Analysis Combination and Pseudo Relevance Feedback in Conceptual Language Model

LIRIS participation at ImageClefMed

Loïc Maisonnasse, Farah Harrathi
Laboratory LIRIS
loic.maisonnasse@insa-lyon.fr, farah.harrathi@insa-lyon.fr

Abstract

This paper presents the LIRIS contribution to the CLEF 2009 medical retrieval task (i.e. ImageCLEFmed). 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. As proposed in [6] last year, we used a conceptual representation for each sentence in the corpus and we proposed a language modeling approach on these representations. We test two versions of conceptual unigram language model; one that use the log-probability of the query and a second one that compute the Kullback-Leibler divergence. We used different concept detection methods and we combine these detection methods on queries and documents. This year we mainly test the impact of the use of additional analysis on queries. But such additional analysis does not show significant improvement. We also test combinations on French queries where we combine translation and analysis, in order to solve the lack of French terms in UMLS, this provide good results close from the English ones. To complete these combinations we proposed a pseudo relevance method. This approach use the n first retrieve documents to form one pseudo query that is used in the Kullback-Leibler model to complete the original query. The results of this approach show that extending the queries with such an approach improves the results.

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, pseudo relevance feedback

1 Introduction

The previous ImageCLEFmed tracks show the advantages of conceptual indexing (see [6]). Such indexing allows one to better capture the content of queries and documents and to match them at an abstract semantic level. On these conceptual representation [5] proposed a conceptual language modeling approach and various ways to merge different conceptual representations of documents or

queries. In this paper we reuse this approach and we extend it in various ways. The rsv value in [5] is computed through a simple query likelihood we also evaluate here the use of a Kullback-Leibler divergence as proposed in many language model approaches. Then we compare combinations of conceptual representations with the divergence rather than combinations with likelihood. In last year participation we used two analyses for documents and queries, as results presented in [5] show that combining analysis on queries is an easy way to improve the results; so we make use this year of two supplementary analysis on queries. Finally we complete this model by proposing a pseudo relevance feedback extension of queries based on our language model approach.

This paper first presents the different extension of our conceptual model. Then we detail the different documents and queries analysis. And finally we show and discuss our results obtain at CLEF 09.

2 Conceptual Model

We rely on a language model defined over concepts, as proposed in [5], which we refer to as *Conceptual Unigram Model*. We assume that a query q is composed by a set $\mathcal{C}$ of concepts, each concept being independent to the others conditionally on a document model. First we compute the rsv of this approach by simply computing the log-probability of the concept set $\mathcal{C}$ assuming a model M_d of the document d as:

$$\begin{aligned} RSV_{log}(q,d) &= \log(P(C|M_d)) & (1)\\ &= \sum_{c_i \in C} \log(P(c_i|M_d)^{\#(c_i,q)}) & (2)\end{aligned}$$

where $\#(c_i, q)$ denotes the number of times concept c_i occurs in the query q. The quantity $P(c_i|M_d)$ is directly estimated through maximum likelihood, using Jelinek-Mercer smoothing:

$$P(c_i|M_d) = (1-\lambda_u)\frac{|c_i|_d}{|*|_d} + \lambda_u \frac{|c_i|_{\mathcal{D}}}{|*|_{\mathcal{D}}} \quad (3)$$

where $|c_i|_d$ (respectively $|c_i|_{\mathcal{D}}$) is the frequency of concept c_i in the document d (respectively in the collection $\mathcal{D}$), and $|*|_d$ (respectively $|*|_{\mathcal{D}}$) is the size of d, i.e. the number of concepts in d (respectively in the collection).

