Textual Context Analysis for Information Retrieval

Mark A. Stairmand
Centre for Computational Linguistics,
UMIST,
Manchester M60 1QD, UK
(marks@ccl.umist.ac.uk)

Abstract

We describe four applications of QUESCOT, a program which analyses and quantifies textual contexts in documents with reference to the WordNet database, and hence ascertains the dominance of topics in a document. Our analysis is based on previous work in lexical cohesion, a feature of texts which contributes to their functioning as a coherent unit. The applications are diverse, but all pertain to information retrieval. Whilst our results suggest that QUESCOT is not well suited to word sense disambiguation and text segmentation, our experimental IR system using QUESCOT as an indexing component produces promising results. We also used QUESCOT representations to automatically generate a resource to supplement WordNet, based on collocational relations between concepts in a document collection. We conclude that QUESCOT is suited to applications based on document-level descriptions, where the degree of granularity allows inaccuracies to be smoothed out.

1. Introduction

In this paper we report on experiments in the domain of information retrieval (IR) exploiting the output of a document analyser which identifies and quantifies distinct textual contexts. We interpret a textual context as a sequence of semantically related items of vocabulary reflecting the exposition of a particular topic, and creating an environment in which further items are interpreted; the term is synonymous with what other researchers have called a 'lexical environment' (e.g. [14]). The basis for such an analysis derives from Halliday and Hasan's category of lexical cohesion [8], and Morris and Hirst's proposal for a textual analysis based on this phenomenon [14].

SIGIR 97 Philadelphia PA, USA
Copyright 1997 ACM 0-89791-836-3/97/7..$3.50

Although the underlying idea is straightforward, it is only relatively recently that large-scale resources suitable for this analysis have become widely available. We use the WordNet lexical database [13] to determine semantic relations in our analysis, and linguistic processing does not extend beyond part of speech tagging and morphological analysis. The program we have produced, QUESCOT[1], produces document representations encapsulating the results of the analysis we describe.

The applications we describe in Section 3 are diverse, but each has the potential to improve the functionality of IR systems. In Sections 3.1 and 3.2 we investigate the suitability of our analysis to the tasks of word sense disambiguation and text segmentation; these applications are both of relevance to IR, as explained below, and are suggested by Morris and Hirst [14] as possible applications for their analysis but not evaluated. More promising results are reported in Section 3.3, where we overview and evaluate COATER, an experimental IR system we have developed which uses QUESCOT as an automatic indexing component. For the final application, described in Section 3.4, we describe how the information held in QUESCOT document representations can be exploited to automatically discover sense-tagged collocations, resulting in a resource which can be exploited for IR purposes. The emphasis of this paper is thus on specific IR based applications, rather than the QUESCOT analysis itself.

2. Overview of QUESCOT

QUESCOT is a program which aims to determine which are the dominant topics in a document by considering the relations between the vocabulary used throughout. This analysis assumes a correspondence between textual context and topic, since the exposition of distinct topics in a document will tend to be evidenced through the use of semantically related terms. Such an analysis has an intuitive appeal for IR. Our notion of context accords with that of Kozima and Ito [11], who state that a context can be specified by a word set consisting of keywords of the context. Having identified clusters of semantically related terms using WordNet, we quantify textual contexts by considering the

[1] QUESCOT: Quantification and Encapsulation of Semantic Context

distribution of these terms throughout the document. The phenomenon of lexical cohesion, which arises from the selection of vocabulary items and the semantic relations between them, described in detail by Halliday and Hasan [8], provides a linguistic basis for the analysis we undertake; it is one of several categories of cohesion which help a text to stick together and function as a coherent unit. Our analysis was motivated by the lexical cohesion analysis based on lexical chains proposed by Morris and Hirst [14]; a lexical chain is a sequence of semantically related words spanning a topical unit of a text. Morris and Hirst claim that lexical chains correspond to the intentional structure of a text, a component of Grosz and Sidner's account of discourse structure [7]. Although our analysis appeals to the notion of a lexical chain as a sequence of semantically related terms, we do not assume a direct correspondence between a chain and another entity.

