Experiments on Using Semantic Distances Between Words in Image Caption Retrieval

Alan F. Smeaton and Ian Quigley
School of Computer Applications
Dublin City University
Glasnevin Dublin 9, IRELAND.

E-mail: Alan.Smeaton@CompApp.DCU.ie

Abstract

Traditional approaches to information retrieval are based upon representing a user's query as a *bag of query terms* and a document as a *bag of index terms* and computing a degree of similarity between the two based on the overlap or number of query terms in common between them. Our long-term approach to IR applications is based upon pre-computing semantically-based word-word similarities, work which is described elsewhere, and using these as part of the document-query similarity measure. A basic premise of our word-to-word similarity measure is that the input to this computation is the correct or intended word sense but in information retrieval applications, automatic and accurate word sense disambiguation remains an unsolved problem. In this paper we describe our first successful application of these ideas to an information retrieval application, specifically the indexing and retrieval of captions describing the content of images. We have hand-captioned 2714 images and to circumvent, for the time being, the problems raised by word sense disambiguation, we manually disambiguated polysemous words in captions. We have also built a collection of 60 queries and for each, determined relevance assessments. Using this environment we were able to run experiments in which we varied how the query-caption similarity measure used our pre-computed word-word semantic distances. Our experiments, reported in the paper, show significant improvement for this environment over the more traditional approaches to information retrieval.

1 Introduction

In information retrieval applications, a user's information need is normally expressed as a query which can be a set or a boolean combination of keywords, a natural language statement or it can be the result of a user-system dialogue. Documents can be similarly represented using some surrogate like a set of index terms or phrases. In most of the traditional approaches to IR implementation, the matching function between the users query and the internal representation of a document will be based on some function of the overlap between query terms and document terms. Much of IR research has been conducted under this premise and could be termed as attempting to match the "bag of terms" representation for a query against the "bag of terms" representation of documents.

There are problems with this approach to information retrieval caused by [Smeaton, 1992]

- the same words or phrases being used to describe different things, termed *polysemy*, e.g. the word "bar" meaning a place for refreshments, a piece of metal, a restriction on a window, etc.
- different words or phrases being used to describe the same thing, e.g. "stomach pain" and "belly ache"
- natural language, the vehicle with which the content of documents and queries is expressed, is fraught with ambiguities at all levels leading to multiple interpretations of words, phrases, etc.

One of the most common ways to address the above problems in IR is to use some kind of query expansion in which the user's initially selected index terms are augmented either manually [Efthimiadis, 1995] or automatically [Smeaton et al., 1983] by the addition of terms related in some way to the query or to known relevant documents. Such query expansion does go some way towards compensating for the IR problems mentioned above but clearly does not solve them all.

Our long-term approach to information retrieval applications is to index queries and documents by the words which occur within them but when computing the degree of similarity between query and document instead of basing similarity on the terms in common between the two we incorporate a quantitative measure of the semantic similarity between index terms into the measure. In this way we hope to significantly address the second and to a lesser extent the first of the 3 basic problems in IR shown above. In order to automatically determine word-word similarity we need some kind of knowledge base and in our case we use a set of hierarchical concept graphs (HCGs) [Richardson, 1994] derived from WordNet and augmented with a statistical analysis of word frequency occurrences in a corpus.

Using HCGs to compute word similarity is computationally expensive however, and even more important for us in our attempt to develop IR applications based on word-word

To appear in *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* Zürich, August 1996.

distances, the word sense disambiguation problem remains unsolved. Principally because of this latter reason, our initial forays into using our approach for document retrieval on the TREC collection did not yield overall improvements in retrieval effectiveness [Richardson et al., 1995] and in fact our results were a disimprovement on our own statistically-based benchmark. These first experiences demonstrated to us that we should have been a little less ambitious in pushing our theory into practice on such a large scale. With this in mind we considered implementing our approach to query-document matching on one of the smaller IR test collections like CACM but decided against doing this because IR as a field has moved on to experiments on much larger collections and results on smaller collections are now regarded as being less significant. In addition, even the older test collections have document representations which are rich in terms of the numbers of index terms and our feeling was that our approach to retrieval would work best when both queries and documents were short.

