MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval

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

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

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

Abstract

This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Retrieval task of ImageCLEF 2007. For this campaign, our challenge was to research on different merging strategies, i.e. methods of combination of textual and visual retrieval techniques. We have focused on the idea of performing all possible combinations of well-known textual and visual techniques in order to find which ones offer the best results in terms of MAP and analyze if the combined results may improve the individual ones. Our system consists of three different modules: the textual (text-based) retrieval module, which indexes the case descriptions to look for those descriptions which are more relevant to the text of the topic; the visual (content-based) retrieval component, which provides the list of case images that are more similar to the topic images; and, finally, the merging module, which offers different operators (AND, OR, LEFT, RIGHT) and metrics (max, min, avg, max-min) to combine and rerank the outputs of the two previous subsystems**.** These modules are built up from a set of basics components organized in four categories: (i) resources and tools for both general-domain and medical-specific vocabulary analysis, (ii) linguistic tools for text-based information retrieval, (iii) tools for image analysis and retrieval, and (iv) ad-hoc tools for result merging and reranking. We finally submitted 50 runs. The highest MAP was obtained with the baseline text-based experiment in English where only stemming plus stopword removal is performed. Neither tagging with UMLS medical concepts nor merging of textual and visual results proved to be of value to improve the precision with regards to the baseline experiment. However, the most interesting conclusion is that experiments that use the OR operator obtain higher MAP values than those with the AND operator.

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, domain-specific vocabulary, thesaurus, linguistic engineering, information retrieval, indexing.

1. Introduction

The MIRACLE team 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 [7] [8], Question Answering, WebCLEF and GeoCLEF tracks.

This paper describes our participation in the ImageCLEFmed task of ImageCLEF 2007. The goal of this task (fully described in [9]) is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text. The task organizers provide a list of topic statements (a short textual description explaining the research goal) in English, French and German, and a collection of images (from one to three) for each topic. The objective is to retrieve as many relevant images as possible from the

given visual and multilingual topics. ImageCLEFmed 2007 extends the experiments of past editions with a larger database and even more complex queries.

Although this task certainly requires the use of image retrieval techniques and our areas of expertise do not include image analysis research, we do take part to promote and encourage multidisciplinary participation in all aspects of information retrieval, no matter whether it is text or content based.

All experiments are fully automatic, thus avoiding any manual intervention. We submitted runs using only text (text-based retrieval) or only visual features (content-based retrieval) and also mixed runs using a combination of both.

2. System Description

Our system is logically built up from three different modules: the textual (text-based) retrieval module, which indexes case descriptions in order to look for the most relevant ones to the text of the topic; the visual (content-based) retrieval component, which provides the list of case images that are more similar to the topic ones; and, finally, the result combination module, which uses different operators to combine the results of the two previous subsystems**. Figure 1** gives an overview of the system architecture.

