Multiplying Concept Sources for Graph Modeling

Loïc Maisonnasse, Eric Gaussier, Jean-Pierre Chevallet
LIG-UJF, IPAL-I2R
loic.maisonnasse, eric.gaussier@imag.fr, viscjp@i2r.a-star.edu.sg

Abstract

This paper presents the LIG contribution to the CLEF 2007 medical retrieval task (i.e. ImageCLEFmed). The main idea in this paper is to incorporate medical knowledge in the language modeling approach to information retrieval (IR). Our model makes use of the textual part of ImageCLEFmed corpus and of the medical knowledge as found in the Unified Medical Language System (UMLS) knowledge sources. The use of UMLS allows us to create a conceptual representation of each sentence in the corpus. We use these sentence representations to create a graph model for each document. As in the standard language modeling approach, we evaluate the probability that a document graph model generates the query graph. Graphs are created from medical texts and queries, and are built for different languages, with different methods. The use of a conceptual representation allows the system to work at a higher semantic level, which solves some of the information retrieval problems, as term variation. After developing the graph model in the first part of the paper, we present our tests, which involve mixing different concepts sources (i.e. languages and methods) for the matching of the query and text graphs. Results show that using language model on concepts provides good results in IR. Multiplying the concept sources further improves the results. Lastly, using relations between concepts (provided by the graphs under consideration) improves results when only few conceptual sources are used to analyze the query.

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

Previous ImageCLEFmed raised the interest of the use semantic resources for information retrieval, essentially in specialised domains such as the medical domain. Indeed, some of the best performing methods from ImageCLEFmed used resources for concept extraction. As concepts can be defined as human understandable unique abstract notions independent from any direct material support, from any language or information representation, indexing at the conceptual level solves term variation problems and conceptual indexing is naturally multilingual. Most of the previously proposed works on concepts integrate concepts in a vector space model. We propose to improve such conceptual indexing in two ways. First we used an advanced representation of the document by using relations between concepts, thus a document is represented as a graph. Secondly we integrate such representation in a more advanced model than the vector space model. We propose to extend the graph language modeling approach developed in [5] by considering that relations between terms or concepts are labeled (both syntactic and semantic relations are generally labeled; the model we present here thus addresses a common situation). This paper first presents a short overview of the use of concepts in medical document indexing and language modeling for complex structures. Then a graph modeling approach is proposed. The different graph extraction processes used for documents and queries are then described. Finally, the different results obtained on the CLEF 2007 medical retrieval task are presented.

2 State of the Art

This section explores previous work on the use of conceptual indexing in the medical domain as well as previous work on the use of structure in language modeling.

2.1 Graphical Representations in the Medical Domain

The usefulness of concepts has been shown in the previous ImageCLEFmed tasks, where some of the best performing systems on text [2] used conceptual indexing methods based on vector space models. On TREC genomics, [12] uses the *Mesh* and *Entrez* databases to select terms from medical publications. They use terms related to concepts, by identifying this terms in document and in queries they improves the results of bag of words. They also made different experiments by adding domain specific knowledge to the query. They add related terms, thesaurus relation, or computed variations. Results show that adding the terms variants of concepts gives the best improvement of results. If authors have directly used concepts instead of terms associated to concepts we supposed that variation of the concept would have been incorporate in the concepts, so terms variation improvement directly is related to concepts. The authors declare that *retrieval on the concept level can achieve substantial improvement over purely term-based retrieval model.*

Other researchers have tried to go beyond the use of concepts by exploiting relations between concepts. [11] evaluates the usefulness of UMLS concepts and semantic relations in medical IR. They first extract concepts and relations from documents and queries. To do so, they hypothesize that if two concepts have a semantic relation in a thesaurus and appear in the same sentence, then the semantic relation holds between the associated words. To select relations, they rely on two further assumptions: (1) interesting relations occur between interesting concepts; (2) relations are expressed by typical lexical markers such as verbs. They implement assumption (1) through a selection based on the IDF of concepts linked by a relation, whereas assumption (2) leads them to use a co-occurrence matrix between lexical verbs and relations. The method is evaluated on a medical collection with 25 queries with relevance assessments provided by medical experts. In their experimentation, concepts and relations are considered in different vector spaces. The experiments show that using both concepts and relations lower the results obtained with concepts alone.

