MIRACLE at ImageCLEFmed 2008: Evaluating Strategies for Automatic Topic Expansion

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3], José C. González-Cristóbal[1,3]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

slana@diatel.upm.es, jvillena@it.uc3m.es, josecarlos.gonzalez@upm.es

Abstract

This paper describes the participation of MIRACLE research consortium at the ImageCLEFmed task of ImageCLEF 2008. The main goal of our participation this year is to compare among different topic expansion approaches: methods based on linguistic information such as thesauri or knowledge bases, and statistical techniques based on term frequency. Thus we focused on runs using text features only. First a common baseline algorithm was used in all experiments to process the document collection: text extraction, medical-vocabulary recognition, tokenization, conversion to lowercase, filtering, stemming and indexing and retrieval. Then this baseline algorithm is combined with different expansion techniques. For the semantic expansion, the MeSH concept hierarchy using UMLS entities as basic root elements was used. The statistical method consisted of expanding the topics using the apriori algorithm. Relevance-feedback techniques were also used.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital libraries. **H.2 [Database Management]**: H.2.5 Heterogeneous Databases; **E.2 [Data Storage Representations]**.

Keywords

Image retrieval, medical domain-specific vocabulary, thesaurus, linguistic engineering, information retrieval, indexing, topic expansion, relevance feedback, ImageCLEF Medical Retrieval Task, ImageCLEF, CLEF, 2008.

1. Introduction

MIRACLE is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise (SME) founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks, including the main bilingual, monolingual and cross lingual tasks as well as in ImageCLEF [5] [6], Question Answering, WebCLEF, GeoCLEF and VideoCLEF (VID2RSS) tracks.

This paper describes our participation in the ImageCLEFmed task of ImageCLEF 2008. In short, the goal of this task is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text [7]. The task organizers provide a list of topic statements (a short textual description explaining the research goal) in English, French and German, and a set of several images for each topic. The objective is to retrieve as many relevant images as possible from the given visual and multilingual topics. ImageCLEFmed 2008 extends the experiments of past editions with a larger database and even more complex queries.

The main goal of our participation this year was to compare among different query expansion techniques using different approaches: methods based on linguistic information such as thesauri or knowledge bases, and statistical techniques based on term frequency. Thus we focused on runs using only text features. All experiments were fully automatic, with no manual intervention.

2. Description of the System

The architecture of our system is composed of four different modules: the textual (text-based) retrieval module, which indexes medical case descriptions in order to search and find the most relevant ones to the text of the topic; the expander module, which performs the expansion of the content of documents and/or topics with related terms using textual or statistical algorithms; the relevance-feedback module, which allows to execute reformulated queries that include the results of an initial seed query; and, finally, the result combination module, which uses OR operator to combine, if necessary, the result lists provided by the previous subsystems**.** Figure 1 gives an overview of the system architecture.

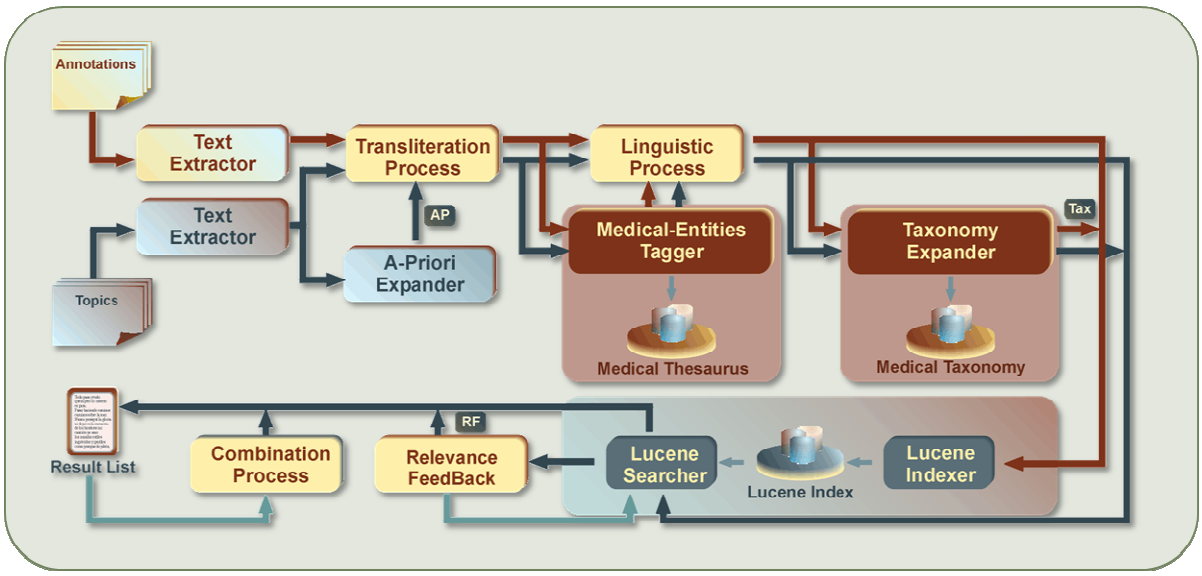

Figure 1. Overview of the system architecture

The system consists of a set of different basic components that can be organized in four categories:

