MIRACLE at ImageCLEFmed 2009: Reevaluating 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 2009. The main purpose of our experiments was to determine if any improvement of the linguistic expansion modules that were developed for the previous CLEF campaign, in terms of precision and recall, was possible. Again, 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 was combined with different semantic expansion techniques. Documents were tagged based on the MeSH concept hierarchy using UMLS entities as basic root elements. Relevance-feedback techniques were also used. Average results were obtained.

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, 2009.

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.

This paper describes our participation in the ImageCLEFmed task of ImageCLEF 2009. 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 [1]. 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, along with their description, for each topic. The objective is to retrieve as many relevant images as possible from the given visual and multilingual topics.

Last campaign, our research goal 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 [2]. Those experiments, in turn, were continuing the research line that was opened in previous campaigns [3] [4]. However, in spite of all our efforts, our best run was the baseline experiment.

Apparently, no strategy for either topic expansion or specially relevance-feedback proved to be useful. However, the post-workshop analysis showed that the main reason for the low precision values obtained in the experiments that included topic expansion techniques was that, in all cases, the OR operator was used to build the reformulated query, i.e., both the original terms and the expanded terms were combined with the OR operator. This implied that documents that contained any of those terms were considered as relevant, no matter if the term belonged to the original topic or it was included in the expansion process.

We concluded that a combination of OR and AND operators should have been used to be sure that documents do contain the original topic terms and, optionally, any of the expanded terms:

(original$_1$ OR expanded$_1$) AND (original$_2$ OR expanded$_2$)

In addition, we found that the reranking algorithm used for combining the different results list could be one of the reasons for the low precision values obtained in the experiments that make use of the relevance-feedback methods.

Thus, the objective of this year's experiments was to try to solve those bugs and be able to analyze and compare the performance in terms of precision and recall of the different query expansion techniques. Again, all runs were based on textual features only. 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 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 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.

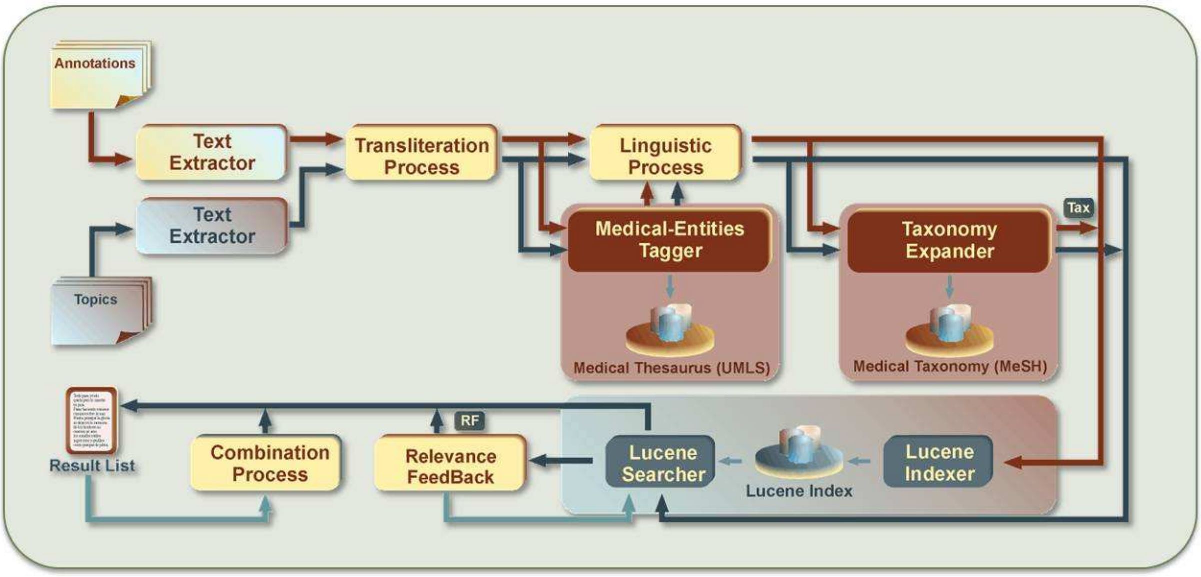

Figure 1. Overview of the system architecture

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

- Resources and tools for medical-specific vocabulary analysis
- Linguistic tools for text analysis and retrieval
- Relevance-feedback tools
- Tools for the combination of result lists

For indexing, 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 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) [5] containing terms in English, French and German (the three different languages involved in the ImageCLEFmed task [1]). 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 in the dictionary.