2.2 Structure Language Modeling

The language modeling approach to IR has first been proposed in [7]. The basic idea is to view each document as a language sample and querying as a generative process. Even though smoothed unigram models have yielded good performance in IR, several works have investigated, within the language modeling framework, the use of more advanced representations. Works like [10] and [9] proposed to combine unigram models with bigram models. Others works, e.g. [4] or [3], incorporated syntactic dependencies in the language model. [3], for example, introduces a dependence language model for IR which integrates syntactic dependencies in the computation of document relevance scores. This model 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 M_d represents a document model. The query is then generated given L and M_d, 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: $L = \text{argmax}_L P\left(L|Q\right)$. $P(Q|M_d)$ 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 [3], each term has exactly one governor in each linkage L, so that the above quantity can be further decomposed, leading to:

$$\log P\left(Q|M_d\right) = \quad \log P\left(L|M_d\right) + \textstyle\sum_{i=1..n} \log P\left(q_i|M_d\right) + \sum_{(i,j)\in L} MI\left(q_i, q_j|L, M_d\right) \tag{2}$$

where MI denotes the mutual information, and:

$$P(L|M_d) \propto \prod_{(i,j)\in L} \hat{P}(R|q_i, q_j) \tag{3}$$

$\hat{P}(R|q_i, q_j)$ in the above equation represents the empirical estimate of the probability that concepts q_i and q_j are related through a parse in document d. As the reader may have noticed, there is a certain ambiguity in the way the linkage L is used in this model. Consequently, this model is not completly satisfying (see [5] for a short discussion of this problem), and we rely on a different model to account for graphical structures in the language modeling approach to IR. We now describe this model, which we will refer to as the *graph model*.

3 Graph Model

We propose a graph modeling approach for IR (which generalizes the one proposed in [5]) in which each relation is labelled with one or more lables. We assume that a semantic analysis of a query q can be represented as a graph $G_q = (C, E)$, where C is the set of terms (or concepts) in q, and E is a relation from $C \times C$ in the set of the label sets EN ($E(c_i, c_j) = \{\textit{labels}\}$ if c_i and c_j are related through a relation labelled with the labels in $\{\textit{labels}\}$, and $\emptyset$ otherwise). The probability that the graph of query q is generated by the model of document d can be decomposed as:

$$P\left(G_q|M_d\right) = P\left(C|M_d\right) P\left(E|C, M_d\right) \tag{4}$$

Assuming that, conditioned on M_d, query concepts are independent of one another (a standard assumption in the language model), and that, conditioned on M_d and C, edges are independent of one another (again a standard assumption), we can write:

$$\begin{aligned} P\left(C|M_d\right) \quad &= \textstyle\prod_{c_i\in C} P\left(c_i|M_d\right) &(5)\\ P\left(E|C, M_d\right) \; &= \textstyle\prod_{(i,j)} P(E(q_i, q_j)|C, M_d) &(6)\end{aligned}$$

Equation 5 corresponds to the standard language model (potentially applied to concepts), and equation 6 carries the contribution of edges. The quantity $P(c_i|M_d)$ of equation 5 is computed through a simple Jelinek-Mercer smoothing:

$$P(c_i|M_d) = (1 - \lambda_u)\frac{D(c_i)}{D(*)} + \lambda_u \frac{C(c_i)}{C(*)} \tag{7}$$

where $D(c_i)$ ($C(c_i)$) is the number of times c_i appears in a document (collection) and $D(*)$ ($C(*)$) in the number of concepts in the document (collection).

