Query and Document Translation by Automatic Text Categorization: A Simple Approach to Establish a Strong Textual Baseline for ImageCLEFmed 2006

Julien Gobeill, Henning Müller and Patrick Ruch
University and Hospitals of Geneva, Switzerland
{julien.gobeill;henning.mueller;patrick.ruch}@sim.hcuge.ch

Abstract

In this paper, we report on the fusion of simple retrieval strategies with thesaural resources in order to perform document and query translation for cross–language retrieval in a collection of medical cases. The collection contains textual and visual contents. In this paper, we focus on the textual contents of the collection, which contains documents in three languages: French, English and German. The fusion of visual and textual content will also be treated. Unlike most automatic categorization systems, which rely on training data in order to infer text–to–concept relationships, our approach can be applied with any controlled vocabulary and does not use any training data. For the 2006 ImageCLEFmed experiments we use the Medical Subject Headings (MeSH), a terminology maintained by the National Library of Medicine and which exists in a dozen languages. The basic idea consists of annotating every textual content of the collection (documents and queries) with a set of MeSH concepts using an automatic text categoriser. Thus, allowing an interlingual mapping between queries and documents. For tuning purposes, the system uses a sample of MEDLINE from the OHSUMED collection. Our results, confirmed that such a simple approach is competitive with best performing cross-language retrieval methods for such a collection. Several simple linear approaches were used to combine textual and visual features

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, Text categorization, multimodal retrieval

1 Introduction

Cross–Language Information Retrieval (CLIR) is increasingly relevant as network–based resources become commonplace. In the medical domain it is of strategic importance in order to fill the gap

between clinical records, written in national languages and research reports massively written in English. Images are also getting increasingly important and varied in the medical domain, and they become available in digital form. Despite the fact that images are language–independent, they are most often accompanied by textual notes in various languages and these textual notes can strongly improve retrieval quality [15]. There are several ways for handling CLIR. Historically, the most traditional approach to IR in general and to multilingual retrieval in particular, uses a controlled vocabulary for indexing and retrieval. In this approach, a librarian selects for each document a few descriptors taken from a closed list of authorised terms. A good example of such a human indexing is found in the MEDLINE database, where records are manually annotated with Medical Subject Headings (MeSH). Ontological relations (synonyms, related terms, narrower terms, broader terms) can be used to help choose the right descriptors, and solve the sense problems of synonyms and homographs. The list of authorised terms and semantic relations between them are contained in a thesaurus. A problem remains, however, since concepts expressed by one single term in one language sometime are expressed by distinct terms in another. We can observe that terminology–based CLIR is a common approach in well–delimited fields for which multilingual thesauri already exist (not only in medicine but also in the legal domain, energy, etc.) as well as in multinational organisations or countries with several official languages. This controlled vocabulary approach is often associated with Boolean–like engines, and it gives acceptable results but prohibits precise queries that cannot be expressed with these authorised keywords. The two main problems are:

- it can be difficult for users to think in terms of a controlled vocabulary. Therefore, the use of these systems – like most Boolean–supported engines — is often performed by professionals rather than general users;
- this retrieval method ignores the free–text portions of documents during indexing.

A detailed task description on the medical image retrieval task can be found in [13]. The non–medical tasks of ImageCLEF are described in [4].

1.1 Translation–based Approaches

A second approach to multilingual interrogation is to use existing machine translation (MT) systems to automatically translate the queries [5] or even the entire textual database [16] [12] from one language to another, thereby transforming the CLIR problem into a mono–lingual information retrieval (MLIR) problem.

This kind of method would be satisfactory if current MT systems did not make errors. A certain amount of syntactic error can be accepted without disturbing results of information retrieval systems but MT errors in translating concepts can prevent relevant documents, indexed on the missing concepts, from being found. For example, if the word *traitement* in French is translated by *processing* instead of *prescription*, the retrieval process would yield wrong results. This drawback is limited in MT systems that use huge transfer lexicons of noun phrases by taking advantage of frequent co–locations to help disambiguation but in any collection of text ambiguous nouns will still appear as isolated noun phrases untouched by this approach.