With this background, we decided to try our retrieval approach on a corpus of image captions which we built ourselves and the experimental results and evaluation of using our word-word similarity measure in caption retrieval are reported in this paper. In the following section we give some brief background on using word distance measures and we sketch out how we transformed WordNet into HCGs and how we use these to compute the similarity between word pairs. Section 3 describes our experimental set-up, the images we used, how they were captioned and indexed and how we collected queries and relevance assessments. In section 4 we present a selection of the most significant of our experimental results and in section 5 we give an outline of other approaches to retrieval of image captions. Section 6 presents our conclusions and analysis.

2 Measuring Semantic Distance Between Words

A number of approaches to measuring conceptual similarity between words have been taken in the past. Tversky's feature based similarity model, [Tversky, 1977], is arguably the most powerful similarity model to date. However, its applicability in our IR situation would require a much richer knowledge base than is available to us. Although our WordNet-derived knowledge base is quite thorough in its coverage of concepts, the number of semantic relation types connecting these concepts is considerably less than would be required for use by a feature based similarity model.

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English, [Miller et al., 1990]. Information in WordNet 1.4, the version we use, is organised around logical groupings called synsets, each of which consists of a list of synonymous word forms and semantic pointers that describe relationships between the current synset and other synsets. A word form can be a single word or two or more words connected by underscores, (referred to as collocations).

In our work we only use the nouns from WordNet as a knowledge base, ignoring the verbs, adjectives and adverbs. Our initial knowledge base consisted of a number hierarchical concept graphs, (HCGs), automatically constructed from WordNet data files with the root concepts chosen as result of a set of experiments to determine which would provide maximum coverage whilst minimising the degree of overlap between HCGs. The resulting HCGs ranged in size from 43950 unique concepts (Entity) to 688 concepts (Shape) [Richardson, 1994].

Our information based approach to measuring semantic similarity is based on work carried out by Resnick, [Resnick, 1993a, Resnick, 1993b] who views noun synsets as a class of words where the class is made up of all words in a synset as well as words in all directly or indirectly subordinate synsets. Conceptual similarity is considered in terms of class similarity. The similarity between two classes or synsets in our case is approximated by the information content of the first class in the hierarchy that subsumes both classes. The information content of a class is approximated by estimating the probability of occurrence of the class in a large text corpus and our case the probabilities of occurrence of classes were computed from an analysis of the frequency of occurrence of WordNet terms using a POS tagged corpus of 278 Mbytes of text. Thus the similarity of two classes can be expressed as:

$$Sim(c_1, c_2) = max_{c_i} \left[log \left(\frac{1}{P(c_i)} \right) \right]$$

where $\{c_i\}$ is the set of classes dominating both c_1 and c_2, $P(c_i)$ is the class probability of class c_i, and $log\left(\frac{1}{P(c_i)}\right)$ is the information content of class c_i. This is described in more detail elsewhere [Richardson, 1994, Richardson et al., 1995].

Despite these apparent strengths of our information based similarity measure, it is not without weaknesses. One of these is apparent in the method of calculating the information content of classes as many polysemous words and multi-worded synsets will have an exaggerated information content value. If one takes for instance the word 'bank', the information content for this word will include all occurrences of bank in the corpus, regardless of word sense. This gives the same (exaggerated) information content value to a 'commercial bank' as a 'river bank'. Also, due to the fact information content values are calculated for synsets as opposed to individual words, it is possible for the information content value to be over-exaggerated in situations where synsets are made up of a number of commonly occurring ambiguous words. For example in the synset {*yield, fruit*}, the information content value of this synset is calculated both from the frequencies of the word *'fruit'* and the word *'yield'*. Given the fact that the information content of a class is defined in terms of the information contents of its subordinate classes, super-classes of classes containing polysemous words are similarly over-valued. This disregard of ambiguous words is a problem given the fact that synsets in our WordNet-derived KB refer to particular senses of words and the KB as a whole tends to include very fine sense distinctions in an attempt to have an exhaustive coverage of concept meanings. Notwithstanding these drawbacks, we were able to evaluate the performance of our measure with human ranking of word pair similarities being the ground truth against which we computed Spearman correlations and our results indicated favourable performance [Richardson et al., 1994].

Computing the similarity between two word senses can only be done if both are in the same HCG, otherwise they are regarded as being dissimilar. Once they have been located in the same HCG we then recursively trace up through the HCG until we find a common parent. Given that HCGs are large we were unable to fit them in main memory and so much IO was required in the computation. To address this we pre-computed the similarity between c.150,000,000 word pairs and we store these as one large file. Thus a similarity value computation now takes two hash function executions (on the words being compared) and one disk lookup. This provided us with fast enough response time to perform our experiments.