Central to QUESCOT's analysis is the WordNet *synonym set*. A synonym set, or synset, represents a concept and comprises all those terms which can be used to express the concept. Synsets have unique identification codes, as the following exemplifies:

Synset Id	Synset & definition
03813880	{hand, manus, hook, mauler, mitt, paw} (def: *the distal extremity of the superior limb*)
06135284	{handwriting, hand, script} (def: *something written by hand*)

Synset 06135284 uniquely identifies the concept "*something written by hand*", which can be represented in natural language by the terms *handwriting*, *hand* and *script*. The synsets in which *hand* appears correspond to the concepts the term can represent, or the senses of the term *hand*. Semantic relations between concepts are encoded in WordNet as pointers between synsets. Pertinent to our analysis is the fact that a term with a corresponding sense tag corresponds directly to a synset identifier, and hence sequences of term/sense tag pairs may be interpreted as sequences of synsets; whilst *hand* maps onto several synsets, '*hand* sense 1' maps onto synset 03813880 only. Reference to the synset, or concept, 03813880 invokes *hand* sense 2, *handwriting* sense 1 and *script* sense 2.

The two basic constructs in QUESCOT are a lexical cluster and a lexical chain. A lexical cluster is made up of those items of vocabulary constituting a distinct textual context, which can be active at various points within a document, and the linear position of those items; we appeal to the notion of a lexical chain to establish in which sections of a document a textual context is active, which in turn helps to eliminate any spurious items since all valid items in a cluster must also pertain to a lexical chain. A lexical cluster therefore embodies one or more lexical chains, which are broken once the distance between successive elements succeeds a pre-determined threshold. Figure 1 shows three distinct textual contexts and the corresponding lexical chains which reflect at which parts of the document the context is active (*appeal*$_{x,y}$ indicates that the term *appeal* in this position pertains to cluster *x*, chain *y*).

Appeal$_{1,1}$........*court*$_{1,1}$..............*drug*$_{2,1}$..................*sentence*$_{1,1}$.
......................................*judge*$_{1,1}$................ *case*$_{1,1}$ ....
...........*hearing*$_{1,1}$.................... ....*cocaine*$_{2,1}$ ....
: : : : : :
....
....
....
: : : : : :
..... *importing*$_{3,1}$......................*narcotics*$_{2,2}$.......... *illegal substances*$_{2,2}$ ....*distribution*$_{3,1}$..........
....*network*$_{3,1}$..................................*supply*$_{3,1}$...*cocaine*$_{2,2}$...........*cocaine*$_{2,2}$................. *heroin*$_{2,2}$
: : : : : :
....
....
....
: : : : : :
verdict$_{1,2}$...........*judge*$_{1,2}$............*prosecution*$_{1,2}$................
..........*importation*$_{3,2}$.......................................*narcotics*$_{2,3}$.............*cocaine*$_{2,3}$..*sentence*$_{1,2}$......
..........*suppliers*$_{3,2}$..............

Figure 1: Textual Contexts revealed in Lexical Clusters

In Figure 1, *cluster 1* embodies the context established by the terms {*appeal, court, sentence, judge, case, hearing, verdict, prosecution*}, and associates each cluster item with its position(s) in the text. There are two distinct sections of the document in which this context is active, reflected through the two lexical chains which the cluster embodies. By considering the distribution of the terms in a cluster throughout the document we quantify the 'strength' of a cluster, and hence a topic, and by normalising the set of scores produced for each cluster identified we hypothesise that the relative dominance of the respective topics within a document can be ascertained. A representation is produced in which the concepts identified in clusters, which we hypothesise are the most salient, are weighted according to the perceived dominance of their context of occurrence within the document. A key aspect of the document representation is that the items are assigned WordNet sense tags, and are hence disambiguated. The output from QUESCOT can be viewed as a set of clusters of concepts represented by weighted WordNet synonym sets. Details about the functionality of QUESCOT can be found in [24].

3. Applications of QUESCOT

In this section we describe four applications which exploit QUESCOT's output, concentrating on results and evaluation. We first investigate QUESCOT's suitability for the applications suggested by Morris and Hirst, word sense disambiguation and text segmentation, before describing two further applications in automatic indexing and automatic resource generation. In the former tasks we do not evaluate their effect on IR, whereas for the automatic indexing experiment we have evaluated our proposal within an experimental IR system. The sense-tagged collocation application describes a method we have implemented, but not yet fully evaluated.

3.1 Word Sense Disambiguation

The fact that language affords the opportunity to express the same concept in different terms and that a single term can express more than one concept is problematic for traditional term-based IR systems, which risk retrieving extraneous information as a consequence [6,12]. Tokens in QUESCOT representations are implicitly disambiguated, as stated above, and in this section we investigate how well QUESCOT performs word sense disambiguation, rather than directly investigating the effect in IR, since our aim is to establish QUESCOT's suitability for these applications.

