SINAI at ImageCLEFmed 2008

M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia,
L.A. Ureña-López, A. Montejo-Ráez
University of Jaén. Computer Science Department
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,magc,maite,laurena,amontejo}@ujaen.es

Abstract

This paper describes the SINAI team participation in the ImageCLEF campaign.

In this paper we only explain the experiments accomplished in the medical task. We have experimented with query expansion and the text information of the collection. For expansion, we carry out experiments using MeSH ontology and UMLS separately. With respect to text collection, we have used three different collections, one with caption and title, other with caption, title and the text of the section where the image appears, and the third with the full article. Moreover, we have experimented with mixed search, textual and visual search, using the FIRE software for image retrieval.

The use of FIRE and MeSH expansion with the minimal collection (only caption and title) obtains the best results in the track.

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

Keywords

Visual and textual retrieval, Indexing, Ontologies, MeSH, UMLS

1 Introduction

This is the fourth participation of the SINAI research group at the ImageCLEF 2008 campaign, specifically in the medical task.

The goal of the medical task is to retrieve relevant images based on an image query. This year, a new collection have been used. It contains images from articles published in Radiology and Radiographics including the text of the captions and a link to the HTML of the full text articles. We have downloaded articles from the web and constructed a new textual collection including the text of the article section where the image appears. Besides, for the experiments, we have created two groups of expanded queries, a group expanded with MeSH ontology[1] and other group expanded with UMLS[2].

For mixed experiment, we have used the list of retrieved images by FIRE[3] [1], which was supplied by the organizers of this track.

[1] http://www.nlm.nih.gov/mesh/
[2] http://www.nlm.nih.gov/research/umls/
[3] http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html

The following section describes the new textual collection. In Section 3, we explain the expansion of the queries. In the next section, we comment the experiments carried out. Finally, conclusions and further work are presented in Section 5.

2 The textual collection

This year, the old collection [2] used in previous years has been discarded. A new collection has been introduced, a subset of the Goldminer[4] collection. The collection contains images from articles published in Radiology and Radiographics including the text of the captions and a link to the HTML of the full text article.

To create the different textual collections, first we have obtained the textual information by following the next steps:

1. Extract a list of articles URLs from the information of the collection given by the organizers. The numbers of articles are lower that the number of images that contains the collection, because an article contains several images.
2. Download all the articles in this list.
3. Filter the downloaded articles to extract different sections in the text: title, authors, abstract, introduction, etc.
4. Mark the position of every image in the filtered articles.

The we have created three different collections. In these collections each document contains information about each image from the original collection. The information is different for each collection. These collections and the section include the following sections:

- **CT**: It contains *caption* of image and *title* of the article.
- **CTS**: Contains caption, title and text of the *section* where the image appear.
- **CTA**: Contains caption, title and text of the full *article*.

3 Query Expansion

One of the purposes of these experiments is to compare the performance of query expansion using two different ontologies: MeSH and UMLS. Experiments with the MeSH ontology have been carried out in the past [3] obtaining good results. The expansion method using MeSH is the same as presented last year.

On the other hand, the UMLS metathesaurus is a repository of biomedical ontologies and associated software tools developed by the US National Library of Medicine(NLM)[5]. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in biomedical literature and health services research. These are referred to as the *source vocabularies* of the Metathesaurus in UMLS literature. One of the source vocabularies is MeSH ontology.

To expand the queries we have used MetaMap program [4] that was originally developed for use in information retrieval. MetaMap uses the UMLS metathesaurus for concept retrieval mapping a input text. For query expansion with MetaMap, we have mapped terms in the query. In order to reduce the number of terms that could expand the query, to make it equal to that of MeSH expansion, we have used MetaMap, restricting the semantic types in the mapped terms [5] as follows:

- bpoc: Body Part, Organ, or Organ Component

[4] http://goldminer.arrs.org/
[5] http://www.nlm.nih.gov/

- diap: Diagnostic Procedure
- dsyn: Disease or Syndrome
- neop: Neoplastic Process

MetaMap gives two types of mapped terms: *Meta Candidates* and *Meta Mapping*. The difference between both mapped terms is that the second are the Meta Candidate with best score. For our expansion we have used the Meta Candidate terms, because these terms provide similar terms with differences in the words. For example, the phrase *"'chest CT image"'* obtains following candidates:

- 793 **Image** (**Medical Imaging** {MSH,MTH,NCI}) [Diagnostic Procedure]
- 734 **Chest CT** (**Chest CT** {MTH, SNOMEDCT, RCD, SNM, SNMI, ICD9CM, MTHICD9, MDR}) [Diagnostic Procedure]
- 604 **Thoracic** (**Dissecting aneurysm of the thoracic aorta** {MTH, CCPSS, ICPC2, ICD10ENG, ICD9CM, MDR, NCI}) [Disease or Syndrome]
- 577 **Breast** (**Breast** {HL7V2.5, LCH, MSH, MTH, NCI, PSY, RCD, SNM, SNOMEDCT, UWDA, CCPSS, LNC, AOD, CSP, SNMI}) [Body Part, Organ, or Organ Component]
- 577 **Breast** (**Entire breast** {MTH, SNOMEDCT}) [Body Part, Organ, or Organ Component]
- 560 **Mammary** (**Mammary gland** {MTH, RCD, SNM, SNOMEDCT, UWDA, MSH, NCI, AOD, CSP, PSY, SNMI}) [Body Part, Organ, or Organ Component]

The first number is the score, the next one is the metathesaurus concept. The preferred name for a metathesaurus concept appears asided in parentheses. Between braces are the different source vocabularies where the concept appears, and between square brackets are the different semantics types of the concept.