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 [5] 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 [6] and afterwards extended using our own sources [2].
6. **Stemming:** This process is applied to each one of the terms to be indexed or used for retrieval. Standard Porter stemmers [7] for each considered language have been used.
7. **Indexing and retrieval:** Lucene [8] was used as the information retrieval engine for the whole textual indexing and retrieval task.

This common baseline algorithm is complemented and combined with semantic expansion techniques. For the semantic expansion, we used the MeSH concept hierarchy [9] 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.

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
Mir	EN, FR, DE	stem + stopwords + tagged with UMLS thesaurus (baseline)
MirTax	EN, FR, DE	baseline + MeSH topic expansion
MirRF0505	EN, FR, DE	baseline + Relevance-Feedback (N=5, M=5)
MirRFTax0505	EN, FR, DE	baseline + MeSH topic expansion + Relevance-Feedback (N=5, M=5)
MirEN	EN	stem + stopwords + tagged with UMLS thesaurus (baselineEN)
MirTaxEN	EN	baselineEN + MeSH topic expansion
MirRF0505EN	EN	baselineEN + Relevance-Feedback (N=5, M=5)
MirRF1005EN	EN	baselineEN + Relevance-Feedback (N=10, M=5)
MirRFTax0505EN	EN	baselineEN + MeSH topic expansion + Relevance-Feedback (N=5, M=5)
MirRFTax1005EN	EN	baselineEN + MeSH topic expansion + Relevance-Feedback (N=10, M=5)

Results are presented in the following tables, which show 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. Overall results achieved for all topics are shown in Table 3. Tables 4, 5 and 6 show the individualized results for visual, mixed and semantic topics, respectively.

Table 3. Results of experiments

	RelRet (2362)	MAP	P5	P10	P30	P100
Mir	842	0.150	0.584	0.472	0.307	0.178
MirTax	843	0.129	0.504	0.396	0.288	0.169
MirRF0505	430	0.075	0.432	0.312	0.212	0.099
MirRFTax0505	447	0.054	0.288	0.216	0.167	0.100
MirEN	912	**0.171**	**0.624**	**0.548**	**0.389**	**0.198**
MirTaxEN	**913**	0.165	0.592	0.516	0.376	0.197
MirRF0505EN	567	0.129	0.592	0.512	0.339	0.151
MirRF1005EN	459	0.089	0.536	0.412	0.235	0.106
MirRFTax0505EN	568	0.102	0.448	0.360	0.227	0.142
MirRFTax1005EN	470	0.071	0.408	0.304	0.183	0.101

Table 4. Results of visual experiments (topics 1-9)

	RelRet (954)	MAP	P5	P10	P30	P100
Mir	447	0.210	**0.733**	0.544	0.322	0.231
MirTax	447	0.183	0.578	0.422	0.296	0.214
MirRF0505	259	0.118	0.644	0.444	0.281	0.139
MirRFTax0505	272	0.079	0.333	0.256	0.237	0.152
MirEN	**504**	**0.228**	0.667	**0.578**	**0.433**	**0.237**
MirTaxEN	**504**	0.217	0.600	0.533	0.396	0.228
MirRF0505EN	285	0.163	0.622	0.544	0.356	0.184
MirRF1005EN	234	0.116	0.600	0.456	0.233	0.121
MirRFTax0505EN	285	0.139	0.533	0.422	0.307	0.181
MirRFTax1005EN	241	0.098	0.422	0.367	0.241	0.127

Table 5. Results of mixed experiments (topics 10-20)

	RelRet (938)	MAP	P5	P10	P30	P100
Mir	280	0.121	0.436	0.382	0.321	0.184

MirTax	281	0.096	0.382	0.309	0.300	0.178
MirRF0505	133	0.060	0.309	0.264	0.212	0.102
MirRFTax0505	137	0.043	0.236	0.200	0.145	0.092
MirEN	291	**0.153**	**0.582**	**0.527**	**0.418**	0.230
MirTaxEN	**292**	0.149	0.564	0.491	**0.418**	**0.235**
MirRF0505EN	234	0.128	0.564	0.500	0.397	0.180
MirRF1005EN	175	0.082	0.491	0.427	0.267	0.129
MirRFTax0505EN	235	0.085	0.309	0.255	0.182	0.161
MirRFTax1005EN	179	0.056	0.345	0.255	0.142	0.111

