Multi-Relation Modeling on Multi Concept Extraction

LIG participation at ImageClefMed

Loïc Maisonnasse, Eric Gaussier, Jean-Pierre Chevallet
Laboratory LIG
loic.maisonnasse@imag.fr, eric.gaussier@imag.fr, jean-pierre.chevallet@imag.fr

Abstract

This paper presents the LIG contribution to the CLEF 2008 medical retrieval task (i.e. ImageCLEFmed). The main idea behind our contribution is to incorporate knowledge in the language modeling approach to information retrieval (IR). On ImageCLEFmed our model makes use of the textual part of the corpus and of the medical knowledge found in the Unified Medical Language System (UMLS) knowledge sources. Last year, we used UMLS to create a conceptual representation for each sentence in the corpus, and proposed a language modeling approach on these representations. The use of a conceptual representation allows the system to work at a more abstract semantic level, which solves some of the information retrieval problems, as the one of terminological variation. We also used different concept extraction methods, and tested how to combine these extraction methods on queries.

This year, we have extended our previous method in two ways: first, we have used, in addition to relations derived from UMLS, co-occurrence relations; second, we have combined concept extraction methods not only on queries, but also on documents. In this paper, we first detail some IR approaches that use advanced index terms. We then develop the graph model used in our submission to ImageCLEFmed 2008, and the different ways use to combine graphs derived from different concept extraction methods. After this, we present our results on this year collection, showing that combined concept extraction on document improves the MAP results and that relations impact more first results precision. Finally, we conclude this work and present some possible extensions.

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;

General Terms

Algorithms, Theory

Keywords

Information retrieval, language model

1 Introduction

Best performing methods from ImageCLEFmed 2007 used advanced indexing schemes, such as conceptual or graph index (see [4]) for representing queries and documents. Such indexing schemes

allow one to better capture the content of queries and documents. They also allow matching documents and queries at an abstract semantic level. However, such representations are sometimes hard to detect from texts and may contain errors that can lower IR results. [4] proposed a graph language modeling approach that consider terms or concepts with labeled relations between them. This model takes into account semantic relations provided by an external resource, but not relations that express lexical links between terms. We extend here the model of [**?**] by integrating co-occurrence relations between terms or concepts.

If this model allows one to take in account advanced representations in an efficient IR model, it does not completely solve the problems associated with the difficulty of detecting such representations in text. We address here some of these problems by combining different concept extraction methods, on both queries and documents.

This paper first presents a short overview on the use of advanced representations in IR. A second section details the graph model used for our contribution and the methods used to combine these representations on both queries and documents. We then describe the graph extraction process used for documents and queries, and finally we present the different results obtained on the CLEF 2008 medical retrieval task.

2 State of the Art

Using semantic resources for indexing has shown promising results on domain-specific colelctions. For example, in previous ImageCLEFmed editions, conceptual indexing based on UMLS provided some of the best systems on text, significantly outperforming standard, keyword indexing (cf. [4, 2]). Similar results have also been obtained on TREC genomics, where[11] uses the *Mesh* and *Entrez* databases to select terms related to concepts from medical publications.

Several works went beyond the use of mere concepts by exploiting relations between them. Some are based on the standard space vector model, as [10] who evaluates the usefulness of UMLS concepts and semantic relations in medical IR, while others have tried to use more advanced models, as the language model of [8], to integrate dependencies between index terms in IR. Along this last line, [1] and [4] have proposed extensions of te language modeling approach that can deal with dependencies, syntactic ones in the case of [1], either syntactic or semantic in the case of [4].

The model of [1] relies on a variable L, defined as a "linkage" over query terms, which is generated from a document according to $P(L|M_d)$, where Md represents a document model. The query is then generated according to $P(Q|L, M_d)$. In principle, the probability of the query, $P(Q|M_d)$, is to be calculated over all linkages Ls, but, for efficiency reasons, the authors make the standard assumption that these linkages are dominated by a single one, the most probable one: $L = argmax LP(L|Q)$. The probability $P(Q|Md)$ is then formulated as:

$$P\left(Q|M_d\right) = P\left(L|M_d\right) P\left(Q|L, M_d\right) \tag{1}$$

In the case of a dependency parser, as the one used in [1], each term has exactly one governor in each linkage L. Then the above quantity can be decomposed, leading to a new one with three terms. This decomposition restricts the use of this model to dependency structure. Furthermore, [6] shows that this decomposition is not completely satisfactory from a theoretical point of view.