Prior to the inclusion of Meta Candidates terms in the queries, the term words are added to a set where repeated words are deleted. All words in the set are included in the query.

4 Experiments

Our main goal is to investigate the effectiveness of different expansions and different sizes in textual collections. Moreover, we have experimented with the influence of mixing visual information with our results.

We have used three textual collections (CT, CTS, CTA) and three sets of topics: original, expanded with MeSH and expanded with UMLS. Besides, we have mixed our textual results with the visual results given by the organizers in order to obtain new results. The visual results have been obtained with the FIRE software. To mix textual and visual results, we have used the same algorithm that applied in 2007 [3]. In previous years the best results reached were obtained with a weight of 0.8 for textual results and 0.2 for visual ones. This year we have only experimented with these weights.

We have submitted only 10 runs, because of the limits imposed by the organizers. The results of these runs are shown in Table 1:

The official baseline results obtain 0,2768 of MAP (Main Average Precission).

Table 2 shows the top ten MAP obtained in all experiments for all of participants.

Experiment	MAP
sinai_CT_Base	0.2480
sinai_CT_Mesh	0.2792
sinai_CT_Umls	0.2275
sinai_CT_Mesh_Fire20	**0.2908**
sinai_CTS_Base	0.1784
sinai_CTS_Mesh	0.1582
sinai_CTS_Umls	0.1429
sinai_CTA_Base	0.1982
sinai_CTA_Mesh	0.2057
sinai_CTA_Umls	0.1781

Table 1: Performance of official SINAI runs

Experiment	MAP
SINAI-sinai_CT_Mesh_Fire20	0.2908
GPLSI-IR-n2_PRF_Type_mesh	0.2881
EXPPRFNegativaMesh	0.2881
SINAI-sinai_CT_Mesh	0.2792
LIG-LIG_COS0506_MPTT_Emi	0.2791
LIG-LIG_MPTT_Emix	0.2781
TEXT-MESS-TEXMXmshTypFIREidfCT	0.2777
TEXTMESSmeshTypeFIREidf_CT	0.2777
TEXTMESSmeshType_CT	0.2777
GPLSI-IR-n2	0.2768

Table 2: Performance of top ten official runs

5 Conclusions and Further Work

This new collection has been used in the ImageCLEFmed2008. Similarly to previous years, the use of textual information improved the results of baseline visual results. In this case, the use of FIRE and MeSH expansion with the minimal collection (only caption and title) obtains the best results in the track.

The use of UMLS expansion obtains worse results than the baseline. Although UMLS Metathesaurus includes MeSH ontology in the source vocabularies, MetaMap adds, in general, more terms in the queries. The MetaMap mapping is different than MeSH mapping, therefore the terms selected to expand are different.

Another conclusion is that it is better to have few textual information but more specific. Including all the section where the image appears is not good approach. Sometimes, a section contains several images, therefore the same information references different images.

Our further work was to obtain a more precise textual information by finding the phrase where a reference to the image exists, that is, by finding HTML tags that reference locally to the figure (for example: A HREF="#F1") or syntactic structures of type *"in figure 1 ..."*. Moreover, we expanded the queries with UMLS Metathesaurus using other algorithm distinct to that used by the MetaMap tool. We investigated new methods to expand with UMLS similar to the expansion with MeSH, that is, less terms but better for information retrieval.

6 Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by the University of Jaén.

References

[1] Deselaers, T., Weyand, T., Keysers, D., Macherey, W., and Ney, H.: FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval. Working Notes of the CLEF Workshop, Vienna, Austria, September 2005.

[2] Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., and Hersh, W.: Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. Working Notes of the 2007 CLEF Workshop. Sep, 2007. Budapest, Hungary.

[3] Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejo-Raez,A., and Ureña-López, L.A.: SINAI at ImageCLEF 2007. In Proceedings of the Cross Language Evaluation Forum (CLEF 2007), 2007.

[4] Aronson, A.R. Effective Mapping of Biomedical text to the UMLS Metathesaurus: the MetaMap Program. Proc.of the AMIA Symposium. Nov.3-7, Washington, DC. pp 17-21. 2001

[5] Chevallet, J.P., Lim, J.H., and Radhouani, S.: Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task. Working Notes of the 2005 CLEF Workshop. Sep, 2005. Vienna, Austria.