHASE-BASED INFORMATION RETRIEVAL[1]

A. T. ARAMPATZIS,* T. TSORIS[2], C. H. A. KOSTER[3] and TH. P. VAN DER WEIDE[4]
Computing Science Institute, University of Nijmegen, Nijmegen, Netherlands

(Received 15 April 1998; accepted 3 June 1998)

Abstract—In this article we describe a retrieval schema which goes beyond the classical information retrieval keyword hypothesis and takes into account also linguistic variation. Guided by the failures and successes of other state-of-the-art approaches, as well as our own experience with the IRENA system, our approach is based on phrases and incorporates linguistic resources and processors. In this respect, we introduce the phrase retrieval hypothesis to replace the keyword retrieval hypothesis. We suggest a representation of phrases suitable for indexing, and an architecture for such a retrieval system. Syntactical normalization is introduced to improve retrieval effectiveness. Morphological and lexico-semantical normalizations are adjusted to fit in this model. © 1998 Elsevier Science Ltd. All rights reserved

1. INTRODUCTION

Information retrieval (IR) has been developed to give practical solutions to people's need for finding the desired information in large collections of textual data. Although IR has existed for more that three decades, most currently available commercial systems are based on simple assumptions which often lead to unsatisfactory effectiveness.

The assumption (implicitly- or explicitly-made) upon which most commercial Information Retrieval systems are based, is that *if a query and a document have a keyword in common, then the document is about the query to some extent* (naive keyword hypothesis). Of course, if there are *more* keywords in common, then the document is more about the query. In that respect, the IR problem is represented by matching the "bag" of keywords in the user's query with the "bag" of keywords representing the documents. This approach suffers from a number of problems which originate from *linguistic variation*:

1. It does not handle cases where different words are used to represent the same meaning or concept in queries and documents. For this phenomenon we use the term *lexical variation*. The result is that a query keyword "film" does not retrieve documents which contain its synonym "movie".
2. It does not distinguish cases where single words have multiple meanings due to *semantical variation*. A singer looking for "bands" will be faced with "radio frequency bands".

[1] A previous version of this work was presented as a paper at RIAO 97 conference in Quebec, Canada and included in RIAO 97 proceedings (Arampatzis *et al.*, 1997a).
[2] E-mail: theofilus@lch.nl
[3] E-mail: kees@cs.kun.nl
[4] E-mail: tvdw@cs.kun.nl

*Corresponding author. Department of Information Systems, Faculty of Mathematics and Computing Science, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands. Tel.: +31-24-3653147; Fax: +31-24-3553450; E-mail: avgerino@cs.kun.nl.

3. It does not deal sufficiently with the problem of *syntactical variation*. A document saying "near to the river, air pollution is a major problem" is not about "river pollution".
4. To make things worse, keywords can, due to *morphological variation*, appear in different numbers, for instance "wolf" and "wolves", or different cases[5] like "man" and "man's".

All these problems hurt a retrieval system in term of *precision* and *recall*. Given a query, a retrieval system retrieves a set of documents as a response to the query. Precision is defined as the proportion of retrieved relevant documents to the retrieved documents. Recall is the proportion of the retrieved relevant documents to all relevant documents in the system. Dealing with the linguistic variation problems 2 and 3 improves precision. On the other hand, taking into account 1 and 4 improves recall. However, usually improving precision decreases recall and vice versa.

The goal of this article is to define a retrieval schema which takes into account all the problems mentioned above. Our approach is based on *linguistically-motivated phrases* and incorporates linguistic resources and processors. Natural language processing (NLP) has been seen by many research groups, including ours, as a way to improve retrieval effectiveness. The results until now have varied and that suggests the need for a closer look at the subject.

The remainder of this paper is organized as follows. In the next section we give an overview of the most important research in this field. In Section 3 we discuss our contribution to this field, the IRENA system, and the results of a small-scale experiment. In Section 4 we introduce the phrase retrieval hypothesis. In Section 5 we suggest an abstract representation of phrases suitable for indexing. The linguistic normalizations we suggest in Section 6 are expected to improve retrieval performance. Conclusions and future work are summarized in Section 7.

2. RELATED RESEARCH

The use of NLP techniques in IR tasks has become an accepted approach to improve retrieval effectiveness. Most of the current systems use stemming techniques to deal with morphological variants of keywords. One of the most reliable stemmers for English is the revised Porter stemmer (Krovetz, 1993). Query expansion with synonyms and broader/narrower terms or word similarity functions have been tried to overcome lexical variation. Word sense disambiguation tries to deal in a degree with semantical variation. The problem of cooccurrence brings the use of syntactic information into retrieval. It seems obvious that by combining techniques like the ones mentioned the quality of retrieval can be improved.

For an excellent literature review on these subjects we refer the reader to the thesis work of Khoo, 1997. However, we choose to review three studies which we consider representative and close-related to the approach we are going to suggest in this article. These three groups in this line of research have obtained negative or at least dubious results; the IR group at Dublin City University, the CLARIT group and the work of Strzalkowski and Carballo, 1996 in TREC-4.

2.1. The IR work at Dublin City University

The IR group at Dublin City University tried the use of indexing structures derived from syntax. We review the approach and results from their participation in TREC-3, since that was their last attempt to use syntactic phrases.

[5] Trivial for English, but crucial for other more inflected languages like German or Greek.