Although QUESCOT accounts for adjectives and verbs in lexical clusters, it only disambiguates nouns with respect to WordNet sense numbers. For evaluation purposes we made use of a semantically tagged corpus of texts, where each word is tagged with a WordNet sense number. In a QUESCOT representation a concept can only appear once, thus a pertinent metric for our evaluation is the number of distinct concepts in a text. If the concept represented as *plane* sense 2 occurs six times in a document, then this constitutes a single occurrence of the concept when we refer to the number of text-distinct concepts. The sense-tagged corpus we used consists of 91 texts and 181303 words. The number of tagged nouns in this corpus is 38543, and the number of text-distinct tagged nouns is 22903[2].

All the texts in the corpus were processed by QUESCOT, and the results were evaluated with respect to the manually-assigned semantic tags; if QUESCOT assigned a WordNet sense number to a noun which corresponds to its tag in the corpus, the word is judged to be correctly disambiguated. It should be borne in mind, however, that some words have only one possible WordNet sense tag, and hence do not require disambiguation; counting such monosemous instances as correct disambiguations artificially inflates the performance of the disambiguation, and we therefore discount these when presenting the results. Running QUESCOT on the corpus, 3602 sense tags were assigned, of which 2731 were correct; 991 of the assignments made were to monosemous nouns. The results are summarised in Figure 2:

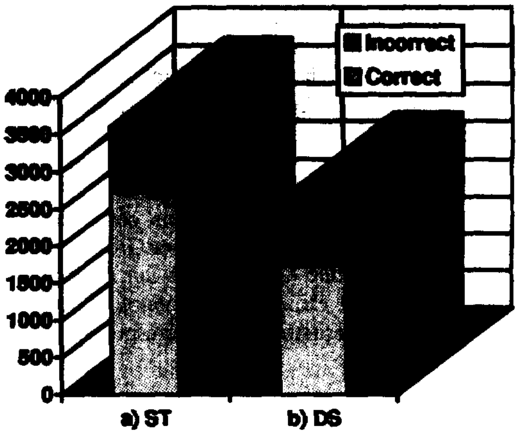


Figure 2: Representations of proportions of (a) ST: sense tagged nouns correctly tagged (b) DS: sense tagged nouns requiring disambiguation correctly disambiguated.

It is of note that the total of nouns disambiguated by QUESCOT in this evaluation is 3602, out of the total of 22093 distinct nouns in the corpus. Thus only 16% of nouns are disambiguated, the remainder not forming part of a QUESCOT chain. This is maybe less significant than it first appears, because those nouns disambiguated are likely to be the most salient as they pertain to a recognised lexical cluster. However, only 67% of those nouns requiring disambiguation within this subset are correctly disambiguated. This figure is in accordance with the figure reported by Okumura and Honda [15], who investigated the disambiguation of Japanese texts using a method based on similar principles. In mitigation, it is widely acknowledged that WordNet makes very fine-grained sense distinctions, and Richardson and Smeaton [19] note that in some cases nouns have different senses simply because the concept occurs at different places in the WordNet hierarchy. Nonetheless, we conclude that QUESCOT is not well suited to this application. The fact that one third of the salient nouns are incorrectly tagged is particularly pertinent in the light of Sanderson's investigation [22], which concluded that the performance of IR systems is "insensitive to ambiguity but very sensitive to erroneous disambiguation".

3.2 Text Segmentation

As full-text documents are increasingly available, attention has recently been turned towards passage retrieval accounting for segments, or fragments, of text [21]. Documents which are 'large and unwieldy' may be unsatisfactory responses to an information request [26] and retrieval of certain document types, in particular long documents and documents summarising many subjects, can be problematic for algorithms not accounting for *where* in a document the text matches the query [3]. Hearst and Plaunt investigate the effect of indexing documents as several independent segments, rather than as a whole document such that terms in segments are "indexed and weighted more accurately

[2] Notice that this differs from the number of distinct concepts in the collection (collection-distinct).

than if they had been a small part of a larger text" [9]. Significant gains in both recall and precision performance were reported. Having established that retrieval may benefit by accounting for segments, attention must then turn to what a segment represents.

The most useful segments of text for IR purposes are those which describe a particular topic, and Morris and Hirst suggest that their lexical chains, "spanning a topical unit of text", can provide a basis for the identification of such segments. We refer to the spans of the lexical chains identified by QUESCOT for this task, following Okumura and Honda [15] who describe a text segmentation application of their Japanese lexical chaining system which establishes where the bulk of one set of chains begins and another ends. Such points are good candidates as segment boundaries. In order to ascertain the effectiveness of QUESCOT for this task, we chose to evaluate it against Hearst's TextTiling algorithm [10] which was used to identify text segments in the IR experiments described in [9] and does not require a source of lexical knowledge.