1.2 Using Parallel Resources

A third approach receiving increasing attention is to automatically establish associations between queries and documents independent of language differences. Seminal researches were using latent semantic indexing [6]. The general strategy when working with parallel or comparable texts is the following: if some documents are translated into a second language these documents can be observed both in the subspace related to the first language and the subspace related to the second one; using a query expressed in the second language, the most relevant documents in the translated subset are extracted (usually using a cosine measure of proximity). These relevant documents are in turn used to extract close untranslated documents in the subspace of the first language. This approach use implicit dependency links and co–occurrences that better approximate the notion of

concept. Such a strategy has been tested with success on the English-French language pair using a sample of the Canadian Parliament bilingual corpus. It is reported that for 92% of the English text documents the closest document returned by the method was its correct French translation. Such an approach presupposes that the sample used for training is representative of the full database, and that sufficient parallel/comparable corpora are available or acquired.

Other approaches are usually based on bilingual dictionaries and terminologies, sometimes combined with parallel corpora. These approaches attempt to infer a word by word transfer function: they typically begin by deriving a translation dictionary, which is then applied to query translation. Of interest for comparison with our experiments, the study reported by Eichmann and al. [7] uses the OHSUMED collection and relies on a transfer lexicon built from the Unified Medical Language System (UMLS[1]). Finally, [19] described a system which combines thesaurus–based approaches and machine translation for translating queries in a monolingual collection.

To synthesise, we can consider that the performance of CLIR systems typically ranges between 60% and 90% of the corresponding monolingual run [23]. CLIR ratios above 100% have been reported [28], however, such results were obtained by computing a weak monolingual baseline.

2 Data and Strategies

The 2006 imageCLEF collection contains a set of 50'000 medical images accompanied by textual reports (mainly pathology and radiology) that are all well-structure in XML. Some reports describe an entire case containing several images and other reports describe a single image. In this article we focus on experiments made on the textual part. A short description will also describe the visual and multi–modal retrieval parts. The text of the collection contains over 40'000 textual documents in three languages: French, English and German. There are 30 English queries, translated into French and German. Because the document collection is multilingual, we decided to map each document and each query to a set of MeSH concepts, thus transforming a fairly usual text retrieval task into a category retrieval task.

Soergel describes a general framework for the use of multilingual thesauri in CLIR [25], noting that a number of operational European systems employ multilingual thesauri for indexing and searching. However, except for very early work [21], there has been little empirical evaluation of multilingual thesauri in the context of free–text CLIR, particularly when compared to dictionary and corpus–based methods. This may be due to the cost of constructing multilingual thesauri, but this cost is unlikely to be more than that of creating bilingual dictionaries or even realistic parallel collections. It seems that multilingual thesauri can be built quite effectively by merging existing monolingual thesauri, as shown by the current development of the Unified Medical Language System (UMLS) or the SemanticMining Multilingual Lexicon[3].

Our approach to CLIR in MEDLINE exploits the UMLS resources and its multilingual components. The core technical component of our cross–language engine is an automatic text categoriser, which associates a set of MeSH terms to any input text. The experimental design is the following:

1. each document and all queries of the imageCLEF collection are annotated by our automatic text categoriser, which contains MeSH in French, English, and German;

2. each query is annotated by three MeSH categories and we evaluate the impact of varying the number of categories for the document collection: three, five and eight categories are tested;

3. the MeSH annotated imageCLEF document collection is indexed using a standard engine.

2.1 MeSH–driven Text Categorization

Automatic text categorization has been studied largely and has led to an impressive amount of papers. A partial list[2] of machine learning approaches applied to text categorization includes naive

[1] See http://umlsks.nlm.nih.gov.

[2] See http://faure.iei.pi.cnr.it/~fabrizio/ for an updated bibliography.

Bayes [11], k–nearest neighbours [29], boosting [22], and rule–learning algorithms [1]. However, most of these studies apply text classification to a small set of classes; usually a few hundred, as in the Reuters collection [8]. In comparison to this our system is designed to handle large class sets [20]: retrieval tools used are only limited by the size of the inverted file, but 10^{5-6} documents is still a modest range [3].