The second approach [4] proposes a graph modeling approach where query and documents are represented as graphs $G = (C, E)$, where C represents the node set of the graph and E the relation set, that they assumed labeled. The relation E is define by an application that indicates the labels associated to such relation. The probability that the graph of query Gq is generated by the model of document M_d is then decomposed as:

$$P\left(G_q|M_d\right) = P\left(C|M_d\right) P\left(E|C, M_d\right) \tag{2}$$

Where $P(C|M_d)$ corresponds to the nodes contribution and $P(E|c, M_d)$ the edges contribution. This second approach is well founded theoretically and can handle different types of graphs.

3 Graph Model

We improve the graph model proposed in [4] in which each relation is labelled with one or more labels. The next sections shows the different improvements of this model.

3.1 Node Contribution

Assuming that, conditioned on M_d, query concepts are independent of one another (a standard assumption in the language model) the node contribution can be decomposed in two different ways:

$$P\left(C|M_d\right) = \begin{cases} \prod_{c_i \in C} P_U\left(c_i|M_d\right) \\ \prod_{c_i \in C} \lambda_t P_U\left(c_i|M_d\right) + (1-\lambda_t) P_{tr}\left(c_i|M_d\right) \end{cases} \quad (3)$$

where $P\left(c_i|M_d\right)$ is the probability of a concept from the query and $P_{tr}\left(c_i|M_d\right)$ is a translation model.

The first method correspond to the standard language model, and is based on the computation of $P_U(c_i|M_d)$. The second one correspond to the usual way to incorporate lexical associations in the language modeling. This method is based on the combination of a standard language model with a translation model, and allows to take in account lexical relations. In both cases, the quantity $P(c_i|M_d)$ of equations 3 is computed through a simple Jelinek-Mercer smoothing:

$$P_U(c_i|M_d) = (1-\lambda_u)\frac{N_d(c_i)}{N_d(*)} + \lambda_u \frac{N_D(c_i)}{N_D(*)} \quad (4)$$

where $N_d(c_i)$ (respectively $N_D(c_i)$) is the number of times that c_i appears in the document d (respectively in the collection), and $N_d(*)$ (respectively $N_D(*)$) the number of concepts in document d (the collection).

The translation model is computed as:

$$P_{tr}\left(c_i|M_d\right) = \sum_{c_t \in Rl} P\left(c_i|c_t\right) P_U\left(c_t|M_d\right) \quad (5)$$

where Rl is the set of concepts lexically related to c_i and $P\left(c_i|c_t\right)$ the probability for a concept c_t to be translated by the query concept c_i. The contribution $P_U\left(c_i|M_d\right)$ still corresponds to a standard unigram language model but applied to the translated concept, with a smoothing parameter different from the one for P_U. We will refer to it as λ'_u.

3.2 Relation Contribution

We assume that E is an application from $C \times C$ in $P(\mathcal{L})$[1] that associates to each relation a set of labels. Thus the edge contribution can be decomposed as:

$$P\left(E|C, M_d\right) = \prod_{i,j \in C, i \leq j} P\left(E(c_i, c_j) = \mathcal{L}|c_i, c_j, M_d\right) \quad (6)$$

where $E(c_i, c_j) = \mathcal{L}$ indicates that a relation exists between c_i and c_j and that this relation is associated to the label set $\mathcal{L}$.

We furthermore decomposed this probability as:

$$P(E(c_i, c_j) = \mathcal{L}_{ij}|c_i, c_j, M_d) = \prod_{label \in \mathcal{L}_{ij}} P(e(c_i, c_j) = label|c_i, c_j, M_d)$$

where $e(q_i, q_j) = label$ indicates that there is a relation between q_i and q_j, the label set of which contains *label*.