In a second approach we compute the rsv of a query q for a document d by using Kullback-Leiber divergence between the document model M_d estimated over d and the query model M_q estimated over the query q, this results in:

$$\begin{aligned} RSV_{kld}(q,d) &= -\mathcal{D}(M_q \| M_d) & (4)\\ &= \sum_{c_i \in C} P(c_i|M_q) \log\left(\frac{P(c_i|M_q)}{P(c_i|M_d)}\right) & (5)\\ &= \sum_{c_i \in C} \log(P(c_i|M_q) * P(c_i|M_d)) - \sum_{c_i \in C} \log(P(c_i|M_q) * P(c_i|M_q)) & (6)\end{aligned}$$

Since the last element of the decomposition correspond to query entropy and does not affect documents ranking, we only compute the following decomposition:

$$RSV_{kld}(q,d) \propto \sum_{c_i \in C} \log(P(c_i|M_q) * P(c_i|M_d)) \quad (7)$$

where $P(c_i|M_d)$ is estimated as in equation 3. $P(c_i|M_q)$ is directly computed through maximum likelihood on the query by $P(c_i|M_d) = \frac{|c_i|_q}{|*|_q}$ where $|c_i|_q$ is the frequency of concept c_i in the query and $|*|_q$ is the size of q.

2.1 Model Combination

We present here the method used to combine different sets of concepts (i.e. concepts obtained from different analyses of queries and/or documents) with the two rsv presented above. We used the results obtain in [5] to select the best combinations on queries and documents. First, we group the different analysis of a query. To do so, we assume that a query is represented by a set of sets of concepts $Q = \{C_q\}$; and that the probability of this set assuming a document model is computed by the product of the probability of each query concept set C_q. Assuming that the first rsv RSV_{log} use the log-probability and that the second RSV_{kld} use a divergence, the combination of the rsv is computed through a sum over the different queries:

$$RSV(Q,d) \quad \propto \quad \sum_{C_q \in Q} RSV(C_q,d) \tag{8}$$

where $RSV(C_q,d)$ is either RSV_{log} (equation 1) or RSV_{kld} (equation 7).

This fusion consider that a relevant document model must generate all the possible analyses of a query Q. The best rsv will be obtained for a document model which can generate all analyses of the queries with high probability.

Second, we group the different analysis of a document $D = \{d\}$. We assume that a query can be generated by different models of the same document M_d^* (i.e. a set of models corresponding to each document d of D). Based on [5] results, we keep the higher probability among the different models, this result in:

$$RSV(Q,D) \quad = \quad argmax_{d \in D} RSV(Q,d) \tag{9}$$

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

2.2 Pseudo Relevance Feedback

Based on the n first results selected for one query set Q obtain by one RSV (equation 8), we compute a pseudo relevance feedback score PRF. This score correspond to the rsv obtain by the pseudo query Q_{fd} constitute by the merging of the n first documents retrieved with the query Q added, with a smoothing parameter, to the results obtained by the original query Q.

$$PRF(Q_{fd},d) \quad = \quad (1-\lambda_{prf})RSV(Q,d) + (\lambda_{prf})RSV(Q_{fd},d) \tag{10}$$

where $RSV(Q,d)$ is either RSV_{log} or RSV_{kld} and $RSV(Q_{fd},d)$ is the same type of rsv apply on the pseudo-query Q_{fd} that correspond to the merging of the n first results retrieved by $RSV(Q,d)$. λ_{prf} is a smoothing parameter that allows to give lower or higher importance to the pseudo query. If different collection analysis are used, we finally merge this results on documents analysis using equation 9.

3 Concepts Detection

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 allow one to derive relations between concepts in documents (and queries).

3.1 Detection Process

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;
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. [4]) and TreeTagger[1]. 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 [3]). We have not used such a refinement here and merely rely on a simple 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[2]. Such a restriction is also used in [7]. 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.

From this method we derived the two same analysis as last year *MP* and *TT* that used respectively MiniPar and TreeTagger POS analysis. We also use one detection without any morphosyntactic analysis that we named *FA*. As this method does not use a POS-tagging, the filtering of empty word is done on the basis of statistical empty word detection. This empty word detection is first based on the hypothesis that empty words are the same over different domains. So we used a corpus from another domain and we select the word witch are common with the medical domain as potential empty words. Then we combine this detection with a filtering based on the Zipf law [10] to determine the final empty word list. The fourth detection method used is MetaMap analysis [1], a tool dedicated to UMLS, that directly provide the four steps.