We strive to eliminate subjective notions from the evaluation by artificially generating segments through concatenating articles, a method used by Reynar [18]. Whilst this evaluation cannot therefore be expected to reflect performance when applied to "real" documents, the method is valid for a *comparative* evaluation of techniques purporting to detect changes in topic flow; an algorithm which performs poorly in this artificial task cannot be expected to perform well when segmenting naturally occurring texts. The rationale underlying this approach is that the topics discussed in distinct newspaper articles are sufficiently diverse that the ability to distinguish between articles can serve as an objective measure of the algorithms' performance. In this evaluation a boundary found by an algorithm is deemed correct if it corresponds to those points at which an excerpt from a new article is introduced. These points correspond to changes in topic flow, and the ability to recognise this change, albeit extreme, serves as a criterion for comparatively measuring the performance of the respective algorithms.

We created twelve such documents by selecting arbitrary multi-paragraph units from several independent general interest articles, and concatenating these to create a document of the required length; we ensured that the units did not traverse section boundaries within a document, such that the paragraphs in a unit all pertained to the same section. The algorithms were evaluated in terms of recall, the number of correct topic boundaries located as a proportion of the number present in a document, and precision, the number of correct boundaries located as a proportion of the total number located. The standard deviation of these measures amongst the set of results generated has been calculated as a means of comparing the consistency of the two algorithms. It is clear that performance for this set of texts is closely matched, with both identifying the vast majority of segment boundaries. The results are summarised graphically in Figure 3:

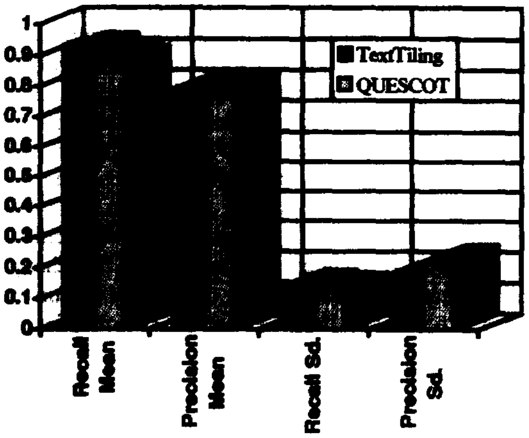


Figure 3:
Performance of QUESCOT and TextTiling for Text Segmentation

Due to the evaluation data used, the figures represent the upper bound of performance which could be expected when applying these methods to text segmentation. Few segment boundaries in expository text are marked by such distinct and marked changes in topic as were evidenced in the artificial documents used in the evaluation.

The purpose of the evaluation was to ascertain whether there was any significant performance difference between the two algorithms for text segmentation. Both algorithms achieve very high precision and recall, suggesting that both successfully recognise the artificially introduced changes in topic, and performance is very closely matched. However, the real significance of this result is the lack of evidence to suggest that QUESCOT outperforms the TextTiling algorithm, despite the significant extra processing involved. Furthermore, the application domain of TextTiling is not restricted, whereas the coverage of WordNet restricts QUESCOT.

3.3 Automatic Indexing and COATER

In this section we describe an automatic indexing application of QUESCOT, hypothesising that accounting for context of occurrence when indexing and retrieving documents will result in performance gains over retrieval methods which are purely term-based. In order to establish whether this method of indexing documents could improve information retrieval performance, we performed a comparative evaluation using the SMART IR system [2]. Further details of this experiment may be found in [23].

3.3.1 Overview of Approach

COATER[3], the experimental IR system we have developed, determines document relevance by establishing how homogenous concepts expressed in a query are with textual contexts evident in a document. The idea underlying the retrieval operation is that for each query concept we establish its context of occurrence, and

[3] COATER: Context-Activated Text Retrieval

then determine how dominant this textual context is within the document. We hypothesise that this provides improved precision performance, since the most relevant documents will evidence dominant textual contexts associated with concepts expressed in the query. Such a retrieval operation is possible within a vector space model of IR using QUESCOT representations, which are akin to indexes with sets of weighted tokens. The score of a lexical cluster is associated with each item in a cluster, so subsequent consultation of a particular item implicitly activates the whole textual context and means the matching process effectively considers the whole context of occurrence. The matching process can thus be viewed as activating a textual context in a document via a query item, and subsequently determining how dominant this context is in the document, which represents how semantically homogenous the query item is with the document. Because the indexing process is based entirely on QUESCOT representations, COATER uses the set of WordNet synsets as its controlled indexing vocabulary. We aim to establish whether COATER provides improved retrieval performance over well-established term-based approaches to text retrieval. There are two aspects to our approach, the controlled indexing vocabulary we use, where concepts are represented uniquely by WordNet synsets, and the criteria for weighting index items.