Our approach is data–poor because it only demands a small collection of annotated texts for fine tuning: instead of inducing a complex model using large training data, our categoriser indexes the collection of MeSH terms as if they were documents and then it treats the input as if it was a query to be ranked regarding each MeSH term. The classifier is tuned by using English abstracts and English MeSH terms. Then, we apply the system on the ImageCLEFmed collection. For tuning the categoriser, the top 15 returned terms are selected because it is the average number of MeSH terms per abstract in the OHSUMED collection. When applied on the ImageCLEFmed collection, the number of categories to be attached to every document will be an important parameter.

2.2 Collection and Metrics

The mean average precision (MAP): is the main measure for evaluating *ad hoc* retrieval tasks (for both monolingual and bilingual runs). Following [9], we also use this measure to tune the automatic text categorization system. We tune the categorization system on a small set of OHSUMED abstracts: 1200 randomly selected abstracts were used to select the weighting parameters of the vector space classifier and the best combination of these parameters with the regular expression–based classifier.

2.3 Visual retrieval techniques

The technology used for the visual retrieval of images is mainly taken from the *Viper*[4] project. Much information about this system is available [26]. Outcome of the *Viper* project is the GNU Image Finding Tool, *GIFT*[5]. This tool is open source and can be used by other participants of ImageCLEF. A ranked list of visually similar images for every query topic was made available for participants and will serve as a baseline to measure the quality of submissions. Feature sets used by *GIFT* are:

- Local color features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode color of each region as a descriptor;
- global color features in the form of a color histogram, compared by a simple histogram intersection;
- local texture features by partitioning the image and applying Gabor filters in various scales and directions, quantised into 10 strengths;
- global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Visual features are quantised and the feature space is similar to the distribution of words in texts. A simple *tf/idf* weighting is used and the query weights are normalised by the results of the query itself. The histogram features are compared based on a histogram intersection [27].

[3] In text categorization based on learning methods, the scalability issue is twofold: it concerns both the ability of these data–driven systems to work with large concept sets, and their ability to learn and generalise regularities for rare events: [9] shows how the frequency of concepts in the collection is a major parameter for learning systems.

[4] `http://viper.unige.ch/`

[5] `http://www.gnu.org/software/gift/`

System or parameters	Relevant retrieved	Prec. at Rec. = 0	MAP
RegEx	3986	.7128	.1601
lnc.atn	3838	.7733	.1421
anc.atn	3813	.7733	.1418
ltc.atn	3788	.7198	.1341
ltc.lnn	2946	.7074	.111

Table 1: Categorization results. For the VS engine, *tf.idf* parameters are provided: the first triplet indicates the weighting applied to the "document", i.e. the concept, while the second is for the"query", i.e. the abstract. The total number of relevant terms is 15193.

3 Methods

In this section, we present the basic classifiers and their combination for the categorization task. Three main modules constitute the skeleton of our system: the regular expression (RegEx) component, the vector space (VS) component, and the visual retrieval part. Each of the basic classifiers implements known approaches to document retrieval. The first tool is based on a regular expression pattern matcher [10], it is expected to perform well when applied on very short documents such as keywords: MeSH terms do not contains more than 5 tokens. The second classifier is based on a vector space engine[6]. This second tool is expected to provide high recall in contrast to the regular expression–based tool, which should privilege precision. The former component uses tokens as indexing units and can be merged with a thesaurus, while the latter uses stems (Porter). Table 1 shows the results of each classifier.

Regular expressions and MeSH thesaurus. The regular expression search tool is applied on the canonic MeSH collection augmented with the MeSH thesaurus (120'020 synonyms). In this system, string normalisation is mainly performed by the MeSH terminological resources when the thesaurus is used. Indeed, the MeSH provides a large set of related terms, which are mapped to a unique MeSH representative in the canonic collection. The related terms gather morpho-syntactic variants, strict synonyms, and a last class of related terms, which mixes up generic and specific terms: for example, *Inhibition* is mapped to *Inhibition (Psychology)*. The system cuts the abstract into 5–token–long phrases and moves the window through the abstract: the edit–distance is computed between each of these 5 token sequences and each MeSH term. Basically, the manually crafted finite–state automata allow two insertions or one deletion within a MeSH term, and ranks the proposed candidate terms based on these basic edit operations: insertion costs 1, while deletion costs 2. The resulting pattern matcher behaves like a term proximity scoring system [17], but is restricted to a 5–token matching window.

