Taking Benefit of Query and Document Expansion using MeSH descriptors in Medical ImageCLEF 2009

Julien Gobeill[1], Douglas Theodoro[2], Emilie Patsche[2], Patrick Ruch[1]
BiTeM group

[1] University of Applied Sciences, Geneva, Switzerland
[2] University and Hospitals of Geneva, Switzerland
julien.gobeill@hesge.ch

Abstract

In 2008, we participated in medical ImageCLEF in order to compare different strategies of query and document expansion. Since we are a group specialized in Natural Language Processing, we discarded the visual aspect and we only dealt with the textual fields of the documents. We chose descriptors belonging to the Medical Subject Headings (MeSH) in order to expand the queries and the documents; therefore, these metadata were supposed to improve the retrieval process, and to be an interlingua between the collection and the topics for the multilingual tasks. MeSH descriptors for query and document expansion could be automatically computed via two strategies. Each document of this collection is provided with several fields describing the image such as title or caption; so we applied a local lexical MeSH categorizer to these fields in order to automatically extract a set of MeSH descriptors. Moreover, as each document is linked to a journal article via a PMID, we harvested the MeSH descriptors assigned to this article in MEDLINE in order to obtain a second set of MeSH descriptors. For the 2008 official runs, we chose to compare both strategies, but we subsequently showed in an unofficial run that combining them, by merging both sets of MeSH descriptors, led to the best performances. Therefore, combining both strategies increased the Mean Average Precision (MAP) of our best English official run from 0.176 in 2008 to 0.321 in 2009. Results for German and French runs were respectively MAP 0.231 and MAP 0.295.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval; I.2.7 - Natural Language Processing

Free keywords
Cross-Language Retrieval, Image Retrieval, Text categorization, Query Expansion, Document Expansion

1 Introduction
Images are getting more important and varied in the medical domain, as they become available in digital form. Despite the fact that images are language-independent, they are often accompanied by textual notes in various languages. These textual notes can strongly improve the retrieval quality [1]. The medical ImageCLEF challenge started in 2004 aiming at retrieving relevant medical images in a multilingual document collection, using visual features (images) or textual features (associated captions, titles and articles).

In the previous medical ImageCLEF, we applied textual strategies based on picture's metadata. In 2008, we particularly used Medical Subject Headings (MeSH) descriptors in order to perform query and document expansion [2]. With the 2008 and 2009 collections, sets of MeSH descriptors for query and document expansion could be automatically computed via two expansion strategies. Each document of this collection is provided with several fields describing the image such as title or caption; so we applied a local lexical MeSH categorizer to these fields in order to automatically extract a first set of MeSH descriptors. This is the *expansion with captions strategy*. Moreover, as each document is linked to a journal article via a PMID, we harvested the MeSH descriptors assigned to this article from MEDLINE in order to obtain a second set of MeSH descriptors. This is the *expansion with MEDLINE strategy*.

In 2008, we applied and compared both expansion strategies. We submitted a baseline run, simply computed by an Information Retrieval step, without any expansion. The Mean Average Precision (MAP) of this baseline run was 0.136. The official run computed with the *expansion with captions strategy* reached a MAP of 0.154 (+13%). By comparison, the official run computed with the *expansion with articles strategy* reached a MAP of 0.176 (+29%). Nevertheless, further experiments and analyses showed that both strategies were highly complementary, as they covered different aspects of the concepts searched in topics [3]. For instance, the diagnostic imaging (such as IRM or CT scan) was better described by the MeSH descriptors harvested from

MEDLINE, while the anatomical parts were better described by the MeSH descriptors extracted from the caption. Therefore, combining these both strategies, by merging both sets of MeSH descriptors, led to great improvements in an unofficial run that reached a MAP of 0.254 (+86%).

In medical ImageCLEF 2009, the size of the collection slightly raised, while the ad hoc task remained unchanged. We then combined both expansion strategies in order to compute our primary run, and take benefits in competition from what our subsequent experiments showed after medical ImageCLEF 2008.

2 Data and Strategies

The 2009 collection consisted of around 75'000 images (67'000 in 2008), along with their title, caption, and a PubMed identifier (PMID). The PMID linked the image to a publication contained in MEDLINE, provided with metadata such as authors and abstract, and a set of manually assigned MeSH descriptors. Thus, MeSH descriptors for a given image could be computed via two different strategies: whether being extracted from its title and caption (*expansion with captions strategy*), or being collected from the assigned MeSH descriptors of the linked publication (*expansion with articles strategy*). Both sets of MeSH descriptors are then used for document expansion.

The 2009 collection is qualitatively identical to the 2008 collection. More descriptions are given in [3] and [4].

2.1 Document expansion with captions strategy

For each image, we aggregated the title and the caption. Then, we applied a local lexical MeSH categorizer in order to automatically extract a set of MeSH descriptors (Fig. 1). The output of the MeSH categorizer was a ranked list of MeSH descriptors along with a confidence score. According to past studies, we chose to keep 5 MeSH descriptors per image [3].