3 Experimental Set-up

The images we used in our experiments came from a CD-Rom entitled "Gifs Galore" which cost US$39. There was no particular reason for using this collection except that it was cheap, available, the images were in a format that suited us (GIF) and there were many thousands of images of different kinds. On the CD-ROM, the images were given filenames up to 8 characters and images were categorised into one of about 50 categories like "Buildings", "Military", "Sports", etc. The manual captioning of the images was done over a 4-week period, about 130 hours in total or an average of just over 2.5 minutes per image. This is less than the time taken to manually index a text document but it must be remembered that reading and comprehending the content of an image can be done much quicker than performing the same exercise on text.

The indexer examined each image and determined the most important physical objects in the image and then assigned it one or more descriptive phrases. Images were presented to the indexer in a randomly generated order to ensure continuous variety. Some images were small variations of others in the collection, e.g. two pictures of the same soldier sitting on the same horse in front of the same castle, but taken at different angles. In such cases where images to be captioned were small variations of others, the indexer was instructed to disregard the second and subsequent images. Similarly the indexer was instructed to disregard images which did not contain obvious physical objects, such as Mandebrot functions or abstract artistic images. These constraints reduced the size of the image set to 2714. Details of the size of captions are included in Table 1

On assigning descriptive phrases to images (a sample are given in Appendix A) the indexer was then asked to disambiguate any polysemous words or collocations used, polysemous with respect to the WordNet lexical database. After some familiarisation with this process, the indexer would disambiguate into more than one WordNet senses due primarily to the very fine sense distinctions used in WordNet. This finding is in line with the findings of others [Voorhees, 1993] as well as with other work we have done using WordNet [Smeaton et al., 1996].

To create a collection of user queries we involved 22 people from our research laboratory. Each was given a work situation in which they played the role of an editorial assistant for a magazine who had been assigned the task of finding images from the company files to illustrate an article to be published in the magazine. The users were asked to decide upon the title for the magazine article and this was done with dialogue assistance from us. They were then asked to think of some number of image descriptions which could be used as pictorial illustrations for that article. Some users thought of only one image, others thought of up to 5 or 6; the average user thought three images sufficient. This whole process of query formulation took an average of 10 minutes per subject.

The text of the subjects' image descriptions were used as our queries and as found with other image querying applications[1] most queries are very short, in many cases

Total number of captions/images	2714
Avg. number of indexing terms per caption	5.89
Avg. number of collocations per caption	0.5
Total number of usable queries	47
Total number of relevant images	474
Avg. relevant captions per usable query	10.1

Table 1: Statistics for the Test Collection

consisting of single phrases, and topic rather than visually based. In order to determine a set of relevant images per query we computed a set of sample images to present to the original user and asked him to perform relevance judgements. This set was almost always presented back to the user within a couple of hours. The set to be judged was determined based on choosing the top-ranked images from a traditional statistically-based IR implementation using words rather than word senses as index terms[2]. This was augmented manually for each query by images taken from whatever of the c.50 categories the original indexer would have suggested as possible starting points for browsing, though throughout our experiments our users were never made aware of the existence of these categories.

The subjects who formed the queries were asked to make relevance judgements based on selecting magazine images they would present to their editor for illustrating the article. Individual queries and sets of images relevant to those queries were to be regarded as independent of each other and thus the same image might have appeared in the relevant set for two queries from the same subject. For some queries (13 of the 60 we processed) there were no relevant images to be found and these queries were subsequently disregarded for our experiments. For the remainder, the average number of relevant images per query was 8.7 though some of this was due to one or two queries with a disproportionally high number of relevant images while at the same time 48% of the queries had less than 5 relevant images. Statistics for the collection are given in Table 1.

The work on creating a collection such as this provided us with an experimental environment with short captions and short queries, the environment in which we believe our word-word distance has greatest potential. A similar environment would be schema resolution in federated distributed databases [Richardson et al., 1996]. However there are pitfalls and drawbacks with the approach we used in constructing this test environment; the collection is relatively small by TREC standards, the notion of manual disambiguation of all caption terms for very large scale image collections may not be realistic, the queries are synthetic and artificial, the exhaustivity of the relevance judgements could be questioned, etc. Our approach of pooling relevance judgements mirrors that taken in the TREC experiments [Harman, 1995] and is the best that could be done in the circumstances.