- Resources and tools for medical-specific vocabulary analysis
- Linguistic tools for text analysis and retrieval.
- Sparse matrix based tools for statistical topic expansion and relevance-feedback.
- Tools for the combination of result lists.

Instead of using raw terms, the textual information of both topics and documents is parsed and tagged to unify all terms into concepts of medical entities. This is similar to a stemming or a lemma extraction process, but the output, instead of the stem or lemma, is the medical entity to which the term relates. The result is that concept identifiers [4] are used instead of terms in the text-based process of information retrieval. For this purpose, a terminological dictionary was created by using a subset of the Unified Medical Language System (UMLS) metathesaurus (US National Library of Medicine) [11] containing terms in English, French and German (the three different languages involved in the ImageCLEFmed task [7]). The final version of the dictionary contains 3,211,169 entries matching 1,215,749 medical concepts. **Table 1** shows the language coverage of terms.

Table 1. Language distribution of terms

Lang	#Terms
EN	3,207,890
FR	2,556
DE	723

Notice that there is a significant different in the number of terms among languages. This might bias the results towards the best covered language, English in this case, which has to be taken into account and further analyzed.

A common baseline algorithm was used in all experiments to process the document collection. This algorithm is based on the following sequence of steps:

1. **Text Extraction:** Ad-hoc scripts are run on the files that contain information about the medical cases so as to extract the annotations and metadata enclosed between XML tags.
2. **Medical-vocabulary Recognition:** All case descriptions and topics are parsed and tagged using the UMLS-based terminological dictionary to identify and disambiguate medical terms.
3. **Tokenization:** This process extracts basic textual components, detecting and isolating punctuation symbols. Some basic entities are also detected, such as numbers, initials, abbreviations, and years. So far, compounds, proper nouns, acronyms or other types of entity are not specifically considered. The outcomes of this process are only single words, years in numbers (e.g. 1995, 2004, etc.) and tagged entities.
4. **Conversion to lowercase:** All terms are normalized by changing all uppercase letters to lowercase.
5. **Filtering:** All words recognized as stopwords are filtered out. Stopwords in the target languages were initially obtained from the University of Neuchatel's resources page [9] and afterwards extended using several other sources [3][2] as well as our own developed resources and knowledge base [6].
6. **Stemming:** This process is applied to each one of the terms to be indexed or used for retrieval. Standard Porter stemmers [8] for each considered language have been used.
7. **Indexing and retrieval:** Lucene [2] was used as the information retrieval engine for the whole textual indexing and retrieval task.

This common baseline algorithm is complemented and combined with different expansion techniques in order to compare the improvement given by semantic- versus statistical-based techniques. For the semantic expansion, we used the MeSH concept hierarchy [10] using the UMLS entities detected in document and topics as basic root elements to expand with their hyponyms (i.e., other entities whose semantic range is included within that of the root entity). Semantic expansion was applied to both topics and documents.

The statistical method consisted of expanding the topics using the Agrawal's apriori algorithm [1]. First, a term-document matrix is built using the UMLS entities found in the document corpus. Then apriori algorithm is used to discover out rules having the UMLS entities identified in the topic as antecedent and a confidence value greater than 0.5. Finally, the topic is expanded with the consequent of those (one-term) rules, i.e., UMLS entities that are related to the topic, according to the document corpus.

Finally, relevance-feedback techniques were also used. The top M UMLS entities of each of the top N result documents were extracted and weighted by a factor that is proportional to their document frequency to reformulate a new query that is executed once again to get the final result list.

3. Results

Experiments are defined by the choice of different combinations of the previous modules with the different topic expansion techniques, and including relevance-feedback or not. Table 2 shows the complete list of submitted runs.