Vector space classifier. The vector space module is based on a general IR engine with the *tf.idf*[7] weighting schema. The engine uses a list of 544 stop words.

As for setting the weighting factors, we observed that cosine normalisation was especially effective for our task. This is not surprising, considering the fact that cosine normalisation performs well when documents have a similar length [24]. As for the respective performance of each basic classifier, table 1 shows that the RegEx system performs better than any *tf.idf* schema used by the VS engine, so the pattern matcher provides better results than the vector space engine for automatic text categorization. However, we also observe in table 1 that the VS system gives better precision at high ranks ($Precision_{at\ Recall=0}$ or *mean reciprocal rank*) than the RegEx system: this difference suggests that merging the classifiers could be effective. The *idf* factor also seems to be an important parameter. As shown in table 1 the four best weighting schemas use the *idf*

[6]The easyIR engine, and the automatic categorization toolkit are available on the author's pages.

[7]We use the SMART representation for expressing statistical weighting factors: a formal description can be found in [18].

Weighting function concepts.abstracts	Relevant retrieved	Prec. at Rec. = 0	MAP
Hybrids: tf.idf + RegEx			
ltc.lnn	4308	.8884	.1818
lnc.lnn	4301	.8784	.1813
nc.ntn	4184	.8746	.1806
anc.ntn	4184	.8669	.1795
atn.ntn	3763	.9143	.1794

Table 2: Combining VS with RegEx.

factor. This observation suggests that even in a controlled vocabulary, the *idf* factor is able to discriminate between content– and non–content–bearing features (such as *syndrome* and *disease*).

Classifier fusion. The hybrid system combines the regular expression classifier with the vector–space classifier. Unlike [9] we do not merge our classifiers by linear combination, because the RegEx module does not return a scoring consistent with the vector space system. Therefore the combination does not use the RegEx's edit distance, and instead it uses the list returned by the vector space module as a *reference* list (RL), while the list returned by the regular expression module is used as *boosting* list (BL), which serves to improve the ranking of terms listed in RL. A third factor takes into account the length of terms: both the number of characters (L_1) and the number of tokens (L_2, with $L_2 > 3$) are computed, so that long and compound terms, which appear in both lists, are favoured over single and short terms. We assume that the reference list has good recall, and we do not set any threshold on it. For each concept t listed in the RL, the combined Retrieval Status Value ($cRSV$, equation 1) is:

$$cRSV_t = \begin{cases} RSV_{VS}(t) \cdot Ln(L_1(t) \cdot L_2(t) \cdot k) & \text{if } t \in BL, \\ RSV_{VS}(t) & \text{otherwise.} \end{cases} \quad (1)$$

The value of the k parameter is set empirically. Table 2 shows that the optimal *tf.idf* parameters (*lnc.atn*) for the basic VS classifier does not provide the optimal combination with RegEx. Measured by MAP. The optimal combination is obtained with *ltc.lnn* settings (.1818) [8], whereas *atn.ntn* maximises the $Precision_{at\ Recall=0}$ (.9143).

3.1 Cross–Language Categorization and Indexing

To translate the ImageCLEFmed textual content (queries and documents), we transform the English MeSH mapping tool described above, attributes MeSH terms to English abstracts. Thus, the English, French, and German version of the MeSH are simply merged in the categoriser. We use the weighting schema and system combination (ltc.lnn + RegEx) as described above. Then, the annotated collection is indexed using the vector–space engine used by the categoriser. For the document indexing, we rely on weighting schemas based on pivoted normalisation: because the documents have a very variable length in the collection such a factor can be important. A slightly modified version of dtu.dtn [2], which has shown some effectiveness for the TREC Genomics, is used for full–text indexing and retrieval. The English stop word list is merged with a French and a German stop word list. Porter stemming is used for all documents.