4 Experimental Results

In evaluating the retrieval performance of various strategies incorporating word-word semantic distances we used the standard evaluation mechanism used in the series of TREC evaluations for document retrieval [Harman, 1995].

[1] This has been found in a manual analysis of 6,000 queries to the image collection searchable through the CIIR [Croft, 1995] as well as a log analysis of queries to the WWW interface to our system which to date has over 21,000 logged searches (Apr 1996).

[2] It may appear that using word senses would be more accurate but in using words rather than senses we eliminated any noise caused by the indexers choices in the sense disambiguation process. Thus the query for *"plants on the seashore"* retrieved images of *"vegetation at the coast"* in computing the set to be judged for relevance.

These calculate precision values interpolated to standard recall points and then average them over the full set of 47 queries.

In our research we performed many experimental runs though in this paper we summarise our results into 8 sets of figures which are presented later in Table 2. In describing some of the retrieval strategies which we used we use the following notation: a query $Q = \{q_1, q_2, \ldots q_m\}$ is matched against a caption $C = \{c_1, c_2, \ldots c_n\}$ where a q_i or a c_j is the original term used only as a representation for its synset, i.e. we represent queries and captions by sets of synsets and thus synonymous terms in a given synset will be treated as equal. $Sim(t_i, t_j)$ is the similarity between the sense-disambiguated form of two terms t_i and t_j.

The first retrieval strategy we implemented (**Run1**) was a straightforward statistically-based tf*IDF match between the word forms or strings, i.e. not using our manual word sense disambiguation of captions or of queries. The retrieval performance of this first run is exceeded by **Run2** where terms in a caption as well as terms in a query are both expanded to include other word strings from their sense-disambiguated synsets. This results shows that bringing in synonyms from synsets improves retrieval but only because both queries and captions are short. We know that adopting such term expansion does not generally work in document retrieval applications [Voorhees, 1993]. The third retrieval strategy we tried uses our word-word similarity outlined in section 2. In **Run3** we take a brute force approach to incorporating this by computing the score between a query and a caption as the sum of the similarities between all pairs of query term and caption term, normalised by dividing by the number of terms in the query times the number of terms in the caption. The formula for computing a score in Run3 is given below. The results of Run3 show a considerable decrease in retrieval performance. Analysis showed that this is due primarily to the noise being brought into the computation by some similarity values. In particular, our measure does not cope with having to compute similarity between words from different HCGs and if a query and caption both contain many terms from the same HCG then the similarity score will be high. For example, Q = {race_cars on a race_track}, D1 = {man in a living_room at a table watching television}, D2 = {cars racing at the grand_prix in Monaco}. Q and D1 contain concepts exclusively from the same HCG (entity) and thus there will be some similarity between all pairs of terms in the match between Q and D1 while for D2, only the term "car" is in that HCG and thus the overall score for D2 will be less than that for D1.

$$Run3(Q,C) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} Sim(q_i, c_j)}{m \times n}$$

Our solution to handling noise introduced into the measure in this way is to apply a word-word simlarity threshold, θ, such that only similarities above this threshold are incorporated into the computation and so the low degree of similarity between "race_cars" and "living_room" which contributes to the inflated score for D1, would not be used. After some trial runs, **Run4** shows the best performance of our global threshold while **Run5** shows the performance when we have a different threshold for each HCG, given that there is a concentration of usage of concepts from some HCGs (like entity) and hardly any use of others (like shape). The results of Run4 and Run5 do show improvements over not using thresholds but we are still far short of out statistically-based initial retrieval strategy, i.e. Run2. In determining the values of the best set of threshold values we even broke up the larger HCGs into smaller ones and determined the best set of threshold values for sub-HCG units but the impact on retrieval effectiveness was marginal.

$$Run4(Q,C) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} Sim'(q_i, c_j)}{k}$$

where

$$Sim'(x,y) = \begin{cases} Sim(x,y) & : \quad iff \quad Sim(x,y) > \theta \\ 0 & : \quad otherwise \end{cases}$$