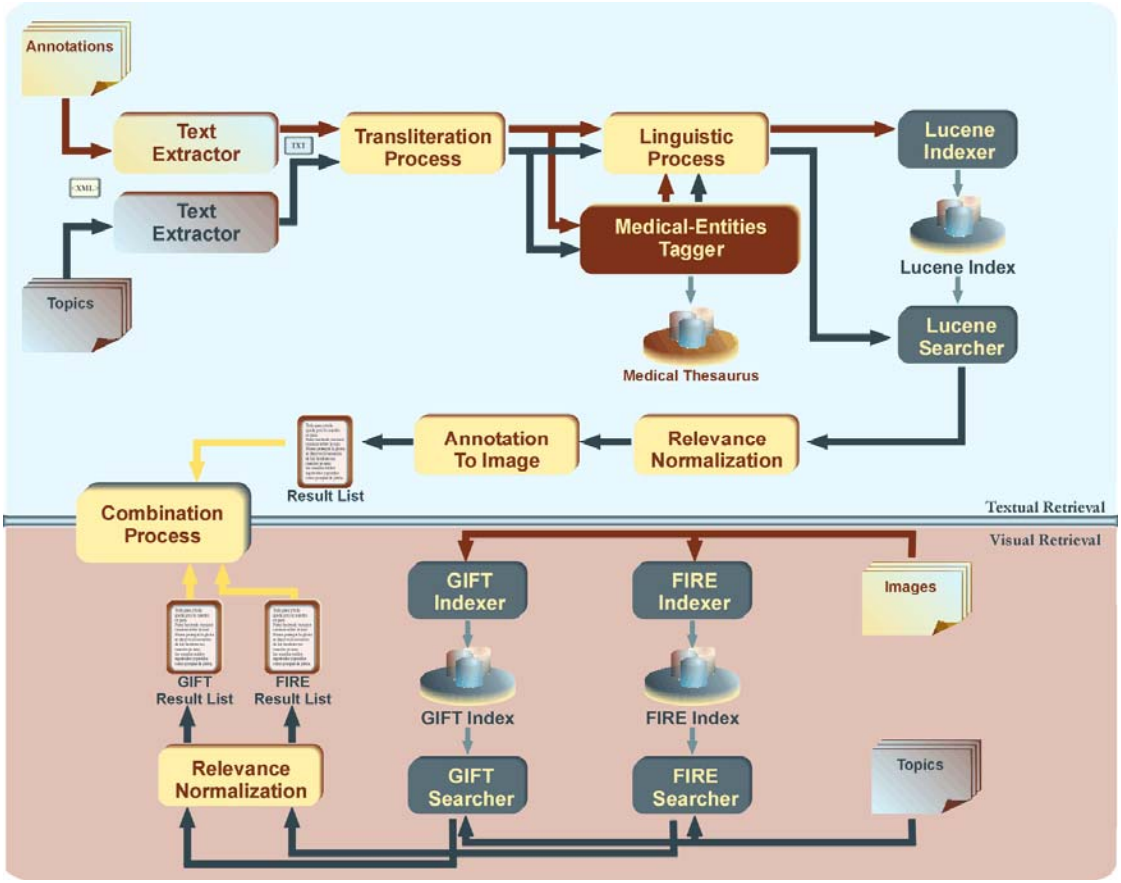

Figure 1. Overview of the system.

2.1. Textual Retrieval

The system consists of a set of different basic components organized in two categories:

- Resources and tools for medical-specific vocabulary analysis
- Linguistic tools for textual analysis and retrieval.

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 consequence of this process is that concept identifiers [5] 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) [12] and incorporating terms in English, Spanish, French and German (the four different languages involved in the ImageCLEFmed task [9]). This dictionary contains 4,327,255 entries matching 1,215,749 medical concepts. **Table 1** shows the language coverage of terms (the same as UML).

Table 1. Language distribution of terms.

Lang	#Terms
EN	3,207,890
ES	1,116,086
FR	2,556
DE	723

For example:

Tagged Topic (M7)
Pathology [***non hodgkins lymphoma***] **UML_C0024305**

Pertinent Tagged document (PathoPic/000041_en)
Primary [***Non Hodgkin's lymphoma***] **UML_C0024305** [lymphoma of the heart] UML_C1332850 41 [***NHL***] **UML_C0024305** UML_C0079745 UML_C1705385
Pertinent Tagged document (PathoPic/000689_en)
[chronic lymphatic leukemia] UML_C0023434 UML_C0023458 689 [CLL] UML_C0023434 UML_C0023458 [***NHL***] **UML_C0024305** UML_C0079745 UML_C1705385

The baseline approach to process the document collection is based on the following steps which are executed in sequence:

1. **Text Extraction:** Ad-hoc scripts are run on the files that contain information about the medical cases in order to extract the annotations and metadata enclosed between XML tags. **Table 2** shows the metadata which was considered from each collection.

Table 2. Metadata extracted from XML annotation files.

Collection	Lang	Metadata
CASImage	FR	Description, Diagnosis, Clinical Presentation, Keywords, Anatomy, Chapter, Title, Age
Endoscopy	EN	Title, Subject, Description
myPACS	EN	Title, Abstract, Keywords, Text-Caption, Discussion, Document-Type, Pathology, Anatomy, Pt-Sex, Months, Years, Days
PathoPICS	DE	Diagnose, Synonyme, Beschreibung, Zusatzbefund, Klinik, Kommentar
	EN	Diagnosis, Synonyms, Description, AddtlFindings, ClinicalFindings, Comment
Peir	EN	Title, Description, RadiographType, DiseaseProcess, ClinicalHistory
MIR	EN	Diagnosis, Brief_History, Images, Full_History, Radiopharm, Findings, Discussion, Followup, Teaching

