Translation by Text Categorisation: Medical Image Retrieval in ImageCLEFmed 2006

Julien Gobeill, Henning Müller, Patrick Ruch

Medical Informatics, University and Hospitals of Geneva, Switzerland
julien.gobeill;henning.mueller;patrick.ruch}@sim.hcuge.ch

Abstract. We present the fusion of simple retrieval strategies with thesaural resources to perform document and query translation by text categorisation for cross–language retrieval in a collection of medical images with case notes. The collection includes documents in French, English and German. The fusion of visual and textual content is also treated. Unlike most automatic categorisation systems our approach can be applied with any controlled vocabulary and does not require training data. For the experiments we use Medical Subject Headings (MeSH), a terminology maintained by the National Library of Medicine existing in 12 languages. The idea is to annotate every text of the collection (documents and queries) with a set of MeSH terms using our automatic text categoriser. Our results confirm that such an approach is competitive. Simple linear approaches were used to combine text and visual features

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 to fill the gap between clinical records in national languages and research reports in English. Images are also getting increasingly important and varied in the medical domain, and they become available in digital form. Despite images being language–independent, they are most often accompanied by textual notes in various languages and these notes can improve retrieval quality [1]. A task description on the medical image retrieval task can be found in [2].

2 Strategies

1. each document and query are annotated by our automatic text categoriser, which contains MeSH in French, English, and German;
2. each query is annotated by 3 terms and each document by 3, 5, 8 categories;

2.1 Terminology–driven Text Categorisation

Automatic text categorisation has been studied largely and led to a large amount of papers. Approaches include naive Bayes, k–nearest neighbours, boosting, rule–learning algorithms. However, most of these studies apply text classification to

a small set of classes. In comparison to this our system is designed to handle large class sets [3], since such retrieval tools used are only limited by the size of the inverted file, but 10^{5-6} documents is a modest range.

Our approach is data–poor because it only demands a small collection of annotated texts for tuning: instead of inducing a complex model using much training data, our categoriser indexes the collection of MeSH (Medical Subject Headings) 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 MeSH terms. 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.

2.2 Regular expressions and MeSH thesaurus

The regular expression search tool is applied on the canonic MeSH collection augmented with a list of MeSH synonyms. In this system, string normalisation is mainly performed by the MeSH terminological resources when the thesaurus is used. 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 [4], but is restricted to a 5–token matching window.

2.3 Vector–space classifier

The vector–space (VS) module is based on a general IR engine with the *tf.idf* weighting schema. The engine uses a list of 544 stop words. For setting the weighting factors, we observed that cosine normalisation was effective for our task. This is not surprising because cosine normalisation performs well when documents have a similar length [5]. For the respective performance of each basic classifier, the RegEx system performs better than any *tf.idf* schema, so the pattern matcher provides better results than the VS engine. However, we also observe 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.

2.4 Classifier fusion

The hybrid system combines the regular expression classifier with the VS classifier. Unlike [6] we do not merge our classifiers by linear combination, because

the RegEx module does not return a scoring consistent with the VS system. Therefore, the combination does not use the RegEx's edit distance, and instead uses the list returned by the VS 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 term length: 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. 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.

2.5 Cross–Language Categorisation and Indexing

To translate the ImageCLEFmed text, we use the English MeSH categorisation tool. French, and German versions of the MeSH are simply merged in the categoriser. We use the weighting schema combination (ltc.lnn + RegEx). Then, the annotated collection is indexed using the VS engine. For document indexing, we rely on weighting schemas based on pivoted normalisation: as documents have a very variable length such a factor can be important. A slightly modified version of dtu.dtn [7] 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.

2.6 Visual and Multimodal Retrieval

Visual retrieval is mainly based on *GIFT*[1] [8]. Features used by *GIFT* are:

- Local color features at various scales by partitioning the images successively into four equally sized regions and taking the mode color of each region;
- global color features in the form of a color histogram, compared by a 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.

The feature space is similar to the distribution of words in texts. A *tf/idf* weighting is used and the query weights are normalised by the results of the query.

To combine visual and textual runs we choose English as language and a number of five terms based on visual observations. The combination is done by normalising the output of visual and textual results and adding them up in various ratios. A second approach for a multimodal combination was to take 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.

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

3 Results and Discussion

3.1 Textual Retrieval Results

The runs were generated using respectively, eight, five and three MeSH terms to annotate the collection. In previous experiments on query translation [9], the optimal number was 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 1. MAP of textual runs.

Results are computed by retrieving 1'000 documents for per query. In Table 1, we observe that the maximum MAP is reached when eight MeSH terms are selected per document. This suggests that a larger number can be selected to annotate a document, although it must be observed that the precision of the system is low beyond the top ranked (one or two) categories. This means that annotating a document with several potentially irrelevant concepts does not hurt the matching power! This result is consistent with known observations made on query expansion: some inappropriate expansion is acceptable and can still improve retrieval effectiveness. The English retrieval results were the second best results of all participants[2]. For other languages it seems to be much harder to obtain good results as the majority of documents is in English.

3.2 Visual and Multi–Modal Runs

Table 2 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 performs 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.

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

Table 2. MAP of visual runs.

A problem shows up in all mixed runs submitted. They are worse than the underlying textual runs even when only a small fraction of visual information

[2] Some groups combined results of three languages, which improved results.

is used. A possible problem is the use of a wrong file for the text runs. English runs perform much better than French and German runs. We need to further investigate to find the reasons and allow for better multimodal image retrieval.

4 Conclusion and Future Work

We presented a cross–language information retrieval engine for the ImageCLEFmed retrieval task, which uses 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 ImageCLEFmed 2006 optimal precision is obtained when selecting three MeSH terms per query and eight per document, whereas a larger number might even improve results further. Visual retrieval shows to work well on visual topics but fails on semantic topics. A problem is the combination of visual and textual features that needs further analysis and best an analysis of the query to find out more about the search goal.

Acknowledgements

The study was supported by the Swiss National Foundation (3200–065228, 205321–109304/1) and the EU (SemanticMining NoE, INFS–CT–2004–507505).

References

1. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine – clinical benefits and future directions. Internation Journal of Medical Informatics **73** (2004) 1–23
2. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Eugene, K., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2007 – to appear)
3. Ruch, P.: Automatic Assignment of Biomedical Categories: Toward a Generic Approach. Bioinformatics **6** (2006) 658–664
4. Rasolofo, Y., Savoy, J.: Term proximity scoring for keyword-based retrieval systems. In: European Conference on Information Retrieval. (2003) 101–116
5. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedigns of the ACM SIGIR Conference. (1996) 21–29
6. Larkey, L., Croft, W.: Combining classifiers in text categorization. In: Proceedings of the ACM SIGIR Conference, ACM Press, New York, US (1996) 289–297
7. Aronson, A., Demner-Fushman, D., Humphrey, S., Lin, J., Liu, H., Ruch, P., Ruiz, M., Smith, L., Tanabe, L., Wilbur, J.: Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In: Proceedings of TREC 2005. (2006)
8. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content–based query of image databases: inspirations from text retrieval. Pattern Recognition Letters **21** (2000) 1193–1198
9. Ruch, P.: Query translation by text categorization. In Kaufmann, A.M., ed.: Proceedings of COLING 2004. (2004) 686–692