For **Run6** we decided to try a different approach. For each query term we determine which is the best-matching or most similar caption term and we compute an overall caption score as the sum of such similarity values, normalised by the number of query terms.[3] The resulting performance of Run6 shows a significant improvement and a result which is slightly better than our starting benchmark. To compliment Run6, in **Run7** we tried the reverse, that is for each caption term we determine, by word-word similarity, the best-matching query term and we compute a caption score as the sum of these best pairwise similarities, normalised by the number of caption terms.

$$Run6(Q,C) = \frac{\sum_{i=1}^{m} max[Sim(q_i, c_j)]}{m} \qquad j = 1 \ldots n$$

$$Run7(Q,C) = \frac{\sum_{i=1}^{n} max[Sim(q_i, c_j)]}{n} \qquad i = 1 \ldots m$$

The results for Run7 show a big drop in performance which can be explained by the fact that more noise relative to a given query, will occur in captions than in queries. During captioning, a set of the most salient features in an image are captured while we know from studies of user queries that they tend to be much more specific and focused with each query term being important [Enser, 1993].

Finally we decided to fuse together the rankings generated from Run6 and Run7 by normalising the set of scores from each into the range [0,1] and simply adding the scores from the different rankings to compute new caption scores. Effectively we have taken two asymmetric metrics and by combining in this way, made them symmetric. The results of this fusion are in **Run8** and these results show a further improvement in effectiveness over Run6 demonstrating that Run6 and Run7 seem to be independent retrieval strategies retrieving different caption sets and hence the reason why the data fusion works so well in this case [Belkin et al., 1995].

$$Run8(Q,C) = Run6(Q,C) + Run7(Q,C)$$

The results of Run8 are an improvement over either Run1 or Run2, our starting points.

5 Retrieval of Image Captions

With the present state of image processing techniques we cannot yet do large scale, domain independent recognition of the content of images though this is one of the most active areas of computing research and has been for decades [Cawkell, 1993]. Whatever about speed and labour costs,

[3] An unnecessary normalisation for this strategy perhaps but we shall see later how this is used.

	Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8
Precision at Recall values:								
0.00	.451	.584	.434	.498	.527	.601	.492	.670
0.10	.422	.530	.363	.449	.474	.561	.450	.620
0.20	.392	.472	.316	.415	.429	.505	.381	.555
0.30	.386	.436	.261	.402	.413	.437	.335	.481
0.40	.359	.404	.240	.368	.385	.409	.310	.460
0.50	.347	.389	.230	.359	.374	.395	.287	.447
0.60	.252	.295	.163	.294	.301	.350	.196	.370
0.70	.228	.253	.131	.285	.282	.332	.183	.333
0.80	.207	.238	.122	.253	.250	.302	.150	.305
0.90	.170	.208	.105	.214	.208	.259	.129	.249
1.00	.139	.190	.098	.181	.175	.230	.120	.219
Av P over all	.286	.323	.206	.320	.327	.372	.257	.405
P at rank 10	.179	.235	.145	.200	.211	.215	.170	.232
P at rank 20	.148	.165	.097	.161	.172	.162	.123	.181

Table 2: Performance figures for experimental runs

the advantages of consistency and standardisation in completely automatic indexing would be obvious. Meanwhile, with the huge interest in multimedia indexing and retrieval, some kind of indexing for retrieval is required and the easiest to produce is the manual description or caption. Such text descriptions are currently the most important for multimedia indexing applications [Croft, 1995].

An extensive study by Enser [Enser, 1993] on the demand from users querying the Hulton Deutsch collection, the largest private image archive in Europe, revealed that users querying an image collection tended to be quite specific in their requests and information needs, much more than when querying, say, an academic library. The implication thus is that users' query terms in image searching will be quite precise and narrow and thus a simple term-term match between query terms and caption terms is generally not likely to be effective.

Rowe and Guglielmo [Rowe et al., 1993] have described the MARIE system in which captions and queries are translated into predicate-calculus expressions equivalent to semantic networks on which subsequent retrieval for the 36,000 captions is based. Their high-scoring retrieval evaluation however, is done in terms of only a 1951-word lexicon and 217 images. The SEYMOUR system [Flank et al., 1995] provides WWW access to a collection of over 300,000 images via their descriptor captions once again using an NLP-based analysis. An exhaustive list of the caption-based image retrieval systems available today, many of them available through WWW, would be impossible but we can see that they seem to be divided into those that retrieve on captions, usually with some relatively sophisticated processing on query, caption or both.