Table 6. Results of semantic experiments (topics 21-25)

	RelRet (455)	**MAP**	**P5**	**P10**	**P30**	**P100**
Mir	115	0.103	**0.640**	**0.540**	**0.247**	**0.104**
MirTax	115	0.103	**0.640**	**0.540**	**0.247**	**0.104**
MirRF0505	38	0.030	0.320	0.180	0.087	0.042
MirRFTax0505	38	0.030	0.320	0.180	0.087	0.042
MirEN	**117**	**0.106**	**0.640**	0.540	**0.247**	**0.104**
MirTaxEN	**117**	**0.106**	**0.640**	**0.540**	**0.247**	**0.104**
MirRF0505EN	48	0.072	0.600	0.480	0.180	0.064
MirRF1005EN	50	0.055	0.520	0.300	0.167	0.056
MirRFTax0505EN	48	0.072	0.600	0.480	0.180	0.064
MirRFTax1005EN	50	0.055	0.520	0.300	0.167	0.056

Independently of the topic type, the highest MAP is achieved with the baseline experiment in English. As in previous campaign, topic expansion using MeSH doesn't seem to be useful and relevance retrieval leads to noticeably worse results. After a preliminary evaluation, the reranking algorithm used for combining the different results list is again the reason for the low precision values obtained in the experiments that make use of the relevance-feedback methods. Obviously other combination operators must be studied, in special those that assign a higher weight to documents that correspond to the initial query and a lower weight to documents found by the relevance feedback query.

If a comparison among the different topic types is made, it can be clearly observed that topics tagged as visual or mixed achieve noticeably better results than semantic topics. This can be explained by the fact that semantic topics have a very low number of terms (i.e., topics are quite short sentences) as compared to other topic types, and this issue negatively affects experiments based on purely textual information retrieval.

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 has to be again invested to research on this issue.

4. Conclusions and Future Work

After a preliminary analysis, it can be observed that the low number of relevant documents retrieved and the low precision (MAP) values achieved in all experiments in general may be caused by the fact that common terms such as body parts ("head", "lungs") that are not directly related to the pathology or diagnosis referenced by the topic, are more frequent in the image description that the actual terms that model or characterize the medical concept ("cancer", "aneurysm"), which produces that the relevance of the result is determined by those wrong terms instead of the others.

For future participations, we will try to isolate the terms that actually describe each medical case (those terms that refer to any pathology or diagnosis technique) and use them to determine the relevance with respect to the topic, for example using a reranking algorithm to calculate the result list.

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. Müller, H.; Kalpathy-Cramer, J.; Eggel, I.; Bedrick, S.; Radhouani, S.; Bakke, B.; Kahn, C. Jr.; Hersh, W. Overview of the CLEF 2009 medical image retrieval track, CLEF working notes 2009, Corfu, Greece, 2009.

2. Lana-Serrano, Sara; Villena-Román, Julio; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFmed 2008: Semantic vs. Statistical Strategies for Topic Expansion. Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Peters, Carol et al (Eds.). Lecture Notes in Computer Science, 2008 (printed in 2009).

3. Villena-Román, Julio; Lana-Serrano, Sara; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. Advances in Multilingual and Multimodal Information Retrieval. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, Vol. 5152, 2008. ISSN: 0302-9743/1611-3349.

4. Martínez-Fernández, José Luis; García-Serrano, Ana M.; Villena-Román, Julio; Martínez-Fernández, Paloma. Expanding Queries Through Word Sense Disambiguation. Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers. Carol Peters et al. (Eds.). Lecture Notes in Computer Science , Vol. 4730, 2007. ISSN: 0302-9743

5. U.S. National Library of Medicine. National Institutes of Health. Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/.

6. University of Neuchatel. Page of resources for CLEF. http://www.unine.ch/info/clef.

7. Porter, Martin. Snowball stemmers and resources page. http://www.snowball.tartarus.org.

8. Apache Lucene project. http://lucene.apache.org.

9. U.S. National Library of Medicine. National Institutes of Health. Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/.