2. **Medical-vocabulary Recognition:** All case descriptions and topics are parsed and tagged using a subset of Unified Medical Language metathesaurus [12] to identify and disambiguate medical terms.
3. **Tokenization:** This process extracts basic text components, detecting and isolating punctuation symbols. Some basic entities are also treated, such as numbers, initials, abbreviations, and years. So far, compounds, proper nouns, acronyms or other entities 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. **Lowercase words:** All document words 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 [11] and afterwards extended using several other sources [2] as well as our own knowledge and resources [8].
6. **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. Standard stemmers from Porter [10] have been used.
7. **Indexing and retrieval:** The information retrieval engine applied for all textual indexing and retrieval task was Lucene [1].

No feedback or any other kind of expansion was used.

Because the textual retrieval module is completely based on information about medical cases, the last step of module is to obtain the images that correspond to each case (block labeled as AnnotationToImage at **Figure 2**).

2.2. Visual Retrieval

For this part of the system, we resorted to two publicly and freely available Content-Based Information Retrieval systems: GIFT (GNU Image Finding Tool) [4] and FIRE (Flexible Image Retrieval Engine) [3] [6]. They are both developed under the GNU license and allow to perform query by example on images, using an image as the starting point for the search process and relying entirely on the image contents.

In the case of GIFT, the complete image database was indexed in a single collection, down-scaling each image to 32x32 pixels. For each ImageCLEFmed query, a visual query is made up of all the images contained in the query. Next, this visual query is used in GIFT to obtain the list of the most relevant images (i.e., images which are more similar to those included in the visual query), along with the corresponding relevance values. Although different search algorithms could be integrated as plug-ins in GIFT, only the provided separate normalization algorithm has been used in our experiments.

On the other hand, we directly used the results of the FIRE system kindly provided by the organizers, with no further processing.

2.3. Merging

The textual and image result lists are then merged by applying different techniques, which are characterized by an operator and a metric for computing the relevance (score) of the result. **Table 3** shows the defined operators: union (OR), intersection (AND), difference (AND NOT), and external join (LEFT JOIN, RIGHT JOIN). Each of these operators selects which images are part of the final result set.

Table 3. Combination operators.

Operators	
OR	$A \cup B$
AND	$A \cap B$
LEFT	$(A \cup B) \cup (A - B)$
RIGHT	$(A \cup B) \cup (B - A)$

Then, results are reranked by computing a new relevance measure value based on their corresponding input results by using different metrics shown in **Table 4**.

Table 4. Score computing metrics.

Metrics	
max	$score = \max(a, b)$
min	$score = \min(a, b)$
avg	$score = \text{avg}(a, b)$
mm	$score = \max(a,b) + \min(a,b) * \frac{\min(a,b)}{\max(a,b) + \min(a,b)}$

3. Experiment Set

Experiments are defined by the choice of different combinations of the previously described modules, operators and score computation metrics. A wide set of experiments was submitted: 8 text-based runs covering the 3 different topic languages, 9 content-based runs (built with the combination of results from GIFT and FIRE), and also 33 mixed runs (built with the combination of textual and visual experiments).

Table 5. Textual experiments.

Run Identifier	Language [(1)]	Method
TxtENN	EN>all	stem + stopwords
TxtENT	EN>all	stem + stopwords + tagged with UMLS thesaurus
TxtFRN	FR>all	stem + stopwords
TxtFRT	FR>all	stem + stopwords + tagged with UMLS thesaurus
TxtDEN	DE>all	stem + stopwords
TxtDET	DE>all	stem + stopwords + tagged with UMLS thesaurus
TxtXN	all>all	stem + stopwords
TxtXT	all>all	stem + stopwords + tagged with UMLS thesaurus

[(1)] [Query language] > [Annotation language]; "all" refers to the concatenation of text in all languages

Table 6. Visual experiments.

Run Identifier	Method [(1)]
VisG	GIFT
VisGFANDavg	GIFT ANDavg FIRE
VisGFANDmax	GIFT ANDmax FIRE
VisGFANDmin	GIFT ANDmin FIRE
VisGFANDmm	GIFT ANDmm FIRE
VisGFORavg	GIFT ORavg FIRE
VisGFORmax	GIFT ORmax FIRE
VisGFORmin	GIFT ORmin FIRE
VisGFORmm	GIFT ORmm FIRE

[(1)] The merging strategy is defined by [Operator] [Metric]

Table 7. Mixed textual and visual retrieval experiments.