[1] $\mathcal{L}$ is the set of all possible labels for a relationship and $\mathcal{P}(\mathcal{L})$ is the set of sets of $\mathcal{L}$.

An edge probability is thus equal to the product of the corresponding single-label relations. Following standard practice in language modeling, one can furthermore "smooth" this estimate by adding a contribution from the collection. This results in:

$$P(e(c_i, c_j) =, label)|c_i, c_j, M_d) = (1 - \lambda_e)\frac{D(c_i, c_j, label)}{D(c_i, c_j)} + \lambda_e \frac{C(c_i, c_j, label)}{C(c_i, c_j)} \quad (7)$$

where $D(c_i, c_j, label)$ ($C(c_i, c_j, label)$) is the number of times c_i and c_j are linked with a relation labeled *label* in the document (collection). $D(c_i, c_j)$ ($C(c_i, c_j)$) is the number of times c_i and c_j are observed together in the document.

3.3 Model combinaison

We present here the methods used to combine different graphs (i.e. different dependency structures obtained from different analyses of the queries and/or documents) in the model presented above. First, we group the different analysis of a query. To do so, we assume that a query is represented by a set of graphs $Q = G_q$; and that the probability of a set of graphs assuming a document graph model is computed by the product of the probability of each query graph:

$$P\left(Q = \{G_q\}\,|M_g\right) = \prod_{G_q} P\left(G_q|M_d\right) \quad (8)$$

This model considers that a relevant document model must generate all the posible analyses of a query Q. The best probabilities will be obtained for a document model which can generate all analyses of the query with high probability.

Second, we group the different analysis of a document. To do so, we assume that a query can be generated by different models of the same document M_d^* (i.e. a set of models). As a result of this generation process, we keep the higher probability among the different models of the document:

$$P\left(C|M_d^*\right) \quad = \quad argmax_{M_d \in M_d^*} \left(\prod_{c_i \in C} P\left(c_i|M_d\right) \right) \quad (9)$$

With this method, documents are ranked, for a given query, according to their best model.

4 Graph Extractions

UMLS is a good candidate as a knowledge source for medical text indexing. It is more than a terminology because it describes terms with associated concepts. This knowledge is large (more than 1 million concepts, 5.5 million of terms in 17 languages). UMLS is not an ontology, as there is no formal description of concepts, but its large set of terms and their variants specific to the medical domain, enables full scale conceptual indexing. In UMLS, all concepts are assigned to at least one semantic type from the Semantic Network. This provides consistent categorization of all concepts in the meta-thesaurus at the relatively general level represented in the Semantic Network. The Semantic Network also contains relations between concepts, which allows one to derive relations between concepts in documents (and queries).

From this information, graphs are produced in two steps: concept detection and then relation detection.

4.1 Concepts Detection

The detection of concepts in a document from a thesaurus is a relatively well established process. It consists of four major steps:

1. Morpho-syntactic Analysis (*POS tagging*) of document with a lemmatization of inflected word forms;

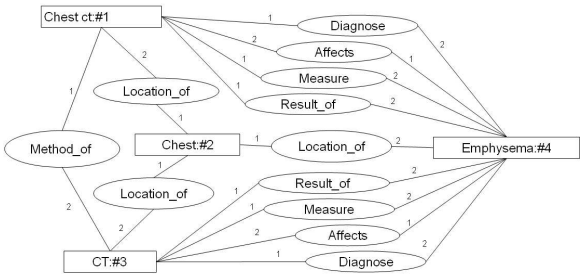

Figure 1: FCG produced for *Show me chest CT images with emphysema*

2. Filtering empty words on the basis of their grammatical class;
3. Detection in the document of words or phrases appearing in the meta-thesaurus;
4. Possible filtering of concepts identified.

For the first step, various tools can be used depending on the language. We used MiniPar(cf. [3]) and TreeTagger[2].