In this approach, documents and quiries were represented by TSA's (tree structure analytics) constructed at the clause level. These TSA's were directly derivable from a morphosyntactic analysis of input text, and were formulated to encode within their structures the most commonly occurring syntactic ambiguities due to PP (prepositional phrase) attachment, conjunction and others. In case of ambiguity, the TSA matching algorithm weights various (syntactic) interpretations at the time of retrieval. This TSA matching algorithm is able to measure the degree of overlap between input phrases which may or may not have been about the same topic, but which used the same words though sometimes in different contexts. The degree of overlap is inferred from the structure roles different words play in phrases, acting as heads, as modifiers or as attachments.

Smeaton *et al.*, 1995 conducted an experiment on category B of TREC-3 (i.e. on 550 Mb of the Wall Street Journal), and reported failure. The implementation was based on a two-stage retrieval. Firstly, a statistically-based prefetch retrieval ranked the collection. Then the computationally expensive language-based processing was applied to the 1000 top-ranked documents in order to rerank them.

The experimental results were disappointing and unexpected (both recall and precision were decreased). The group posed some possible reasons for the poor results:

- The language analyzer used was of poor quality.
- The type of language used in TREC topic descriptions is very different to that used in document texts (interrogative vs descriptive language), and the two types of language should had been treated differently.
- Maybe the combination of independent retrieval strategies (prefetch using $tf \times IDF$ and TSA-based weighting in this case) would have bootstrapped the performance of individual strategies (this has been shown before by a number of groups in TREC-3 and elsewhere). Maybe the TSA-based retrieval could have retrieved documents not retrieved by the term weighting strategy, especially if those had a few words in common with a query but those words played the same or similar structural roles in query and in document.

The results led the authors to conclude that the approach of using syntax to determine structural relationships between words and to use them as a part of an information retrieval strategy, does not work. Since then, the group has abandoned this strategy and it concentrated on the use of NLP resources (such as machine-readable dictionaries and knowledge bases) to improve retrieval.

2.2. The CLARIT *work*

The CLARIT system (Evans *et al.*, 1993) has several NLP techniques integrated with the vector space retrieval model. These techniques include morphological analysis, robust noun-phrase parsing, and automatic construction of first order thesauri. CLARIT's indexing emphasizes phrase-based indexing with different options for decomposing noun phrases into smaller constituents, including single words.

The goal of the CLARIT TREC-5 NLP track[6] (Zhai *et al.*, 1996) was to test two hypotheses:

1. The use of lexical atoms, such as "hot dog", to replace single words for indexing would increase both precision and recall.
2. The use of syntactic phrases, such as "junior college" to supplement single words would increase precision without hurting recall and using more such phrases results in greater improvement in precision.

[6] The latest literature we found available about the participation of CLARIT in TREC NLP track was from TREC-5, since the proceedings of TREC-6 have not been published by the time of this study.

For the first hypothesis, as lexical atoms were considered the high frequency word pairs that tended not to be separated by other words within the context of noun phrases. The only pairs considered were formed by two nouns or one adjective followed by a noun. From both TREC-5 and the preliminary experiments with TREC-4 topics, it was shown that the use of lexical atoms leads to a slight but consistent improvement in average precision. On the other hand, using lexical atoms did not consistently improve recall and initial precision. In fact, it increased either recall or the initial precision. The inconsistent influence of lexical atoms may indicate a need for a better control over the selection of phrases that are used for replacing single words.

For the second hypothesis, syntactic phrases were obtained from noun phrases. The noun phrase parser used an expectation maximization algorithm to obtain statistical evidence of word modifications from the noun phrases in the corpus (Zhai, 1997). In simple words, they applied statistical methods to assign structure to those noun phrases which had an ambiguous structure (all noun phrases of more than two words). The three automatic official runs of the experiment corresponded to the following three levels of term combinations: (a) single word only, (b) single word + head modifier pair + full NP, and (c) single word + head modifier pair + adjacent subphrase + full NP. These experiments in supplementing single words by various combination of syntactic phrases in the indexing process showed a consistent and significant improvement in retrieval performance. However, the impact of adding phrases into the index space varied according to the query topic. Thus, while adding phrases helped some topics it hurt some others.

2.3. Natural language information retrieval: TREC-4 report

The approach of Strzalkowski and Carballo, 1996 in TREC-4 was successful. They built an NLP module around a statistical full-text indexing and search backbone. The NLP module was used to (a) extract content-carrying phrases from documents, and (b) process user's natural language requests into effective search queries.

All TREC-4 texts were processed with a syntactic parser. Phrases were extracted from the parse trees and used as compound indexing terms in addition to single keywords. They also, like the CLARIT group in the previous section, applied statistical methods to resolve structural ambiguity. These phrases were head-modifier pairs.

The user's natural language request was also parsed to identify indexing terms. Highly ambiguous, usually single-word terms were dropped, provided that they also occurred in compound terms. They also used similarity relations (synonymy, hypernymy, hyponymy, etc.) to add other terms. For example, "unlawful activity" is added to a query containing the compound term "illegal activity" via a synonymy link between "illegal" and "unlawful".

Two types of morphological normalization were performed: (a) inflected word-forms were reduced to their root forms as specified in the dictionary and (b) nominalized verb forms were converted to the root forms of corresponding verbs (e.g. "implementation" was converted to "implement").

Their experiments showed a substantial improvement in precision when phrasal terms are used. Especially, it was achieved a sharp increase of precision near the top of the ranking, which bring further gains in performance via automatic feedback. They also cautiously suggest that NLP can be effective in creating appropriate queries out of user's natural language request which can be frequently imprecise or vague. The benefit, however, from linguistic processing was tied to the length of the query: the longer the query, the larger the improvement.