Table 2. Description of experiments

Run Identifier	Language	Method
MirBaselineEN	EN	stem + stopwords + tagged with UMLS thesaurus
MirAPEN	EN	baseline + Apriori topic expansion
MirTaxEN	EN	baseline + MeSH topic expansion
MirRF0505EN	EN	baseline + Relevance-Feedback (N=5, M=5)
MirRF1005EN	EN	baseline + Relevance-Feedback (N=10, M=5)
MirRFTax1005EN	EN	baseline + MeSH topic expansion + Relevance-Feedback (N=10, M=5)
MirRFTax1005FR	FR	baseline + MeSH topic expansion + Relevance-Feedback (N=10, M=5)
MirRFTax1005DE	DE	baseline + MeSH topic expansion + Relevance-Feedback (N=10, M=5)

Results are presented in the following table, which shows the run identifier, the number of relevant documents retrieved, the mean average precision (MAP), and the precision at 5, 10, 30 and 100 first results. The best results are highlighted in bold.

Table 3. Results of experiments

	RelRet	MAP	P5	P10	P30	P100
MirBaselineEN	1861	**0.266**	**0.507**	**0.467**	0.390	**0.258**
MirAPEN	1773	0.250	0.487	0.457	**0.393**	0.244
MirTaxEN	**1867**	0.246	0.380	0.373	0.368	0.240
MirRF0505EN	1372	0.105	0.280	0.243	0.241	0.153
MirRFTax1005EN	1260	0.069	0.153	0.130	0.140	0.108
MirRF1005EN	1248	0.071	0.220	0.160	0.149	0.1193
MirRFTax1005DE	461	0.048	0.087	0.090	0.059	0.038
MirRFTax1005FR	823	0.066	0.127	0.107	0.090	0.076

The highest MAP is obtained with the baseline experiment in English. Moreover, MAP values are similar in practice for experiments using topic expansion, and noticeably worse (0.105 against 0.266) in the case of relevance-feedback. This shows that no strategy for either topic expansion or specially relevance-feedback has proved to be useful.

As in previous participation, the value for early precisions (P5, P10) quickly decreases as more documents are considered for the calculation and therefore decreasing the final MAP value. This shows that, although the first results may be appropriate, we probably fail to filter non-relevant documents out of the result list, or perhaps to sort out relevant documents that are "more difficult" to find. Some effort will be invested to research on this issue.

4. Conclusions and Future Work

A preliminary analysis of the results, given the low precision values obtained in the experiments that make use of the relevance-feedback methods, shows that the reranking algorithm used for combining the different result lists is likely to be the main reason for the disappointing results. However, this impression has to be confirmed with a more in-depth analysis. Another probable cause is the choice of the OR operator to combine the terms in the topic to build up the query. Due to time constraints to prepare this report, we were unable to repeat our experiments with the AND operator, but we think that MAP values should be significantly higher using this operator.

In addition, experiments using French and German languages get a very low precision. A possible explanation is that the process of entity unification (detection) for those languages is poor, due to the reduced coverage of the knowledge base. We will try to complete and expand the thesaurus for those languages with other available resources.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-03 and by Madrid R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

[1] Agrawal, Rakesh; Srikan, Ramakrishnan. Fast algorithms for mining association rules. In Proceedings of the International Conference on Very Large Data Bases, pp. 407-419, 1994.

[2] Apache Lucene project. On line http://lucene.apache.org [Visited 10/08/2008].

[3] CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/ ~gjones/CLEF2005/Multi-8/ [Visited 10/08/2008].

[4] González, José C.; Villena, Julio; Moreno, Cristina; Martínez, J.L. Semiautomatic Extraction of Thesauri and Semantic Search in a Digital Image Archive. Integrating Technology and Culture: 10th International Conference on Electronic Publishing, ELPUB 2006, Bansko, Bulgaria, 14-16 June 2006.

[5] Martínez-Fernández, J.L.; Villena-Román, Julio; García-Serrano, Ana M.; Martínez-Fernández, Paloma. MIRACLE team report for ImageCLEF IR in CLEF 2006. Proceedings of the Cross Language Evaluation Forum 2006, Alicante, Spain. 20-22 September 2006.

[6] Martínez-Fernández, J.L.; Villena-Román, Julio; García-Serrano, Ana M.; González-Cristóbal, José Carlos. Combining Textual and Visual Features for Image Retrieval. Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, Vol. 4022, 2006. ISSN: 0302-9743.

[7] ImageCLEF Medical Retrieval Task. On line http://ir.ohsu.edu/image/ [Visited 14/08/2008].

[8] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 10/08/2008].

[9] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 10/08/2008].

[10] U.S. National Library of Medicine. National Institutes of Health. On line. Medical Subject Headings http://www.nlm.nih.gov/mesh/ [Visited 10/08/2008].

[11] U.S. National Library of Medicine. National Institutes of Health. On line. Unified Medical Language System. http://www.nlm.nih.gov/research/umls/ [Visited 10/08/2008].

[12] Villena-Román, Julio; Lana-Serrano, Sara; Martínez-Fernández, José Luis; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.