3.2 Visual and multimodal retrieval

For the visual retrieval, one quantisation of four grey levels was used that has shown to be efficient [14]. To combine visual and textual runs we choose English as the main language and a number of

[8] For the augmented term frequency factor (noted a, which is defined by the function $\alpha + \beta \times (tf/max(tf))$, the value of the parameters is $\alpha = \beta = 0.5$.

five terms. Visual inspection of some results indicated that this might lead to good results. The combination is simply done linearly by normalising the output of visual and textual results and then adding them up in various ratios. A second approach for a multimodal combination was to take the results from the visual retrieval side and increase the value of those results in the first 1000 images that also appear in the visual results by simple re–ranking.

4 Results and Discussion

We submitted several runs including runs combining textual and visual features. The visual runs have a low overall performance but do perform well on the visual topics. The mixed runs had a problem in the combination part and are in large part broken. The text retrieval was based on the cases and thus needed to be expanded towards

4.1 Textual retrieval results

The runs were generated with the very same strategy but using respectively, eight, five and three MeSH categories to annotate the document collection. The runs were generated automatically and do not use visual features. For each of them, queries were expanded using the top three MeSH categories provided by the categoriser. Following previous experiments dedicated to query translation [19], the optimal number of categories for query translation is around two or three.

Run	MAP	Run	MAP	Run	MAP
GE-8EN	0.2255	GE-8DE	0.0574	GE-8FR	0.0417
GE-5EN	0.1967	GE-5DE	0.0416	GE-5FR	0.0346
GE-3EN	0.1913	GE-3DE	0.0433	GE-6FR	0.0323

Table 3: MAP of textual runs.

Evaluations are computed by retrieving the first 1'000 documents for each query. In Table 3, we observe that the maximum MAP is reached when eight MeSH terms are selected per document. This result suggests that a large number of categories can be selected to annotate a document, although it must be observed that the precision of the system is low beyond the top one or two categories. It means that annotating a document with several potentially irrelevant concepts does not hurt the matching power of our interlingual concepts! This result is somehow consistent with known observations made on query expansion: a certain number of inappropriate expansion is acceptable and still can improve retrieval effectiveness of modern search engines. This statement should be emphasised in the case of the ImageCLEFmed collection, which can be regarded as a small collection (about 40'000 cases for 50'000 images), where recall plays a more important role than in larger collections, which can rely on information redundancy with a less important role for recall.

The English retrieval results were the second best group results of all participants. For languages other than English it seems to be much harder to obtain very good results as the majority of documents in the collection is in English.

4.2 Visual and multi-modal runs

Table 4 shows the results for our visual run and the best mixed runs. The visual run is performance–wise in the middle of the submissions and the best purely visual runs are approximately 30% better. GIFT performance better in early precision than other systems with a higher MAP. For the visual topics the results are very satisfactory whereas semantic topics do not perform well.

One big problem shows up in all the mixed runs submitted. They are not as good as the underlying textual runs even when only a very small percentage of visual information is used. One problem that might have caused this is the use of a wrong file for the textual runs. The English

Run	MAP
GE-GIFT	0.0467
GE-vt10	0.12
GE-vt20	0.1097

Table 4: MAP of visual runs.

runs perform much better than French and German runs, and so mixing up these two can cause such trouble. We need to further investigate into this to find the reasons and allow for better multimodal image retrieval.

5 Conclusion and Future Work

We have presented a cross language information retrieval engine for the ImageCLEFmed image retrieval task, which capitalises from the availability of a multilingual controlled vocabulary to translate user requests and documents. The system relies on a text categoriser, which maps queries into a set of predefined concepts. For the ImageCLEFmed 2006 collection, optimal precision is obtained when selecting three MeSH categories per query and eight MeSH categories per documents, whereas a larger number leads to best MAP. Visual retrieval shows to work well on the visual topics but fails on the semantic topics. Further experiments are needed to determine the best number of concepts. The use of a dynamic threshold will be evaluated in the future. Another problem is the combination of visual and textual features that really needs further analysis beyond simple linear combinations.