In their subsequent participations in TREC elaborated further the techniques described here, but NLP has not been proven yet as effective as they would have hoped to obtain better indexing and better term representation of queries. Using linguistic terms still does help to improve precision, however, the gains remain quite modest (Strzalkowski *et al*., 1997).

The inconsistent results in combining NLP and IR make difficult to reach a conclusive statement. The observed inconsistency in these three studies can be assigned to:

1. the choice and representation of indexing terms,
2. insufficient dealing with the problems of linguistic variation, and
3. the quality of NLP.

These suggest further investigation and better modeling. In the next section and before we elaborate further, we discuss our own small experience in this field, the IRENA system, and the results of a small-scale experiment.

3. THE IRENA SYSTEM

The experimental IRENA (information retrieval engine based on natural language analysis) system was built at the University of Nijmegen, The Netherlands. It was developed to study the influence of NLP techniques on precision and recall in document retrieval systems by means of NLP techniques. The NLP component dealt with the morphological and lexical part of the English language to improve recall, and with syntax to improve precision. The retrieval approach taken was based on noun phrases. We make here a short review of the approach. For extended details, the reader should refer to Arampatzis *et al*., 1997a and Arampatzis and Tsoris, 1996.

3.1. Noun phrase as a unit of cooccurrence

An ideal retrieval strategy would be based on some measure of the "nearness" of one noun phrase (in the query) to another (in the document). Although similar measures had been developed (e.g. *logical nearness* in Bruza, 1993; Bruza and IJdens, 1994), we investigated in IRENA other, more heuristic strategies that fitted for our *noun phrase cooccurrence hypothesis*.

Our basic premise was that words occurring in the same noun phrase (NP) tend to share some semantical relation. If two or more nouns and their respective adjectives had found in a single NP, then we assumed that these nouns shared some relatedness, even without knowing what they stand for. For example in the phrase

> ...tracks were recorded *at the BBC studios for later radio programs...*
>
> the nouns "radio", "programs" and the proper name "BBC" which reside in the same underlined NP[7] of the sentence are semantically related. Therefore, searching for the programs of the BBC radio station with the query "radio programs on BBC", will retrieve documents containing phrases like the one above and not documents with other forms of cooccurrence like:

Document 1: *the transmission of his first* **radio programs** *resembled the early years of the creation of* **BBC** *empire which...*

Document 2: *ten musicians from the* **BBC** *Symphony Orchestra* were interviewed *in several* **radio programs** *of L.A. stations...*

These phrases are rejected due to the syntactic information that the three words of the query do not all reside in the same NP. The last real cases clearly showed that extra

[7] Syntactically, the prepositional phrase (PP) "for later radio programs" belongs to the verb phrase in this example. The problem here is that a simple-minded parser cannot resolve this PP-attachment. We used an NP-parser only which was not able to parse verb phrases. In this respect, we went for the *longest possible NP* resolving structural ambiguity by means of syntactic under-specification.

linguistic processing is superior compared to a *proximity search* that requires words in the user's query to be just close to each other in the document.

Additionally, we experimented with the improvement of recall by expanding queries with morphological (only inflectional) variants of the keywords (singular, plural and genitives), and lexical variants of the keywords (only synonyms). The synonyms were obtained from WORDNET (Miller, 1995). We did not try to disambiguate word senses to select the right synonym-set in WORDNET. We merely used all the synonym-sets which a keyword belongs to (all possible synonyms in any context).

For the syntactical analysis needed we used the AGFL parser generator system (Derksen, 1997) to produce a parser for the English NP. The parser syntactically analyzed the NP queries to extract adjectives and nouns. On the side of documents the parser checked for cooccurrence of keywords or their variants in the same NP.

Of course we expected some exceptions which do not conform to the NP cooccurrence hypothesis. We encountered some phrases during the execution of the experiment where the terms of a query occurred in one single NP of a document's sentence, but were not semantically related. An example:

Query: soundtracks of films

Text: ...In this album, there is a good background, but there is something missing. Either a solo voice or instrument. Or at least a film. Soundtrack without pictures so to speak....

Searching for film soundtracks in general we came across this text. The noun "picture" is a synonym of the noun "film" and belongs to the same NP as "soundtrack". But the meaning of the last sentence is merely that this album could be a soundtrack of a movie but it was not. Notice that the prepositions (of/without) were not taken into account.

3.2. The experiment

We conducted a small experiment using a manually collected corpus of 6.7 Mb of music texts (e.g. magazine articles, FAQ's about artists, interviews, reviews, etc.).

Forty-four NP queries were submitted to the system and an average of 2.6 keywords per query were extracted. The expansion with synonyms resulted in 4.1 more keywords, that is about 1.6 synonyms per initial keyword. The morphological expansion of all keywords and their synonyms added an average of 14.2 more search words. Consequently, for every query, on average 20.9 keywords were submitted to the retrieval subsystem. The precision-recall results are summarized in Table 1 and graphically in Fig. 1. We were confronted here with the classic problem of calculating recall. So we defined a measure called *relative recall*, and wherever we refer to recall in this experiment we mean relative recall. Arampatzis *et al.*, 1997a have proved that if *generality*(q) is approximately constant for all the submitted queries q, then relative recall is close to the real recall. The results are also normalized for readability: we