Run Identifier	Method	Merging strategy
MixGENT[Merging]	VisG+TxtENT	ANDmax, ANDmin, ANDavg, ORmax, ORmin, ORavg, ORmm, LEFTmax, LEFTmin, LEFTmm, RIGHTmax, RIGHTmin, RIGHTmm
MixGFRT[Merging]	VisG+TxtFRT	ORmax, ORmm, LEFTmax, LEFTmm, ANDmin
MixGDET[Merging]	VisG+TxtDET	ORmax, ORmm, LEFTmax, LEFTmm, ANDmin
MixGFANDminENT[Merging]	VisGFANDmin+TxtENT	ORmax, ORmm, LEFTmax, LEFTmm, ANDmin
MixGFORmaxENT[Merging]	VisGFORmax+TxtENT	ORmax, ORmm, LEFTmax, LEFTmm, ANDmin

4. Results

Results are presented in the following tables. Each of them shows the run identifier, the number of relevant documents retrieved, the mean average precision (MAP), the R-precision and the precision at 10, 30 and 100 first results.

Table 8 shows the results of the text-based experiments. The highest MAP is obtained by the baseline experiment in English where only stemming plus stopword removal is performed. Surprisingly for us, tagging with UMLS thesaurus has proved to be of no use with regards to the simplest strategy. This issue has to be further investigated in case that there is some problem with the generation of the result sets.

Table 8. Results for textual experiments.

	RelRet	MAP	R-prec	P10	P30	P100
TxtENN	2,294	0.3518	0.389	0.58	0.4556	0.36
TxtXN	2,252	0.299	0.354	0.4067	0.3756	0.2943
TxtENT	2,002	0.274	0.2876	0.45	0.3822	0.2697
TxtXT	1,739	0.2005	0.2118	0.3267	0.2889	0.2263
TxtFRN	898	0.1107	0.1429	0.2733	0.1989	0.133
TxtFRT	970	0.1082	0.1138	0.2533	0.1911	0.1297
TxtDET	694	0.0991	0.0991	0.23	0.1222	0.0837
TxtDEN	724	0.0932	0.1096	0.18	0.1356	0.097

Experiments using French and German languages achieve a very low precision (respectively, a decrease to 31% and 28% with regards to English). This result is similar to other experiments carried out in other CLEF tracks and may be attributed to deficient stemming modules.

The evaluation for the content-based experiments is shown in **Table 9**.

Table 9. Results for visual experiments[(1)].

	RelRet	MAP	R-prec	P10	P30	P100
VisG	532	0.0186	0.0396	0.0833	0.0833	0.047
VisGFANDmm	165	0.0102	0.0255	0.0667	0.05	0.0347
VisGFANDmax	165	0.0099	0.0251	0.06	0.0511	0.0343
VisGFANDavg	165	0.0087	0.0214	0.0467	0.0556	0.0343
VisGFANDmin	165	0.0081	0.0225	0.0367	0.0478	0.0333

(1) Evaluations for some experiments with OR operator are missing

In general, MAP values are very low, which reflects the complexity and difficulty of the visual-only retrieval for this task. The best value (5% of the top ranked textual experiment) is obtained with the baseline visual experiment, which just uses GIFT. However, probably due to an oversight by the task organizers, the evaluations for the experiments with the OR operator (4 runs) are missing in the Excel files provided. Thus, no definitive conclusion can be extracted about the usage of any merging strategy, as the restrictive AND operator filters out many images (165 instead of 532 relevant images retrieved).

Finally, **Table 10** in next page shows the evaluation for the mixed runs. Although the MAP of the best ranked mixed experiment is lower than the MAP of the best textual one (77%), we cannot conclude that the combination of textual and visual results with any kind of merging strategy fails to improve the precision because. The same as before, some experiments with OR operator (11 runs) are missing from the table, thus, it is impossible to extract any valuable conclusion on this issue.

However, observe that the best ranked runs are those with the RIGHT operator, which implicitly includes an OR (see definition in **Table 4**). In addition, the use of this operator (visual RIGHT textual) shows that textual results are preferred over visual results (RIGHT prioritizes the second result list).