1a)

<figureID>110931</figureID>
<title>Diagnosis of thyroid cancer in children: value of gray-scale and power doppler US</title>
<caption> Figure 1a. Transverse gray-scale US images. (a) Papillary thyroid carcinoma. Image in a 14-year-old girl depicts 9-mm subcapsular hypoechoic nodule (arrows). (b) Follicular thyroid adenoma. Image in a 12-year-old boy depicts 15-mm nodule (arrows) that is separated from the capsule by intervening thyroid parenchyma. C = carotid artery, E = esophagus, T = thyroid gland, Tr = trachea. </caption>

1b)

Score	MeSH descriptor
112222	*thyroid glands*
104202	*thyroid cancer*
75010	*carotid arteries*
74248	*tracheas*
54489	*esophagus*
40848	*diagnosis*
34960	*capsules*
27948	*carcinomas*
……..	……

Figure 1: the expansion with captions strategy. a) the caption and the title for a given image (ID 110931) of the collection. The MeSH categorizer is applied on these fields. b) the ranked list of MeSH descriptors outputted by the categorizer for this image. Only the 5 first descriptors were kept.

A deep analysis of the MeSH descriptors extracted via this strategy showed that they were particularly relevant to describe diseases and anatomical concepts [3].

2.2 Document expansion with articles strategy

For each image, we extracted the associated PMID. Then, using the PubMed e-utils [5], we collected the assigned MeSH descriptors for this publication in MEDLINE (Fig. 2). For this set of MeSH descriptors, all descriptors were kept for document expansion.

2a)

<figureID>110931</figureID>
<pmid>15770036</pmid>

2b)

Adenocarcinoma, Follicular ; Adenoma ; Adolescent ; Calcinosis ; Carcinoma, Papillary ; Child ; Diagnosis, Differential ; Female ; Goiter, Nodular ; Humans ; Image Enhancement ; Image Processing, Computer-Assisted ; Imaging, Three-Dimensional ; Male ; Neoplasm Staging; Neovascularization, Pathologic ; Prospective Studies ; Sensitivity and Specificity ; Thyroid Neoplasms ; Thyroid Nodule ; Ultrasonography, Doppler

Figure 2: the expansion with articles strategy. a) the PMID for a given image (ID 110931) of the collection. PubMed e-utils are used in order to collect the MeSH descriptors assigned to this PMID. b) the set of MeSH descriptors collected for this image.

A deep analysis of the MeSH descriptors collected via this strategy showed that they were particularly relevant to describe diagnostic imaging concepts [3]. Since the manual assignment of MeSH descriptors is ruled by best practices guides, the diagnostic imaging technique is more likely to be described in a unique standardized MeSH descriptor. For instance, a X-ray CT scan can be mentioned in the caption as a CT scan, or a X ray tomography, or a radiograph, while MEDLINE assignors often choose the MeSH descriptor *Tomography, X-Ray Computed* in order to describe it.

2.3 Query expansion

Our MeSH categorizer was used in order to compute MeSH descriptors for the queries and to perform query expansion. According to past studies [3], we kept only extracted MeSH descriptors belonging to the C and A MeSH trees, describing respectively diseases and anatomical concepts. Only 3 MeSH descriptors were kept for each query. For French and German queries, we used respectively a French and a German version of the MeSH in order to extract the descriptors, then we selected the corresponding MeSH descriptor in the English version. The MeSH behaves then as an intermediate language between English and other languages. French and German queries were translated into english by Google translator [6].

Moreover, we applied a set of manually generated rules in order to describe the diagnostic with the correct MeSH descriptor, according to the best practices in MEDLINE. For instance, for English, when the word radiograph was mapped, we expanded the query with the MeSH descriptor *Tomography, X-Ray Computed*.

3 Methods

Our MeSH categorizer was developed with EasyIR, a local Natural Language Processing toolkit. EasyIR combines a fuzzy mapping module with a vector-space model. Our MeSH categorizer compares a textual input with indexed MeSH descriptors and their synonyms, and then outputs a ranked list of MeSH descriptors along with a confidence score. More information on EasyIR is available in [7].

The Information Retrieval process was performed with Terrier [8]. We chose classic settings that showed their competitiveness in previous competitions. We used Porter stemming, and the default stopwords list available in Terrier. The weighting scheme was BM25.

Document expansion was simply performed by adding the selected MeSH descriptors at the end of the document. According to past studies [3], we chose to write the MeSH descriptors along with their unique identifier (Fig. 3).