Table 1. Precision-recall results

Window	K		KM		KSM	
	precision (%)	recall (%)	precision (%)	recall (%)	precision (%)	recall (%)
NP	100.00	6.31	95.65	23.14	91.38	27.90
2	79.17	19.98	76.45	49.44	71.61	58.41
3	74.60	24.75	70.62	58.41	66.33	69.43
4	72.97	28.40	68.05	66.83	63.40	78.40
6	70.21	34.71	65.22	76.79	59.59	91.58
8	69.44	39.48	63.87	84.22	56.38	100.00

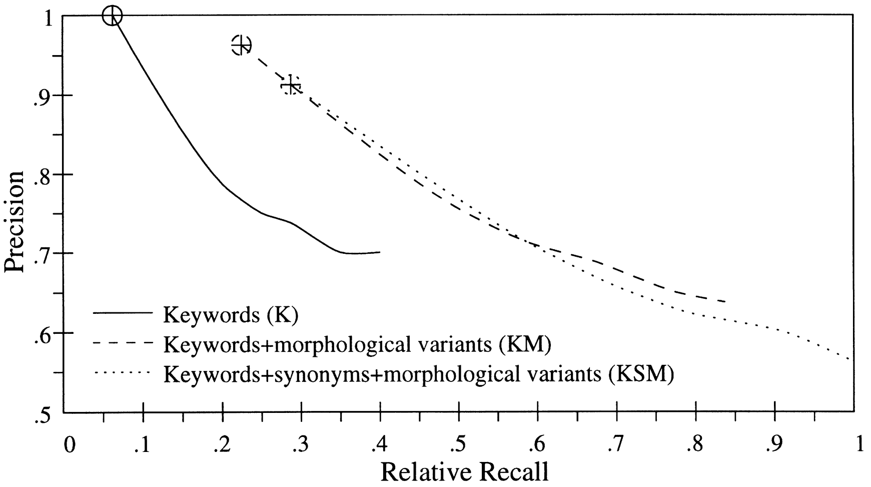

Fig. 1. Precision-recall graphs.

assumed that the configuration of the system which retrieved the most documents had achieved 100% recall. We compared 3 different kinds of searches, these were, K, KM and KSM.

K means keywords only, KM keywords *and* their morphological variants and KSM keywords *and* all their synonyms *and* morphological variants of all the previous. We restricted the retrieved set to these documents in which all query keywords or any kind of their variants cooccurred. The ranking was based on the size of cooccurrence window. Thus, documents which presented term cooccurrence in an NP were ranked higher, followed by the documents with term cooccurrence within 2 text lines, then 3 text lines, and so on. Test runs showed that using this kind of ranking precision decreased and recall increased by going down rank positions. In Fig. 1, the points which are included in circles represent precision and recall for NP cooccurrence.

The use of NP cooccurrence resulted in very high precision levels, above 90%, in all search types especially in keyword search; in this case, precision was even 100%, possibly because of a tendency of the user and formulator of the queries to think primarily in terms of keywords, rather than phrases. Conversely, recall with NP cooccurrence was extremely low compared to simple cooccurrence in a text window. As window size increases from 2 to 8 lines, it seems that better recall is gained at the price of a slight drop in precision. However, it is found that by increasing the size to more than 16 lines, precision is dramatically lowered to 25–35%. Upon enlarging the window, keywords may appear in different paragraphs with possibly different subjects, which accounts for this large drop in precision. A window size of 4 to 8 lines gives reasonable levels of precision and recall.

Expanding queries with lexical and morphological variants led to a marked increase in recall, up to 60% (but the query set cannot be considered as representative). The decrease of precision, which in the worst case was 13%, can be considered as insubstantial compared to the recall gained.

The small-scale experiment with IRENA has (again) proved that lexical and morphological expansion of queries is indispensable for high recall and results in an insubstantial average loss of precision, hence is highly recommended. This holds in spite of the fact that the natural language used was English, which is weak in morphology and poor in syntax. Experiment has to show whether this also holds for highly inflected languages like e.g. modern Greek.

The NP cooccurrence criterion has proved to be successful in determining whether keywords are semantically related and achieves a much better precision than *proximity search*. The low recall obtained suggests the generalization of the NP cooccurrence hypothesis to wider classes of phrases to delimit the semantic relatedness between words (verbal phrases, anaphora). At any rate, the NP cooccurrence criterion can also be used in the future for relevance feedback.

The dramatically low recall achieved could be interpreted in two different ways: One could argue (Gay and Croft, 1990; Smeaton, 1997) that use of the noun phrase shows no promise in improving the performance of IR systems. We argue, on the other hand, that we should retain the noun phrase as a unit of cooccurrence, but should investigate the possibilities of enhancing the recall without loosing too much precision, by taking into account linguistic variation and anaphora.

In the next sections we formulate the phrase hypothesis for retrieval and introduce *syntactical normalization* to deal with the rich choice of alternative syntactic formulation for one same phrase, such as: *air pollution*, *polluted air*, *pollution of the air*, *air is polluted*, etc. The query expansion mechanisms we employed in IRENA are replaced by *morphological* and *lexico-semantical normalization* techniques, which are more general.

4. THE PHRASE HYPOTHESIS

The evidence suggests that noun phrases should be considered as a semantical unit, rather than text windows. In this respect, we should focus on the noun phrase as the *unit of retrieval* instead of the single word. The most important reasons are:

- noun phrases play a central role in the syntactic description of *all* natural languages, functioning as subject, object and in preposition phrases.
- In artificial intelligence, noun phrases are considered as references to (or descriptions of) complicated concepts (Winograd, 1983). By others, as *picture producers*.