COATER was evaluated alongside the SMART information retrieval system, and the critical aspect of our evaluation was that both systems made use of the same cosine correlation matching function [20] for matching query representations to document representations; any difference in performance could thus be attributed to the indexing method used rather than to the matching function. The evaluation involved indexing a collection of 84678 documents with both systems, and generating a set of queries representing user information requests. Our aim was purely to demonstrate the feasibility of the method of indexing, as we do not claim that COATER represents a fully-functional IR system. In particular its recall performance is restricted by WordNet's coverage, which was reflected in the very simple queries we used for evaluation. For each of the twelve queries the top three documents retrieved by both COATER and SMART were retained and labelled as 'Group A' and 'Group B' respectively. We then obtained relevance judgements from eleven subjects, who were asked to categorise each document for its relevance to the query to which it pertained on a four point scale. Hypothesising that our method would be more able to retrieve those documents in which the main topic of the document related to the query, we made it possible to distinguish degrees of relevance as follows:

- *Category 1:*

Document is relevant to the query, and main theme in document is the topic in the query.

- *Category 2:*

Document is relevant to the query, but the topic in the query is not the main theme.

- *Category 3:*

Document as a whole is not relevant to the query, but some aspects of the query were addressed.

- *Category 4:*

Document is not at all relevant to query.

The evaluation was thus based on 864 relevance judgements. The results show that when retrieving a small number of documents, the precision performance achieved by COATER was substantially higher than that of the SMART system for the queries used in the evaluation. Of the 10 subjects who expressed a preference, all agreed that those documents presented as Group A, retrieved by COATER, were generally more relevant than those presented as Group B which were retrieved by the SMART system. Overall, 81% of the documents retrieved by COATER were deemed to be relevant by the subjects (falling into either category 1 or category 2), compared to 57% of those retrieved by SMART. The results by category are summarised graphically in Figure 4.

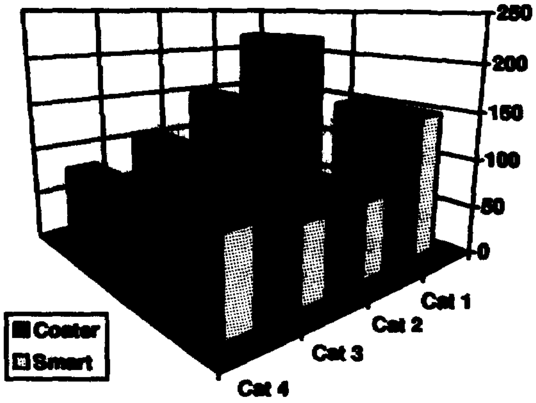


Figure 4:
Summary of COATER vs. SMART comparative evaluation

These results are encouraging, and indicate that the indexing method described leads to improved performance for the simple queries used in the evaluation over the SMART system using term frequency-based (tf/idf) measures alone (see e.g. [20]). However, the result should be tempered by the fact that WordNet's coverage severely restricts the recall performance. The lack of coverage of proper names is of particular concern in an IR scenario, and this must be addressed if the method described is to be used in a 'real-word' scenario. These restrictions currently prohibit evaluation of COATER with standard IR test collections.

3.4 Discovery of sense-tagged collocations: A resource for query expansion and topic identification

Resnik [17] points out that the lack of corpora annotated with word sense information prohibits the extension of distributional grouping methods to word senses. QUESCOT document representations constitute a partially sense tagged corpus, and we consider how these may be exploited to produce concept groupings, based on collocational relations between WordNet concepts; such groupings are tailored to a specific document collection and can be exploited for IR purposes. A collocational relation holds between two elements which co-occur within a specified range at a frequency much greater than that suggested by chance alone; often the precise relation between the terms can be difficult to specify such as *hospital/doctor*, *vicar/religion*.