<doc>
<docID>110931</docID>
<title>Diagnosis of thyroid cancer in children: value of gray-scale and power doppler US</title>
<caption> Figure 1a. Transverse gray-scale US images. (a) Papillary thyroid carcinoma. Image in a 14-year-old girl depicts 9-mm subcapsular hypoechoic nodule (arrows). (b) Follicular thyroid adenoma. Image in a 12-year-old boy depicts 15-mm nodule (arrows) that is separated from the capsule by intervening thyroid parenchyma. C = carotid artery, E = esophagus, T = thyroid gland, Tr = trachea. </caption>
<mesh>
thyroid glands D013961; thyroid cancer D013964 ; carotid arteries D002339 ; tracheas D014132 ; esophagus D004947 ; adenocarcinoma, follicular D018263 ; adenoma D000236 ; adolescent D000293 ; calcinosis D002114 ; carcinoma, papillary D002291 ; child D002648 ; diagnosis, differential D003937 ; female D005260 ; goiter, nodular D006044 ; humans D006801 ; image enhancement D007089 ; image processing, computer-assisted D007091 ; imaging, three-dimensional D021621 ; male D008297 ; neoplasm staging D009367 ; neovascularization, pathologic D009389 ; prospective studies D011446 ; sensitivity and specificity D012680 ; thyroid neoplasms D013964 ; thyroid nodule D016606 ; ultrasonography, Doppler D018608

</mesh>
</doc>

Figure 3: the final document representation for the image 110931.

4 Results and Discussion

In medical ImageCLEF 2009, we took benefits from what we showed in subsequent experiments from ImageCLEF 2008, i.e. combining our both expansion strategies led to great improvements [3].

Language	English	French	German
Best MAP in 2008	0.176	0.115	0.118
Best MAP in 2009	0.321	0.295	0.231

Table 1: Mean Average precision (MAP) for the best run in the three languages, in 2008 and in 2009.

Our best English run reached a Mean Average Precision (MAP) of 0.321 in 2009. By comparison, our best English run in ImageCLEF 2008 was computed with the *expansion with articles strategy* and achieved a MAP of 0.176. Thus, we confirmed in competition conditions that the merge of both expansion strategies significantly improves the performance of the document expansion. Nevertheless, we cannot assume that this rise is only due to the merge of both expansion strategies, because runs computed exactly in same conditions reached a MAP of 0.254 for the 2008 ImageCLEF. It seems that the task was slightly easier in 2009 than in 2008.

The MAP of the French run (0.295) is not so distant from the English run. We assume that it is due to the language-independent property of our document expansion strategy, because the MeSH plays the role of an interlingua between the different languages. Even if our MeSH categorizer was not evaluated for each language, we assume that it is because the MeSH categorizer has weaker performances for German than the German run was more outdistanced (MAP of 0.231). Moreover, additional French and German runs, computed without Google translator, reached respectively a MAP of 0.275 and a MAP of 0.22.

Finally, we applied a new expansion strategy that was to expand the MeSH descriptors with the set of all their synonyms provided by the MeSH. For instance, when we expanded the query 10 by the MeSH descriptor *Pulmonary Edema* (D011654), we extended the query with all its synonyms too, such as *Wet Lung*. This *synonyms strategy* run reached for English a MAP of 0.205 (- 37%). It seems that expanding with all synonyms is too brutal.

5 Conclusion

In 2009 medical ImageCLEF, we evaluated a new document expansion strategy, by combining the both strategies applied in 2008: to expand with MeSH descriptors automatically extracted from the image's metadata, and to expand with the MeSH descriptors collected from the image's associated MEDLINE publication. We then confirmed in competition conditions that this global expansion strategy led to great improvements compared to a baseline run computed without document expansion. Moreover, this strategy shows interesting language independent properties. However, the best run for all teams reached a MAP of 0.43 in 2009, while it reached a MAP of 0.29 in 2008. Hence, our improvements between 2008 and 2009 must be analyzed with caution, as it seems that the 2009 task was significantly easier than the 2008 one.

6 References

1. Müller H., Michoux N., Bandon D. and Geissbuhler A., "A review of content-based image retrieval systems in medicine - clinical benefits and future directions" , International Journal of Medical Informatics, pp. 73:1-23, 2004
2. Gobeill J., Ruch P. and Zhou X., "Text-only Cross-language image search at medical ImageCLEF 2008", CLEF working notes 2008, Aarhus, Denmark, 2008.
3. Gobeill J., Ruch P. and Zhou X., "Query and Document Expansion with Medical Subject Headings Terms at Medical ImageCLEF 2008", CLEF 2008 Proceedings, Lecture Notes in Computer Sciences, in process.
4. Müller H., Kalpathy-Cramer J., Eggel I., Bedrick S., Radhouani S., Bakke B., Kahn C. and Hersh W., "Overview of the CLEF 2009 medical image retrieval track", CLEF working notes 2009, Corfu, Greece, 2009.

5. http://www.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html.
6. http://translate.google.com.
7. Gobeill J., Müller H. and Ruch P., “Query and Document Translation by Automatic Text Categorization: A Simple Approach to Establish a String Textual Baseline for ImageCLEFmed 2006”, CLEF working notes 2006, Alicante, Spain, 2006.
8. Ounis I., Lioma C., Macdonald C. and Plachouras V.. “Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web”, Novatica/UPGRADE Special Issue on Next Generation Web Search, vol 8, pp 49-56, 2007.