Consequently, a characterization using noun phrases captures more of the conceptual content of a document.

Although noun phrases are good approximations of concepts, all phrases corresponding to concepts, but not being noun phrases are missed. This observation points to the necessity to consider other phrases than only noun phrases for retrieval. We are prompted to take linguistically meaningful phrases as retrieval *terms*: the noun phrase including its modifiers (for the reasons mentioned above), and the verb phrase including its subject and other complements. The verb phrase describes a situation or process by relating a main verb to a number of NP's and other phrases, thus it is a good extension beyond NP's. Our *naive* retrieval hypothesis is formalized as follows.

Definition 4.1 (Naive phrase hypothesis). If a document and a query have a phrase in common, then the document is about the query.

We use this definition as a naive starting point, upon which we will build our framework. It is an evolution of the keyword hypothesis (see Section 1). Although the keyword hypothesis is also naive, most of the current systems are based on it and try to improve retrieval effectiveness by applying other mechanisms (e.g. keyword stemming, relevance feedback, etc.) over it. This is how we are going to proceed as well. First, an abstract representation of phrases which is suitable for indexing is needed.

5. REPRESENTATION OF PHRASES

Most of the approaches which use phrases for retrieval do not work with complete sentence grammars but with some form of phrase pickers. The phrases are extracted by using part-of-speech taggers. *Tagging* takes into account the syntactic context and can be seen as a weaker form of parsing, thus the resulting representation is a sequence of words which may form e.g. an NP. The NP cooccurrence approach, considering the NP as a set of words, may give too little detail. For example,

> the hillary clinton health care bill proposal would contain "bill clinton", but it is obvious that this NP does not refer to him.

Parsing takes into account both syntactic context and structure, but noun phrase representations as parse-trees result in too much detail, at least for indexing. Nevertheless, Hull *et al.*, 1996 have shown that linguistically motivated light parsing can slightly improve retrieval results over the classic IR approximation to noun phrase recognition.

As a starting point in this research, we suggest an intermediate representation of noun and verb phrases, eliminating elements and structures which are assumed not relevant to IR, so as to be more suitable for indexing.

Definition 5.1 (Noun phrase). A core noun phrase NP, from an abstract point of view, has the general form:

$$\text{NP} = \text{det pre* head post*},$$

where *det* (determiner) is the article, quantor, number, etc., *pre* (premodifier) the adjective, noun or coordinated phrase, *head* usually a noun, *post* (postmodifier) the prepositional phrase, relative clause, etc., and the asterisk (*) denotes zero or more occurrences.

Determiners are of little interest from an IR point of view, and therefore are eliminated. Pre- and postmodifiers may recursively include other NP's. Relative clauses are also dropped.

Definition 5.2 (Verb phrase). A verb phrase VP, from an abstract point of view, has the general form:

$$\text{VP} = \text{subj kernel comp*},$$

where *subj* (subject) is an NP (in the wide sense, including personal names, personal pronouns etc.), *kernel* (verbal clause) the inflected form of some verb, possibly composed with other auxiliary verb-forms and adverbs, *comp* (complements like object, indirect object, preposition complement, etc.) an VP or prepositional phrase (PP) and the asterisk (*) denotes zero or more occurrences, depending on the transitivity of the verb (e.g. intransitive verbs have no complements, transitive verbs have an object, ditransitive have an object and indirect object).

Adverbs are considered redundant, thus are eliminated.

The phrases, as defined in Definitions 5.1 and 5.2, could be used in their literal form as units of retrieval, although the performance is then expected to be inferior than that of keywords. It is well known that, as the size of corpus grows, the number of keywords grows with the square root of the size of corpus. One could expect that the same holds for phrases, but the number of such enriched terms grows even faster. So does the likelihood of there being different phrases corresponding to the same concept. On one hand we would like to use phrases to achieve precision, but on the other hand recall will be too low, because the probability of a phrase reoccurring literally is too low. To achieve also recall we shall introduce in the next sections a number of linguistic normalizations and some forms of fuzzy matching of phrases.

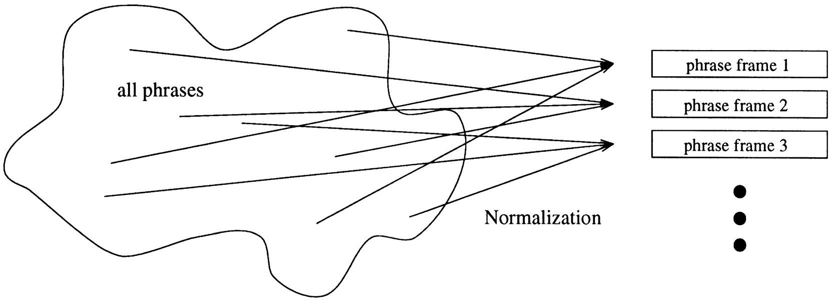

Fig. 2. Normalization.

6. LINGUISTIC NORMALIZATION

The goal of normalization is to map different but semantically equivalent phrases onto the same phrase (Fig. 2). This will be done as far as possible without deep semantic analysis. The normalization is applied in three steps:

1. *syntactical normalization* by means of rearrangement of elements,
2. *morphological normalization* by means of lemmatization, and
3. *lexico-semantical normalization* by means of fuzzy matching.

The resulting representation after step 1 is called *phrase frame* (PF).