Once the documents are analyzed, the second and third steps are implemented directly, first by filtering grammatical words (prepositions, determinants, pronouns, conjunctions), and then by a look-up of word sequences in UMLS. This last step will find all alternatives, present in UMLS, of a concept. One can certainly improve this simple lookup by identifying potential terminological variants (see for example [**?**]). We have not used such a refinement here and merely rely on a simlpe look-up. It should be noted that we have not used all of UMLS for the third step: the thesauri NCI and PDQ were not taken into account as they are related to areas different from the one covered by the collection[3]. Such a restriction is also used in [5]. The fourth step of the indexing process is to eliminate a number of errors generated by the above steps. However, the work presented in [9] shows that it is preferable to retain a greater number of concepts for information retrieval. We thus did not use any filtering here.

We finally obtain two variations of concept detection:

- (MP) uses our term mapping tools with MiniPar.
- (TT) uses our term mapping tools with TreeTagger.

4.2 Relations Detection

After concept detection, we add conceptual relations between concepts. The relations used are those defined in the Semantic Network. We made the hypothesis that a relation exists in a document if two concepts are detected in the same sentence and if a relation between these concepts is defined in the Semantic Network. For finding relations, we first tag concept with their semantic type and then add semantic relations that link concepts with corresponding tags. A sample result of the relation extraction process for a sentence can be viewed on figure 4.2. We do not make any further disambiguisation on relations. Finally, for each concept extraction method, we obtain one graph for each document and for each query.

4.3 Coocurence Extractions

We want here to extract lexical links from the collection. We made the standard assumption that similar concepts occur in the same context (i.e. they co-occur with the same concepts). Based

[2]www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3]This is justified here by the fact that these thesauri focus on specific issues of cancer while the collection is considered more general and covers all diseases.

on this assumption, a standard method consists in building a context vector for each concept of the collection, and to compute a similarity between concepts using context vectors. In this work, we assume that a concept is in the context of another if these two concepts appear in the same sentence. Thus we compute a context vector for each concept of the collection based on a mutual information score. The weigth of the dimension c_j from the context vector of c_i is computed as:

$$MI\,(c_i, c_j) = log(\frac{P(c_i, c_j)}{P(c_i) * P(c_j)}) \quad (10)$$

$$= \frac{Nph(c_i, c_j) * Nph}{Nph(c_i) * Nph(c_j)} \quad (11)$$

where $Nph(c_i, c_j)$ is the number of times that the two concepts c_i and c_j appear in the same sentence, $Nph(c_i)$ is the number of times that c_i appear in a sentence, and Nph is the number of sentences in the collection. For efficiency and based on experimental results, we only keep the 200 highest dimensions in the context vector. We then calculate the similarity between concepts throug the cosine of their context vectors. We consider a concept c_i related to another concept c_j if c_i is in the 200 nearest neighbors (as defined by the cosine similarity) of c_j. This method provides a first set of concepts Rl.

We used the concepts in Rl to compute the translation probabilities, by dividing the cosine of the concept by the sum of the cosine of all the retained concepts:

$$P\,(c_i|c_t) = \frac{cos(V_{ctxt}(c_i), V_{ctxt}(c_t))}{\sum_{c_j \in Rl} cos(V_{ctxt}(c_i), V_{ctxt}(c_j))} \quad (12)$$

where $V_{ctxt}(c)$ is a context vector built with mutual information, Rl is the list of the N selected concepts and cos is the cosine between vectors.

5 Evaluation

We show here the results obtain for this methods on the corpus CLEFmed 2007 [7] and on the test one the CLEFmed 2008 corpus.

5.1 Model Variations

This year the track ImageCLEFmed is based on a new collection. On this collection, we submit 10 runs these runs explore differents variation of our relational model and the differents analysis merging methods. Last year results show that merging queries improves the results. As consequence, this year we do not test query graphs combination and we allways use the two graphs detected on a query.

We test 4 model variations :

- (UNI) that only use node contribution (as define in **??**).
- (RET) that use the node contribution and a relation contribution.
- (COS) that use the node contribution with translation.
- (RC) that use the node contribution with translation and a relation contribution.