As noted by Besser [Bessner, 1990] among others, manual interpretation of image content will always be open to criticism of subjectivity and there are systems available in which image retrieval is based on actual image content. The QBIC system[4] [Niblack, 1993] is one such system. Starting with a sample image as a query, similar images are found by matching based on colour percentages, colour layout, textures or the shapes of images and their objects. The user is allowed to control the relative importance of each of these 4 retrieval methods into one unified retrieval as demonstrated with their WWW system on a collection of 1,900 images. The VIRAGE online demo[5] on 10,000 images also uses a user-controlled combination of colour, composition, texture and structure to yield retrieval using a sample image as a starting point. Finally, the Excalibur XRS Image Server[6] uses adaptive pattern recognition techniques and a neural network to identify patterns in an image and 'clue' patterns from images whose content is known, e.g. face, fingerprint, etc. are matched against images to be indexed in order to determine their representation.

In general, most of these approaches to image retrieval have not been evaluated quantitatively to determine their retrieval effectiveness in the sense that the traditional approach to information retrieval research would have done. In addition, methods such as those mentioned above tend to be computationally expensive and thus in the meantime the short-term prospects for caption-based indexing and retrieval look sound. The long-term potential for caption-based retrieval with respect to developing techniques in image processing is more difficult to judge but it is believed by many that a combination of caption and image based retrieval holds the best promise. The experiments that we have reported in this paper show that techniques like ours can be developed which are demonstrably more effective than a simple keyword matching approach.

6 Conclusions

The purpose of conducting the experiments reported in this paper was not to develop the ultimate retrieval strategy for image captions though our results do show an approach which would be worth pursuing for this application. Yet retrieval performance compared to other caption-based retrieval approaches is difficult to define and we do not attempt to do that. In conducting our experiments we have shown that the notion of pre-computing word-word similarities based on a WordNet-derived knowledge base and using these in an information retrieval application, is worthwhile though it remains to be seen whether the improvements can

[4] http://www.qbic.almaden.ibm.com/~qbic/qbic.html

[5] http://www.virage.com/online

[6] http://xrs.com/products/image.intro.html

be maintained in scaling up, not the number of documents but to the greater number of *average* terms per document.

In performing our experiments we neatly circumvented the problems created by having to do word sense disambiguation on documents. In our case we manually disambiguated both queries and captions but in practice for larger applications, manual sense disambiguation of documents is not practical yet automatic methods are not as accurate as manual. Our approach to this is to index documents not by uniquely defined sets or bags of index terms or of word senses but by sets of fuzzy sets of word senses where each possible sense of a word has a membership value for its inclusion in the set based on output from an automatic sense disambiguator. For a document about banking and finance we might have a value of 0.45 for the *"Bank: financial institution that accepts, deposits and channels money into lending activities"* sense and 0.40 for the *"Bank: building in which commercial banking is transacted"* but 0.15 for the *"Bank: stock of something from which it is possible to draw"*. This is the focus of our current work.

References

[Bessner, 1990] Bessner, H. (1990). Visual Access to Visual Images: The UC Berkeley Image database project. *Library Trends*, 38(4).

[Belkin et al., 1995] Belkin, N.J., Kantor, P. & Shaw, J.A. (1995). Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management*, 31(3), 431-, 1995.

[Cawkell, 1993] Cawkell, A.E. (1993). An Introduction to Image Processing and Picture Management. *Journal of Document and Text Management*, 1(1).

[Croft, 1995] Croft, W.B. (1995). Text and Multimedia: Possibilities and Challenges. *Presented at the MIRO Workshop*, Glasgow, Scotland.

[Efthimiadis, 1995] Efthimiadis, (1995). User Choices: A New Yardstick for the Evaluation of Ranking Algorithms for Interactive Query Expansion. *Information Processing & Management*, 31(4), 605-620.

[Enser, 1993] Enser, P.G.B. (1993). Query Analysis in a Visual Information Retrieval Context. *Journal of Document and Text Management*, 1(1).

[Flank et al., 1995] Flank, S., Martin, P., Balogh, A. & Rothey, J. (1995). PhotoFile: A Digital Library for Image Retrieval. In. *Proc. International Conference on Multimedia Computing and Systems*, Washington D.C., 292-295.

[Harman, 1995] Harman, D. (1995). The TREC Conferences. In: *Proc. HIM'95 (Hypertext - Information Retrieval - Multimedia)*, Konstanz, 9-23.