To simplify structural matching of phrases, we are going to decompose PF's into *binary terms* (BT's) (Section 6.3). The overall architecture of our phrase-based retrieval system is given in Fig. 3. The lexico-semantical normalization can be incorporated in the matching function (fuzzy matching), or it can been seen as a separate process.

6.1. Syntactical normalization

Syntactical normalization is achieved by flattening the syntactic structure, and it is highly language dependent. According to the linguistic principle of *headedness*, any phrase has a single word as a head. This head is usually a noun (the last noun before the postmodifiers) in NP's, the main verb in the case of VP's. The rest of the phrase consists of modifiers.

The resulting representation after syntactical normalization is a phrase frame (*PF*), basically a head–modifier (*h–m*) pair.

$$\mathrm{PF} = [h, m]$$

The head h gives the central concept of the phrase and the modifiers m serve to make it more precise. Conversely, the head may be used as an abstraction of the phrase, loosing precision but gaining recall. Modifiers in the form of phrases are recursively defined as phrase frames: $[h_1, [h_2, m]]$. The modifier part might be empty in case of a bare head. This case is denoted by $[h]$. The head may serve as an index for a list of phrases with occurrence frequencies:

```
engineering 1026
, of software 7
, reverse 102
, software 842
, ...
```

The frequency of a bare head will include that of its modified occurrences.

These head-modifier pairs are produced by applying normalizations. At this point, only the obvious and easily expressed normalizations should be attempted. Restricting ourselves to English, the following normalizations are possible.

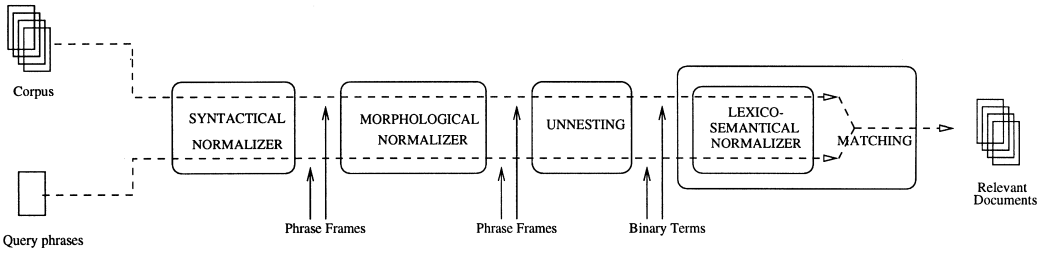

Fig. 3. A phrase-based retrieval architecture.

6.2. *Noun phrase*

1. *Mapping of syntactical variations*. For example, "air pollution" and "pollution of the air" have the same meaning and therefore one should be mapped on the same PF. Obviously there are two possibilities:

   a. mapping a postmodifier with "of" onto a premodifier. This, after eliminating "the", would map both NP's onto "air pollution".
   b. mapping the premodifying noun to a postmodifier. This means that the missing preposition must be deduced. To this end, information about the *subcategorization of the noun* is needed. Of course it is possible to take "of" in all case, but is unfair to a noun phrase like "software construction conference" which should be turned into

      conference on construction of software

      rather than

      conference of construction of software

2. Mapping of noun phrases to verb phrases. Nominalized verbs can be mapped to root forms of verbs. For instance, "implementation" can be turned to "implement".

To take up the last examples, and assuming option (a), "conference on construction of software" and "pollution of the air" will become

`[conference, on software construction]` and `[pollution, air]`

in PF representation. Converting further the modifier in the first PF yields

`[conference, on [construction, software]]`

The preposition "on" optionally can be kept for further semantic analysis, although its use is currently dropped for simplicity in our schema. Deep semantic analysis is beyond the scope of our retrieval schema, but it must be noted that "the spaceman *on* the ship" has a different view than "the spaceman *outside* the ship" and "the spaceman *without* ship" is probably not even in space. In this respect, prepositions should be considered at one point. The use and meaning of prepositions can always be postponed until the matching of PF's.

To conclude, the noun phrase presents only few opportunities for syntactical normalizations.

6.3. *Verb phrase*

For the VP, many more normalizations can be found which preserve the meaning, or rather do not loose information which is obviously relevant for retrieval purposes:

1. *Word-order normalization*. The elements of each VP must be brought in some standard form like,

   kernel subj [comp]*

   including some specific ordering of the complements.
2. *Mapping to active form*. Passive VP's may be brought to active forms.
3. *Elimination of time and modality*. The kernel can be reduced to the infinitive. It may make sense though to retain an indication of certain modalities (present or future, wish or fact).
4. *Mapping of verb phrases to noun phrases* (nominalization). There are cases where it is possible to map a VP to a semantically equivalent NP. Predicative adjectives can be brought together with their corresponding subject, e.g. the predicative sentence

   "the apple is red"

   can be turned into the NP

   "red apple"

   Verbs can also be turned to corresponding nouns (nominalization). For instance, the verb "implement" could be mapped to "implementation". Since the opposite transformation is also possible for nominalized verbs (as mentioned before in Section 6.1.1), a choice has to be made on the basis of experimentation.

These normalizations are rather language dependent and the final decision of what has to be included in the PPs is left to the linguists and system designers. For example, in the phrase

"spacemen travelling to moon with spacecraft"

combining the verb with the object and the preposition complement as head-modifier pairs may produce frames like

`[travel, to moon]`, `[travel, with spacecraft]`

However, the verb should be combined with the subject in reverse order

`[spaceman, travel]`,

since it seems more meaningful than `[travel, spaceman]`.