A resource which relates to a specific document collection is useful in two ways. When a topic is treated in a document there is a tendency for words which relate to a particular vocabulary domain to be used, although the precise relation between these words can be difficult to establish and may not be accounted for in a resource based on specific relations between concepts, such as WordNet. Thus a means of automatically deriving such relations from a document collection can potentially lead to better identification of topics in documents, an important criterion for successful information retrieval. Second, query expansion processes can have recourse to extrinsic sources of lexical knowledge [25], and a resource tailored to a document collection should lead to more accurate and comprehensive query expansion. We have implemented CASSAN[4] to elicit synset co-occurrence information having processed a large document collection with QUESCOT.

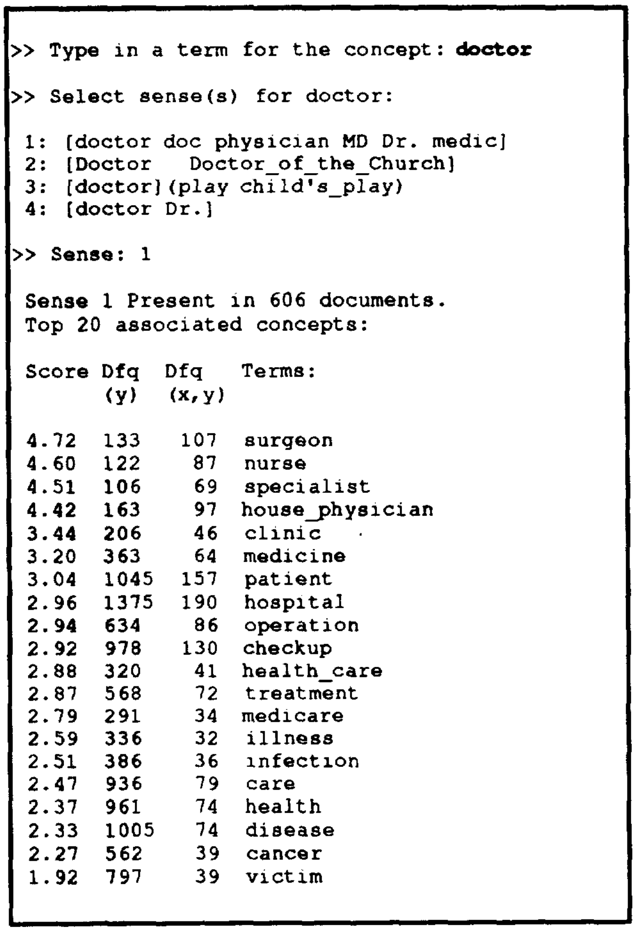
Analysing each of the 84678 document representations available from the previous application, it is possible to determine which synsets tend to co-occur frequently within a document, whether or not they occur in the same textual context, and subsequently to form clusters of related synsets based on these observations. Thus although QUESCOT does not link *explosion* and *bomb*, since this link can not be derived from WordNet, CASSAN will identify that textual contexts including the concept *bomb* frequently co-occur in documents with textual contexts containing *explosion*, suggesting a relation between the corresponding synsets. Importantly, QUESCOT's sparse representations allow us to consider associations which occur anywhere in a document, in contrast with other term-based co-occurrence analyses which are usually restricted to identifying relations within a fixed window of text [1,4]. We calculate, for every synset s, the number of documents in which s occurs, and the number of documents in which s co-occurs with every other synset. When we come to determine the strength of association between synsets x and y, we know $P(x)$, the probability of x occurring in any document, $P(y)$, the probability of y occurring and $P(x,y)$, the probability of x and y occurring together in the same document.

Using a measure based on the concept of mutual information [4] the association between synonym sets can be quantified. If we know the joint and independent probabilities of two items x and y, then the mutual information between the items is:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

If there is an association between x and y, then the joint probability $P(x,y)$ will be much larger than chance, $P(x)\,P(y)$, and hence $I(x,y) >> 0$.

CASSAN locates the synsets (concepts) most associated with a specified synset (concept). The operation of the interactive version is shown in Figure 5, where we seek the concepts most associated with the concept of a *medical doctor* (the column *score* represents the mutual information (MI score), *Dfq (y)* is the number of documents in which the associated concept y occurs, and *Dfq(x,y)* is the number of documents in which *both* concepts occur).



```
>> Type in a term for the concept: doctor

>> Select sense(s) for doctor:

 1: [doctor doc physician MD Dr. medic]
 2: [Doctor   Doctor_of_the_Church]
 3: [doctor](play child's_play)
 4: [doctor Dr.]

>> Sense: 1

 Sense 1 Present in 606 documents.
 Top 20 associated concepts:

 Score Dfq  Dfq   Terms:
       (y)  (x,y)

 4.72  133   107  surgeon
 4.60  122    87  nurse
 4.51  106    69  specialist
 4.42  163    97  house_physician
 3.44  206    46  clinic
 3.20  363    64  medicine
 3.04  1045  157  patient
 2.96  1375  190  hospital
 2.94  634    86  operation
 2.92  978   130  checkup
 2.88  320    41  health_care
 2.87  568    72  treatment
 2.79  291    34  medicare
 2.59  336    32  illness
 2.51  386    36  infection
 2.47  936    79  care
 2.37  961    74  health
 2.33  1005   74  disease
 2.27  562    39  cancer
 1.92  797    39  victim
```