The quantities $P(E(q_i, q_j)|C, M_d)$ of equation 6 can be decomposed as:

$$P(E(q_i, q_j)|C, M_d) = \prod_{label \in E(q_i, q_j)} P(R(q_i, q_j, label)|q_i, q_j, M_d) \tag{8}$$

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

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:

$$P(R(c_i, c_j, label)|C, M_d) = (1 - \lambda_e)\frac{D(c_i, c_j, label)}{D(c_i, c_j)} + \lambda_e \frac{C(c_i, c_j, label)}{C(c_i, c_j)} \tag{9}$$

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.

The above model can be applied to any graphical representation of queries and documents, and relies on only two terms, which are easy to estimate. We now show how this model behaves experimentally.

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). Unfortunately, UMLS is constitute of different sources (thesaurus, terms lists), and is neither complete, nor consistent. UMLS is a "meta thesaurus", i.e. a merger of existing thesaurus. It is not an ontology, because there is no formal description of concepts, but its large set of terms and variation restricted to medical domain only, enable full scale conceptual indexing system. 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. This enables to detect general semantic relation between concepts that are define in this network.

Graphs are produced in two steps: concept detection and then relation detection. For concept detection, we use three different methods. The first one use MetaMap [1] but as MetaMap is only for English we use it only on the English part of the collection. No equivalent tools are available for French and for German, so we use our own mapping tools. This tool selects in UMLS all concepts that have a textual instance in the document. We show in [8] that such a strategy is better than extracting only precise concepts and can challenge or improve text based IR. This concept detection uses a syntactic analysis of sentences and a term mapping. The syntactic analysis is provided by MiniPar (for English) or by treeTagger (for all languages). To improve term mapping, we carried out some filtering on word and/or on UMLS. First we consider that stop word can not be instance of a concept even if such instance exists in UMLS. We eliminate some specific thesaurus/ontologies from UMLS (those that are very precise or that are not relevant for our task). At last, we remove from UMLS some hierarchies that we consider irrelevant for our task (ex. hierarchy corresponding to "Geographic area"). Such filtering reduces terms ambiguity and consequently improve the mapping. Results of concept extraction for a sentence can be view on figure 4.

We therefore have three concept extraction methods:

- (1) MetaMap
- (2) Mapping tools with MiniPar

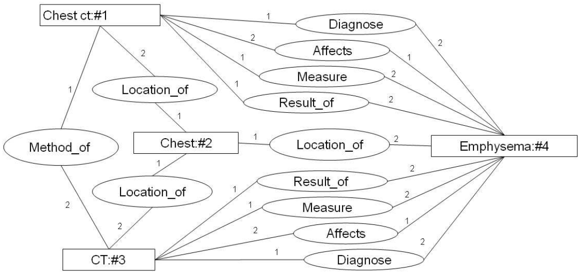

Figure 1: FCG produced for *Show me chest CT images with emphysema*

- (3) Mapping tools with TreeTagger

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 semantic graph 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 we add semantic relations that link concepts with corresponding tag. Result of relation extraction for a sentence can be view on figure 4. We don't make any further disambiguisation on relations.

Since we compute the frequency $D(c_i, c_j, name)$ $(C(c_i, c_j, name))$ as the number of times that c_i and c_j are linked in sentence on document (collection) and the probabilities $D(c_i, c_j)$ $(C(c_i, c_j))$ as the number of times that they appear in the same sentence in a document (collection). Assuming our relation extraction method the probability of a relation on a document will be 1 if the two concepts appear in a same sentence of the document and 0 otherwise.

5 Evaluation

We show here the results obtains for this methods on the corpus CLEFmed 2007 [6].

5.1 Corpus Analysis

We choose to analyses the English part of the collection with MetaMap(1). The French and the German part of the collection are analysed with TreeTagger.