6.4. Morphological normalization

Until now, in keyword-based systems, morphological normalization has been mostly done by means of *stemming*, which without considerable aid from a lexicon may reduce a word to a different word (executive/execute) or even to a nonword (police/polic). Stemming is rather ineffective when applied to more inflected languages than English because of the ambiguities it introduces. Rather than using stemming we will use *lemmatization*. Every inflected word-form will be replaced by its dictionary entry. Verb forms are going to be reduced to the infinitive and all inflected forms of nouns to the nominative singular. This normalization will not be done by a *tagger* or *stemmer*, but it will be guided by the syntax analysis, taking into account the context of a word. After all, the best tagger is the grammar.

6.5. Unnesting

In order to simplify the structural matching of phrases, and also to raise recall, we follow the strategy of unnesting all complicated phrase frames. A composed term like [*a*, [*b*, *c*]] will be decomposed into two frames [*b*, *c*] and [*a*, *b*] using *b* as an abstraction for [*b*, *c*]. When this decomposition is applied recursively, it results in *binary terms* (BT's). For example, the phrase

"man visited conference on software engineering"

will give the frame

`[visit, [conference, on [engineering, software]]]`

which is further unnested to

`{[visit, conference], [conference, on engineering], [software, engineering]}`

Of course it is important that the parser should be able to deduce the right dependency structure in complicated phrases.

6.6. Lexico-semantical normalization

This kind of normalization depends on the observation that certain relations can be found between the meaning of individual words. The most well-known of those *lexico-semantical* relations are:

- *synonymy* and *antonymy*,
- *hyponymy* and *hypernymy* (the *is-a* relation),
- *meronymy* and *holonymy* (the *part-of* relation).

Even simple relations like synonymy have been proven quite effective (Arampatzis *et al.*, 1997a).

Two important aspects which must be considered for this kind of normalization are *polysemy* and *collocations*. A word is polysemous if its meaning depends on the context. All the terms to which a certain word can lead by using the above-mentioned relations, are actually dependent on the initial meaning of the word. For example, "note" can be meant as a being a short letter, or as a musical note. Using the synonymy relation for the first meaning we can obtain "brief", while "tune" is obtained in the second case. This suggests the use of a *word-sense disambiguator* which takes into account the conceptual context of a word.

Collocations are two or more words which usually appear together, e.g. "health care" and have a certain meaning. These collocations in our approach are considered as single units, since we are more interested in "health care" than "health" and "care" separately. This is also useful in retrieval from a semantic point of view. For instance, when using WORDNET in expanding a query with hypernyms, the notion "*health care*" obtains "*social insurance*" which cannot be obtained in any case by expanding the two separate words. Furthermore, some cases of part-of-speech and structural ambiguity can be resolved by parsing collocations together. In this respect, where we refer to nouns in this article, we also mean noun collocations. In previous approaches in retrieval, these have been obtained using statistical methods dependent on their frequency of use in a certain corpus (Zhai *et al.*, 1996 refer to them as "lexical atoms"). Another method, which fits a domain independent approach is to use, as a starting point, an online thesaurus like WORDNET (Miller, 1995) in combination with statistical methods.

Three possibilities can been seen for lexico-semantical normalization:

1. *Semantical clustering* in analogy with stemming. For instance, several synonyms in a context are reduced to one *word cluster*. This approach is rather aggressive and suffers from the same drawbacks as stemming the documents. For example two "synonyms" are always overlapping in meaning and they do not actually mean the same. The convention to call them "synonyms" depends on the degree of overlap.

2. *Semantical expansion*, extending a term with all its synonyms, hypernyms, hyponyms, antonyms, meronyms and holonyms. The derived terms must be weighted according to their relation with the initial term. Antonyms must be supplied in a NOT-fashion.
3. *Incorporation of a semantical similarity function into the retrieval function (Fuzzy Matching)*. Based on a semantical *taxonomy*, an *ontology* or a *semantical network* we can define a *semantical similarity function* for binary terms as

$$\text{similarity: } \mathrm{BT} \times \mathrm{BT} \longmapsto [0, 1]$$
$$\text{similarity } ([\mathrm{h}_1, \mathrm{m}_1], [\mathrm{h}_2, \mathrm{m}_2]) = \mathrm{sim}(\mathrm{h}_1, \mathrm{h}_2)\mathrm{sim}(\mathrm{m}_1, \mathrm{m}_2)$$

As an example, using the relations *SYN*onymy and *HYP*ernymy for two words (or collocations) x and y, one could try:

$$\mathrm{sim}(x, y) = \begin{cases} 1 & x = y \\ 0.9 & y \in \mathrm{SYN}(x) \\ 0.5 & y \in \mathrm{HYP}(x) \text{ or } x \in \mathrm{HYP}(y) \\ 0 & \text{otherwise} \end{cases}$$

Since the similarity of heads is more important than this of modifiers, the above *sim* factors must be further weighted.

7. CONCLUSIONS AND FUTURE WORK

Our research is concerned with the improvement of the effectiveness of retrieval and filtering systems by using phrases. In this article, we have described a retrieval model based on phrases. At this point in time, there is no conclusive evidence that NLP-support can provide an important contribution to the quality of IR. It is our intention to contribute to this goal by increasing the level of linguistic information used. In particular, apart from the noun phrase, we wish to include the information present in the verb phrase and its complements, in combination with syntactic normalization and fuzzy matching.