For each model, we test it on the analysed collection obtain with MiniPar (MP) and on the collection analysed by MiniPar and TreeTagger (MPTT) using the combination methods proposed in this paper.

We also submit two other run, one that use a unigramme model with an extended image description that integrates the text that corresponds to the paragraph where the image is referred. A second one use the COS model but use coocurence computed on the previous ImagesCLEFmed collections.

Table 1: best results for mean average precision (MAP) and precision at five documents (P@5)

	2006-2007[4]		2008	
model	MP	MP TT	MP	MP TT
MAP				
UNI	0.2978	0.3057	0.2542	**0.2781**
RET	0.3037	0.3127	0.2556	0.274
COS	0.3058	**0.3152**	0.2443	0.2728
RC	NA	NA	0.2368	0.267
P@5				
UNI	0.487	0.490	0.447	0.433
RET	0.531	**0.516**	0.467	0.460
COS	0.510	0.510	0.447	0.467
RC	NA	NA	0.467	**0.480**

5.2 Results

From each method we use the bests parameters obtained on ImageCLEFmed corpus for MAP and we use these parameters on the new collection. We compare the variation between the results on the two corpus for the MAP and the P5D.

The best results obtained on the new medical collection are those of the unigramme model with a collection analysis by MiniPar and TreeTagger. On 2008 collection, integrating relations only improves the results when lexical relations are used on the collection analized by MiniPar. In the others cases no improvement are obtained with relations and thus combination of the two types of relation did not improved the results. On the P@5 the use of relations improves the results even more if the two analysis are used.

The results of our two other runs show that using coocurence computed on the past collection gives better results than the coocurences learned on the 2008 collection. This run gives us our best MAP result (0.2791). The other run that uses part of the article, provides surprising low results (0.1908). This can be due to the fact that the text added is considered as equivalent to the image caption, but it can be less precise or less image related. Thus we think taht this approach could provide good results if we adapt our model to take in account this new text.

6 Conclusion

We proposed here a framework for using semantic resources in the medical domain. We describe a method for using relations in language modeling, and for merging different document or query versions in this framework. Results show that relation are useful to maintain good results on the first retrieved documents, when mixing different detection trends to improve the recall. This paper shows the robustness of our method on a new corpus, where they provide good results.

References

[1] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Research and Development in Information Retrieval*, 2004.

[2] Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Xiong Wei, Daniel Raccoceanu, Diem Le Thi Hoang, Roxana Teodorescu, and Nicolas Vuillenemot. Ipal knowledge-based medical image retrieval in imageclefmed 2006. In *Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain*, 2006.

[3] D. Lin. Dependency-based evaluation of MiniPar. In *Workshop on the Evaluation of Parsing Systems, Granada, Spain, May.* ACM, 1998.

[4] Eric Gaussier Loc Maisonnasse and Jean Pierre Chevallet. Multiplying concept sources for graph modeling. In *CLEF 2007, LNCS 5152 proceedings*, 2008.

[5] Y. Huang HJ. Lowe and WR. Hersh. A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. In *Conference of the American Medical Informatics Association*, 2003.

[6] Loic Maisonnasse, Eric Gaussier, and Jean-Pierre Chevallet. Revisiting the dependence language model for information retrieval. In *Research and Development in Information Retrieval*, 2007.

[7] Henning Müller, Thomas Deselaers, Eugene Kim, Jayashree Kalpathy-Cramer, Thomas M. Deserno, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[8] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, 1998.

[9] Said Radhouani, Loic Maisonnasse, Joo-Hwee Lim, Thi-Hoang-Diem Le, and Jean-Pierre Chevallet. Une indexation conceptuelle pour un filtrage par dimensions, experimentation sur la base medicale imageclefmed avec le meta thesaurus umls. In *COnference en Recherche Information et Applications CORIA'2006*, pages 257–271, mars 2006.

[10] VolkSemantic M. Vintar S, Buitelaar P. Relations in concept-based cross-language medical information retrieval. *In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2003.

[11] Neil Smalheiser Vetle Torvik Jie Hong Wei Zhou, Clement Yu. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Research and Development in Information Retrieval*, 2007.