Acknowledgements

The study has also been partially supported by the Swiss National Foundation (Grants 3200–065228 and 205321–109304/1), the European Union (SemanticMining Network of Excellence, INFS–CT–2004–507505) via an OFES Grant (No 03.0399, cf. http://www.genisis.ch/~natlang/semm/).

References

[1] C Apté, F Damerau, and S Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251, 1994.

[2] A Aronson, D Demner-Fushman, S Humphrey, J Lin, H Liu, P Ruch, M Ruiz, L Smith, L Tanabe, and J Wilbur. Fusion of Knowledge-intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents. In *TREC 2005*, 2006.

[3] R Baud, M Nystrom, L Borin, R Evans, S Schulz, and P Zweigenbaum. Interchanging lexical information for a multilingual dictionary. *AMIA Symposium Proceedings*, 2005.

[4] P Clough, M Grubinger, T Deselaers, A Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[5] M Davis. Free resources and advanced alignment for cross-language text retrieval. In *In proceedings of The Sixth Text Retrieval Conference (TREC6)*, 1998.

[6] S Dumais, T Letsche, M Littman, and T Landauer. Automatic cross-language retrieval using latent semantic indexing. In D Hull and D Oard, editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[7] D Eichmann, M Ruiz, and P Srinivasan. Cross–language information retrieval with the UMLS metathesaurus. In *SIGIR Conference*, pages 72–80, Melbourne, Australia, 1998.

[8] P Hayes and S Weinstein. A system for content-based indexing of a database of news stories. *Proceedings of the Second Annual Conference on Innovative Applications of Intelligence*, 1990.

[9] L Larkey and W Croft. Combining classifiers in text categorization. In *SIGIR*, pages 289–297. ACM Press, New York, US, 1996.

[10] U Manber and S Wu. GLIMPSE: A tool to search through entire file systems. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 23–32, San Fransisco CA USA, 17-21 1994.

[11] A McCallum and K Nigam. A comparison of event models for naive bayes text classification, 1998.

[12] J McCarley. Should we translate the documents or the queries in cross-language information retrieval. *ACL*, 1999.

[13] H Müller, T Deselaers, T Lehmann, P Clough, and W Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[14] Henning Müller, Antoine Geissbuhler, and Patrick Ruch. ImageCLEF 2004: Combining image and multi–lingual search for medical image retrieval. In *Cross Language Evaluation Forum (CLEF 2004)*, Springer Lecture Notes in Computer Science (LNCS), Bath, England, 2005.

[15] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine – clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23, 2004.

[16] D Oard and P Hackett. Document translation for cross-language text retrieval at the university of Maryland. In *In Proceedings of The Sixth Text Retrieval Conference (TREC6)*, 1998.

[17] Y Rasolofo and J Savoy. Term proximity scoring for keyword-based retrieval systems. In *ECIR*, pages 101–116, 2003.

[18] P Ruch. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. *COLING 2002*, 2002.

[19] P Ruch. Query translation by text categorization. *COLING 2004*, 2004.

[20] P Ruch, R Baud, and A Geissbühler. Learning-Free Text Categorization. *LNAI 2780*, pages 199–208, 2003.

[21] G Salton. Automatic processing of foreign language documents. *JASIS*, 21(3):187–194, 1970.

[22] R Schapire and Y Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[23] P Schäuble and P Sheridan. Cross-language information retrieval (CLIR) track overview. In *In Proceedings of The Sixth Text Retrieval Conference (TREC6)*, 1998.

[24] A Singhal, C Buckley, and M Mitra. Pivoted document length normalization. *ACM-SIGIR*, pages 21–29, 1996.

[25] D Soergel. Multilingual thesauri in crosslanguage text and speech retrieval. In D Hull and D Oard, editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[26] David McG. Squire, Wolfgang Müller, Henning Müller, and Thierry Pun. Content–based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13-14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

[27] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[28] J Xu, A Fraser, and R Weischedel. Cross-lingual retrieval at bbn. In *TREC*, 2001.

[29] Y Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1:67–88, 1999.