Experimentation with phrases will help in achieving these goals. This requires the use of natural language resources and tools. Effective and efficient natural language analysis of large corpora requires highly advanced parsing techniques; a problem which is beyond the scope of this article. Arampatzis *et al.*, 1997b describe syntactic analysis techniques specially adjusted for text filtering and elaborate also robustness and ambiguity issues.

The issue of phrase representation in Section 5 and the linguistic normalization techniques in Section 6 are important research problems which definitely deserve careful studies and experimental evidence. They are presently being investigated in two projects:

1. the information filtering project PROFILE[8], at the University of Nijmegen. The phrase hypothesis is applied to filtering, and the effect of syntactic normalization is being investigated in the context of English documents.
3. the DoRo project (ESPRIT HPC), which aims at the development of a system for the automatic classification and routing of full-text documents.

Our current task is to experimentally evaluate these techniques, by performing a well-defined set of experiments on a document classification system as a testbed. In articles we will inform you of the results.

[8] For more information see: http://hwr.nici.kun.nl/ ~ profile.

REFERENCES

Arampatzis, A. T. & Tsoris, T. (1996). A Linguistic Approach to Information Retrieval. M.Sc. thesis, Patras University, Patras.

Arampatzis, A. T., Tsoris, T. & Koster, C. H. A. (1997a). IRENA: information retrieval engine based on natural language analysis. In *Proceedings of RIAO'7 Computer-Assisted Information Searching on Internet*, pp. 159–175. McGill University, Montreal.

Arampatzis, A. T., van der Weide, T. P., van Bommel, P. & Koster, C. H. A. (1997b). Syntactical Analysis for Text Filtering. Technical Report CSI-R9721, Computing Science Institute, University of Nijmegen, Nijmegen.

Bruza, P. D. (1993). *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia*. Ph.D. thesis, University of Nijmegen, Nijmegen.

Bruza, P. D. and IJdens, J. J. (1994) Efficient Probabilistic Inference through Index Expression Belief Networks. In *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence (AI94)*, pp. 592–599. World Scientific.

Derksen, C. F. (1997). Manual for the AGFL system — the GEN parser generator version 1.6. http://www.cs.kun.nl/agfl/.

Evans, D. A., Lefferts, R. G., Grefenstette, G., Handerson, S. H., Hersh, W. R. & Archbold, A. A. (1993). CLARIT TREC design, experiments and results. In *The First Text Retrieval Conference (TREC-1)*, ed. D. K. Harman, pp. 251–286; 494–501. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-207, Washington, DC.

Gay, L. S., & Croft, W. B. (1990). Interpreting nominal compounds for information retrieval. *Information Processing and Management, 26*(1), 21–38.

Hull, D. A., Grefenstette, G., Schulze, B. M., Gaussier, E., Schutze, H. & Pedersen, J. O. (1996). Xerox TREC-5 site report: routing, filtering, NLP and Spanish tracks. In *The Fifth Text Retrieval Conference (TREC-5)*, ed. D. K. Harman and E. M. Voorhees. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-238, Gaithersburg, MD.

Khoo, C. S.-G. (1997). The use of relation matching in information retrieval. *LIBRES: Library and Information Science Research, 7*(2), 0.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ed. R. Korfhage, E. Rasmussen and P. Willett, pp. 191–202. ACM Press, Pittsburgh, PA.

Miller, G. A. (1995). WORDNET: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Smeaton, A. F. (1997). Using NLP or NLP resources for information retrieval tasks. In *Natural Language Information Retrieval*, ed. T. Strzalkowski. Kluwer Academic Publishers.

Smeaton, A. F., O'Donnell, R. & Kelledy, F. (1995). Indexing structures derived from syntax in TREC-3: system description. In *The Third Text Retrieval Conference (TREC-3)*, ed. D. K. Harman, pp. 55–67. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-225, Gaithersburg, MD.

Strzalkowski, T. & Carballo, J. P. (1996). Natural language information retrieval: TREC4 Report. In *The Fourth Text Retrieval Conference (TREC-11)*, ed. D. K. Harman, pp. 245–258. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-236, Gaithersburg, MD.

Strzalkowski, T., Lin, F. & Carballo, J. P. (1997). Natural language information retrieval: TREC-6 report. In *The Sixth Text Retrieval Conference (TREC-6)*, Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication, Gaithersburg, MD (to appear).

Winograd, W. (1983). *Language as a Cognitive Process*. Addison-Wesley, Reading, MA.

Zhai, C. (1997) Fast Statistical Parsing of Noun Phrases for Document Indexing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC (to appear).

Zhai, C., Tong, X., Frayling, N. M. & Evans, D. A. (1996). Evaluation of syntactic phrase indexing — CLARIT NLP track report. In The *Fifth Text Retrieval Conference (TREC-5)*, ed. D. K Harman and E. M. Voorhees. Department of Commerce, National Institute of Standards and Technology (NIST) Special Publication 500-238, Gaithersburg, MD.

PERGAMON

Information Processing and Management 35 (1999) 587

INFORMATION PROCESSING & MANAGEMENT

Erratum

Phrase-based information retrieval, A.T. Arampatzis, T. Tsoris, C.H.A. Koster and Th.P. Van der Weide. *Information Processing & Management*, 34, 693–707 (1998)[☆]

The publisher regrets that an error occurred in the title of the above paper, the correct and definitive title for the paper is "Phrase-based information retrieval" and not "Phase-based information retrieval" as it appeared in the published issue.

[☆] PII of original article: S0306-4573(98)00030-2

PII: S0306-4573(99)00014-X