Figure 5: A CASSAN Analysis

The analysis reveals those concepts deemed to be strongly associated with the synonym set {doctor doc physician MD Dr. medico}, based on mutual information scores, and the synsets involved appear to be mutually related[5]. This analysis is automatically applied to each synset in the QUESCOT representations of documents and clusters are formed based around the most *sticky* concepts, those which attract a high number of strongly associated concepts and are intuitively an appropriate basis for concept clusters. We create clusters consisting of those synsets which associate with these sticky synsets with $I(x,y) > 2.5$, our threshold for determining which synsets are strongly related[6]. The synset representing 'doctor' in Figure 5 would thus represent a sticky synset. Our proposal is that the clusters formed may be used to supplement the lexical

[4] CASSAN: Concept Association Analysis

[5] This intuition is confirmed to some extent by Philips ([16]), who found that words often inter-collocate.

[6] Church and Hanks ([4]) observed that pairs with $I(x,y) > 3$ are generally interesting, whereas smaller values are generally not. We relaxed this criterion slightly in order that a reasonable number of clusters could be generated from the data available.

information in WordNet, so that collocational relations between WordNet concepts may be derived. Using this process on the QUESCOT representations we had generated for the previous application, 274 clusters of 12 concepts each were produced. Although the majority of the clusters seem to be coherent, we have yet to evaluate them fully or to investigate how efficiently they supplement the data available in WordNet.

4. Discussion and Conclusion

Each application described above is based solely on the output of the QUESCOT textual context analysis we have described, the analysis itself being application independent. Indeed, this is part of the appeal since several aspects of an IR system can potentially benefit from a single analysis. Our investigations have shown, however, that the analysis is not well suited to word sense disambiguation and provides no added value compared with a simpler approach to text segmentation. Even within the reduced subset of nouns for which QUESCOT could propose a sense tag, only 67% of nouns were correctly disambiguated which represents very mediocre performance. Similarly, the results from the comparative text segmentation evaluation suggest that a process based on QUESCOT does not outperform a more appealing dedicated segmentation method which does not depend on a specific lexical resource, and hence can be applied to documents in any domain. We conclude that analyses dedicated to the respective applications are therefore preferable.

Using QUESCOT as an automatic indexing component in an IR system produced more positive results, suggesting that accounting for textual context can improve precision performance. The system we developed is based on a vector-space model of IR which means established matching functions can be applied. Recall performance, however, is restricted by the coverage of the WordNet database and this currently prohibits the use of our system in a "real-world" IR scenario. The vast majority of information requests we have encountered are based around proper names, which are generally not represented in WordNet. It would be interesting to investigate the effect of using a high recall IR engine as a preliminary retrieval stage, and then considering textual contexts within the subset of retrieved documents. Alternatively, a proper name recognition and classification stage, for example exploiting techniques proposed by Coates-Stevens [5], could be incorporated into the QUESCOT analysis.

A consequence of QUESCOT's dependence on WordNet is that any gaps in the coverage of WordNet are reflected in QUESCOT's results. One notable area in which WordNet is lacking is collocational information, and our final application investigates how a supplementary resource, tailored to a specific document collection, can be automatically derived. A resource detailing which concepts in a document collection are related is useful for information retrieval from that document collection, in particular for improved topic-identification and query expansion purposes. We have implemented a procedure for automatically generating such a resource from the QUESCOT representations. Although it is not immediately clear how to evaluate such a resource, the majority of concept clusters appear coherent, and imply relations between concepts which it would not be possible to derive from WordNet.

The process by which QUESCOT representations are derived is relatively unconstrained, and inexact in nature. Successful applications of the analysis are based on document-level descriptions where the granularity of the analysis allows local inaccuracies to be smoothed out, as in the case of the automatic indexing and collocation analysis applications. The sense disambiguation application operates at a finer level of granularity and, because the weights assigned to items are not accounted for, there is no scope for smoothing out inaccuracies. We therefore feel that although invalid sense tags can occur in document representations, a QUESCOT analysis can be successfully exploited to improve precision performance in IR if the weights assigned to items in document representations are accounted for, and can furthermore provide a useful source of sense-tagged collocation relations when applied to a large corpus of text.