[Miller et al., 1990] Miller, G.A., Beckwith, R., Felbaum, C., Gross, D. & Miller, K., (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), 235 - 244.

[Niblack, 1993] Niblack, W. *et al.* (1993) The QBIC Project: Querying Imagesby Content Uing Color, Texture and Shape. *Storage and Retrieval of Image and Video Databases*, SPIE Vol 1908, 173-187.

[Resnick, 1993a] Resnick, P. (1993) Selection and Information: A Class Based Approach to Lexical Relationships. *PhD dissertation, University of Pennsylvania.*

[Resnick, 1993b] Resnick, P. (1993) Semantic Classes and Syntactic Ambiguity. *ARPA Workshop on Human Language and Technology*, Princeton University, USA.

[Richardson, 1994] Richardson, R. (1994). A Semantic Based Approach to Information Processing. *PhD thesis, Dublin City University.*

[Richardson et al., 1994] Richardson, R., Smeaton, A.F. & Murphy, J. (1994). Using WordNet for Conceptual Distance Measurement. In: *Proc. BCS-IRSG Colloquium*, Glasgow, Scotland.

[Richardson et al., 1995] Richardson, R. & Smeaton, A.F. (1995). Using WordNet in a Knowledge-Based Approach to Information Retrieval. *Dublin City University School of Computer Applications Working Paper CA-0395.*

[Richardson et al., 1996] Richardson, R. & Smeaton, A.F. (1996). An Information Retrieval Approach to Locating Information in Large Scale Federated Database Systems. In: *Proc. 2nd International workshop on Applications of Natural Language to Information Systems*, Amsterdam, The Netherlands.

[Rowe et al., 1993] Rowe, N.C. & Guglielmo, E.J. (1993). Exploiting Captions in Retrieval of Multimedia Data. *Information Processing & Management*, 29(4), 453-461.

[Smeaton et al., 1996] Smeaton, A.F., Kelledy, F. & O'Donnell, R. (1996). Thresholding Postings Lists, Query Expansion by Word-Word Distances and POS Tagging of Spanish Texts. In: *Proceedings of TREC-4 (to appear).*

[Smeaton et al., 1983] Smeaton, A.F. & van Rijsbergen, C.J. (1983). The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. *The Computer Journal*, 26(3), 239-246.

[Smeaton, 1992] Smeaton, A.F. (1992). Progress in the Application of Natural Language Processing to Information Retrieval Tasks. *The Computer Journal*, 35(3), 268-278.

[Tversky, 1977] Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84(4), 327-352.

[Voorhees, 1993] E.M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In. Proc. 16th ACM-SIGIR Conference, Pittsburgh (USA), June 1993, 171-180.

Appendix A: Some of the image captions[7]

- A cannon on a battlefield
- A canoe on a lake beside a boathouse
- Fisherman in a canoe on a lake with mountain_peak behind
- Map of Canterbury golf_course
- A duck on the water
- Map of a national_park in California

[7] Note the automatic determination of collocations as in "mountain_peak", "golf_course", etc. in both captions and queries.

- Pathway to a house on the beach on Cape_Cod
- Fishing_boats in a harbour by a village with the mountains behind
- A space_capsule in outer_space
- Twilight in Washington
- The smiling captain of a fishing_boat with a bridge behind
- A photograph of Capitol_Hill in Washington
- A red Porche by a balcony
- A cartoon of a dragon and a gladiator in the mind of a computer programmer
- A carver with a white beard
- A brunette model in a red hat and a swimsuit at a car wash
- A female clown painting a boy with war_paint
- A castle beside gardens with a fountain
- A castle with a moat and a drawbridge

Appendix B: Some of the queries used

The title of the motivating magazine article agreed between user and us is given first, followed by the text of some of the queries we collected to illustrate that particular article:

- "Danger in motor sport"
  - race_car driving on a race_track
  - motor_car crashes
  - spectators at a race_track
- "Renovation of buildings because of pollution, what is the point ?"
  - buildings and offices of historical importance
  - traffic in a busy city street
  - person working on scaffolding beside building
- "Fun with watersports"
  - people waterskiing
  - sunbathers on the beach

There is an online demo of the image caption-based retrieval system available by following a link from http://www.compapp.dcu.ie/~asmeaton

Acknowledgements

Some of the work reported here was funded by EC structural funds through the National Software Directorate.