Another conclusion that can be drawn from these results is that the textual retrieval is the best strategy for this task. We think that this is because many queries include semantic aspects such as medical diagnoses or specific details present in the image, which a purely visual retrieval cannot tackle. This issue will be considered for future participations.

The best experiment at ImageCLEFmed 2007 reaches a MAP value of 0.3962, 112% better than ours. Despite this difference, MIRACLE participation is ranked 3rd out of over 12 groups, which is indeed considered to be a very good position.

Table 10. Results for mixed textual and visual retrieval experiments[(1)].

	RelRet	MAP	R-prec	P10	P30	P100
MixGENTRIGHTmin	2002	0.274	0.2876	0.45	0.3822	0.2697
MixGENTRIGHTmax	2045	0.2502	0.2821	0.3767	0.35	0.29
MixGENTRIGHTmm	2045	0.2486	0.2817	0.3733	0.3578	0.289
MixGFANDminENTORmm	1972	0.1427	0.1439	0.22	0.2	0.1793
MixGFANDminENTORmaxt	1972	0.1419	0.1424	0.2067	0.1911	0.177
MixGFRTORmm	697	0.0372	0.064	0.1433	0.1244	0.084
MixGFRTORmax	693	0.0322	0.0611	0.14	0.1233	0.0747
MixGENTLEFTmm	532	0.0279	0.0485	0.12	0.0944	0.0643
MixGDETLEFTmm	532	0.024	0.043	0.1	0.09	0.0577
MixGFRTLEFTmm	532	0.0236	0.0416	0.09	0.0889	0.058
MixGENTANDavg	162	0.0234	0.0341	0.17	0.1056	0.047
MixGENTANDmin	162	0.0229	0.0341	0.17	0.1056	0.047
MixGDETANDmin	247	0.0213	0.0415	0.12	0.0989	0.0447
MixGFRTANDmin	176	0.0209	0.037	0.1167	0.1044	0.0487
MixGFRTLEFTmax	532	0.0191	0.0398	0.0833	0.0856	0.0487
MixGDETLEFTmax	532	0.0189	0.0408	0.0867	0.0844	0.048
MixGENTLEFTmax	532	0.0186	0.0397	0.0833	0.0833	0.0473
MixGENTANDmax	162	0.0175	0.0332	0.1533	0.1044	0.047
MixGENTLEFTmin	532	0.0155	0.0339	0.0767	0.0822	0.0433
MixGFANDminENTANDmin	67	0.0114	0.0152	0.1233	0.0622	0.0207
MixGFANDminENTLEFTmm	165	0.0099	0.024	0.0533	0.0544	0.0363
MixGFANDminENTLEFTmax	165	0.0081	0.0225	0.0367	0.0478	0.0333

[(1)] Evaluations for some experiments with OR operator are missing

5. Conclusions and Future Work

The highest MAP is obtained with the baseline text-based experiment in English where only stemming plus stopword removal is performed. Neither tagging with UMLS medical concepts nor merging of textual and visual results have proved to be of value to improve the precision with regards to the baseline experiment. However, evaluations for some of our experiments were missing, so this issue cannot be confirmed and has to be further investigated. In addition, experiments using French and German languages get a very low precision. This result is similar to other experiment carried out in other CLEF tracks and may be attributed to deficient stemming modules. We will invest more effort in these languages in future participations.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by the Madrid's 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] Apache Lucene project. On line http://lucene.apache.org [Visited 10/08/2007].

[2] CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/~gjones/CLEF2005/Multi-8/ [Visited 10/08/2007].

[3] Deselaers, T.; Keysers; D.; Ney, H. FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In CLEF 2004, LNCS 3491, Bath, UK, pp 688-698, September 2004.

[4] GIFT: The GNU Image-Finding Tool. On line http://www.gnu.org/software/gift/ [Visited 10/08/2007].

[5] 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.

[6] FIRE: Flexible Image Retrieval System. On line http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html [Visited 10/08/2007].

[7] 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.

[8] 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.

[9] Müller, Henning; Deselaers, Thomas; Kim, Eugene; Kalpathy-Cramer, Jayashree; Deserno, Thomas; Clough, Paul; Hersh, William. Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.

[10] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 10/08/2007].

[11] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 10/08/2007].

[12] U.S. National Library of Medicine. National Institutes of Health. On line http://www.nlm.nih.gov/research/umls/ [Visited 10/08/2007].