References

[1] Brown, P., deSouza, P., Mercer, R., Pietra, V., Lai, C., Class-Based n-gram Models of Natural Language, Computational Linguistics, 18 (4), pp. 467-479, 1992

[2] Buckley, C., Implementation of the SMART Information Retrieval System, Technical Report TR85-686, Computer Science Department, Cornell Univeristy, 1985

[3] Callan, J.P., Passage-Level Evidence in Document Retrieval, In: Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR-94), pp. 302-310, 1994

[4] Church, K.W., Hanks, P., Word Association Norms, Mutual Information and Lexicography, Computational Linguistics, 16 (1), pp. 22-29, 1990

[5] Coates-Stephens, S., The Analysis and Acquisition of Proper Names for Robust Text Understanding, PhD Thesis, Department of Computer Science, City University, London, 1992

[6] Furnas, G., Landauer, T.K., Gomwz, L.M., Dumais, S.T., The Vocabulary Problem in Human-System Communication, Communications of the ACM, 30 (11), pp. 964-971, 1987

[7] Grosz, B.J., Sidner, C.L., Attentions, Intentions and the Structure of Discourse, Computational Linguistics, 12(3), pp. 175-204, 1986.

[8] Halliday, M.A.K., Hasan, R., Cohesion in English, Longman, London, 1976

[9] Hearst, M., Plaunt, C., Subtopic Structuring for Full-length document access, In: Proceedings of the 16th International Conference on Research and Development in Information Retrieval (SIGIR-93), pp. 59-68, 1993

[10] Hearst, M., Multi-paragraph segmentation of Expository Text, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994

[11] Kozima, H., Ito, A., Context-Sensitive Measurement of Word Distance by Adaptive Scaling of Semantic Space, Proceedings of Recent Advances in Natural Language Processing (RANLP-95), pp.161-168, 1995

[12] Krovetz, R., Croft, W.B., Lexical Ambiguity in Information Retrieval, ACM Transactions on Information Systems, 10(2), pp. 115-141, 1992

[13] Miller, G.A., Special Issue, WordNet: An On-line Lexical Database, International Journal of Lexicography, 3 (4), 1990

[14] Morris, J., Hirst, G., Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, Computational Linguistics, 17 (1), pp. 21-45, 1991

[15] Okumura, M., Honda, T. Word Sense Disambiguation and Text Segmentation based on Lexical Cohesion, Proceedings 15th International Conference on Computational Linguistics (Coling-94), pp. 755-761, 1994

[16] Philips, M., Aspects of Text Structure: An Investigation of the Lexical Organisation of Text, North Holland, Amsterdam, 1985

[17] Resnik, P., Disambiguating Noun Groupings with Respect to WordNet Senses, Proceedings of the Third Workshop on Very Large Corpora, 1995

[18] Reynar, J., An Automatic Method of Finding Topic Boundaries, Proceedings 32nd Annual Meeting of the ACL, 331-333, 1994

[19] Richardson, R., Smeaton, A.F., Automatic Word Sense Disambiguation in a KBIR Application, Working Paper CA-0595, School of Computer Applications, Dublin City University, 1995

[20] Salton, G., Buckley, C., Term-weighting Approaches in Automatic Text Retrieval, Information Processing and Management, 24 (5), pp. 513-523, 1988

[21] Salton, G., Allan, J., Selective Text Utilization and Text Traversal, International Journal of Human Computer Studies, 43 (3), pp. 483-497, 1995

[22] Sanderson, M., Word Sense Disambiguation and Information Retrieval, Proceedings of the 17th ACM-SIGIR Conference, pp. 142-151, 1994

[23] Stairmand, M.A., Black, W.J., Contextual and Conceptual Indexing using WordNet-derived Lexical Chains, Proceedings of the 18th BCS-IRSG Colloquium on Information Retrieval Research, pp. 47-65, 1996

[24] Stairmand, M.A., A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval, PhD Thesis, Department of Language Engineering, University of Manchester Institute of Science and Technology, 1996

[25] Vorhees, E.M., Query Expansion using Lexical Semantic Relations, Proceedings of the 17th ACM-SIGIR Conference, pp. 61-69, 1994

[26] Wilkinson, R., Effective Retrieval of Structured Documents, Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR-94), pp. 302-310, 1994