For queries we propose to regroup different analysis, a query is therefore a set of graph $Q = \{G_q\}$. The probability of a query assuming a document graph model is obtained by the product of the probability of generating each graph.

$$P\left(Q = \{G_q\} | M_g\right) = \prod_{G_q} P\left(G_q | M_d\right) \tag{10}$$

We propose to group the analysis as following:

- (E) one English graph extracted by (1)
- (E_Mix) English graphs extracted by (1)(2)(3)
- (EFG) English graph extracted by (1) with French and German graph extracted by (3)
- (EFG_Mix) English graphs extracted by (1)(2)(3) with French and German graph extracted by (3)

For example, in group *EFG_Mix* a query is represented by 5 different graphs.

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

	unigram model				graph model				
		2005-2006		2007			2005-2006		2007
	λ_u	text	image	image	λ_u	λ_e	text	image	image
MAP									
E	0.2	0.2468	0.2284	0.3494	0.2	0.9	0.2463	0.2277	0.3638
E_Mix	0.1	0.2610	0.2359	0.3784	0.1	0.9	0.2620	0.2363	0.3779
EFG	0.1	0.2547	0.2274	0.3643	0.1	0.9	0.2556	0.2313	0.3733
EFG_Mix	0.1	0.2673	0.2395	0.3962	0.1	0.9	0.2670	0.2394	0.3969
P@5									
E	0.2	0.4618	0.4436	0.4533	0.2	0.9	0.4582	0.4400	0.4867
E_Mix	0.1	0.4727	0.4582	0.4767	0.1	0.8	0.4800	0.4582	0.4767
EFG	0.2	0.4582	0.4364	0.5333	0.1	0.8	0.4618	0.4473	0.5667
EFG_Mix	0.1	0.4836	0.4691	0.5	0.1	0.8	0.4909	0.4655	0.5

5.2 Results

We first evaluate our system on the two previous years of CLEFmed, we select the best performing methods at the textual level. At this level, we consider a textual annotation pertinent if one of its associated images is pertinent at the image level. Table 5.2 shows the results obtained on CLEF for the different collections, with the best parameters evaluated on the textual part of CLEFmed 2005 and 2006. We evaluate the results with mean average precision (MAP), since it gives an overview of results, and with precision at 5 documents (P@5), since this measure shows system precision on first results.

Results show that the best performing method, for MAP, is the one that uses all concept sources presented in this paper (EFG_mix). Using different concept sources for the query improves the overall results of IR. Such method helps in finding all query concepts and improves the recall. But for precision at five documents, best results are obtained with EFG that use one concept source per language. Using only the three languages provides the best concepts. Adding other concept sources may add some false concepts that lower the precision.

For graph indexing, MAP results show a similar behaviour to concepts alone. The only difference is on EFG where relation results are better than concept results, on MAP and P@5. This confirms our idea that concepts and relations are well extracted in this method that is why P@5 results are the best and why using relation improves concept results.

6 Conclusion

We proposed here a framework for using semantic resources in medical domain. We describe a method for creating graph representation of text and we propose a graph modeling approach for IR. On this framework we evaluate the impact of the multiplication of concept extraction sources on query. Results show that graph indexing can be useful for improving the first results of the system and that multiplying concept sources improves the overall results of IR. In this paper we only work on the query, in future work, we intend to evaluate multiplication of concept sources for document. Our relation extraction method is simple; using a more precise method should improve results.

References

[1] A. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21, 2001.

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

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

[4] Changki Lee, Gary Geunbae Lee, and Myung Gil Jang. Dependency structure language model for information retrieval. In *ETRI journal*, 2006.

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

[6] 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.

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

[8] 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.

[9] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321, New York, NY, USA, 1999. ACM Press.

[10] M. Srikanth and R. Srikanth. Biterm language models for document retrieval. In *Research and Development in Information Retrieval*, 2002.

[11] 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.

[12] 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.