We finally obtain four variations of concept detection:

- (MP) uses our term mapping tools with MiniPar.
- (TT) uses our term mapping tools with TreeTagger.
- (MM) that use MetaMap.
- (FA) uses our term mapping tools without morphosyntactic analysis.

From these analyses, we use the two first one to analyse the collection and we pick some to analyse the query depending of the runs.

This year we also test this combination approach on French queries, where we first detect concepts with our term mapping tools with the French version of TreeTagger. Then we translate the French queries from French to English with Google API[3] and we extract concepts from this English translation with the MP and the TT analysis. Thus we obtain three concept sets that correspond to the French queries and we use them to query the collection.

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

[2] 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.

[3] http://code.google.com/intl/fr/apis/ajaxlanguage/documentation/

	MPTT		MMMPTT		MMMPTTFA	
	2008	2009	2008	2009	2008	2009
log-probability	0.280	0.420	0.276	-	-	0.412
KL-divergence	0.279	-	0.281	0.410	-	0.416

Table 1: Results for different query analysis combination, for the two unigram models

4 Evaluation

We train our methods on the corpus CLEFmed 2008 [2] and we run the best parameters obtained on CLEFmed 2009 corpus[8].

4.1 Model Variations

On this year collection, we submit 10 runs, these runs explore different variations of our model. Previous year results show that merging queries improves the results, we test this year the impact of adding new analysis only on the queries.

So we first test 3 model variations:

- (UNI.log) that use the conceptual unigram model (as define in 1).
- (UNI.kld) that use the conceptual unigram model with the divergence (as define in 7).
- (PRF.kld) that combine the conceptual unigram model with a pseudo relevance feedback (as define in 10).

For each model, we test it on the collection analysed by two detection methods, MiniPar and TreeTagger (MPTT), using the model combination methods proposed in section 2.1 and we test it with the three following query analysis:

- (MPTT) that groups MP and TT analysis,
- (MMMPTT) that groups the two preceding analysis with MM one,
- (MMMPTTFA) that groups the three preceding analysis with FA one.

4.2 Results

From each method we use the bests parameters obtained on ImageCLEFmed 08 corpus for MAP and we use these parameters on the new 09 collection. We first compare the variation between the results on the two rsv define for MAP and for different query merging on, table 1.

Results show that the two rsv give close results on 2008 queries. On 2009 queries, our best result is obtained with the log-probability and with two analyses (MPTT) on the query. Using the four analyses (MMMPTTFA), the log-probability is slightly better than the KL-divergence but the results are close

As presented before, we test our combination model on French queries, from these queries we obtain different concept sets by merging detection methods and by translating, or not, the query to English in order to find the UMLS concepts that are not linked with French terms. This method obtains the good results of 0.377 in MAP. This shows that the combinations methods can be used on translation methods.

We then test our pseudo relevance feedback method for this we query with RSV_{kld} and we process the relevance feedback, the results are presented in table 2. The results, we achieve on 2008 queries, show that the best results are obtain with the pseudo query build on the 100 first documents initially retrieve. On 2008, merging more analysis of the query improve the results. Transposed to 2009 the results also show good results, but the best results are obtained by using only two analyses (MPTT).

size of the pseudo query (n)	MPTT 2008	MPTT 2009	MMMPTT 2008	MMMPTT 2009	MPTTFA 2009	MMMPTTFA 2009
20	0.279	-	0.281	-	-	-
50	0.289	-	0.290	-	-	-
100	0.292	**0.429**	0.299	0.416	0.424	0.418

Table 2: Results for different size of pseudo relevance feedback with the Kullback-Leiber divergence and with different query analysis

5 Conclusion

Using the conceptual language model provides good performance in medical IR, and merging conceptual analysis is still improving the results. This year we explore a variation of this model by testing the use of a Kullback-Leiber divergence and we improve it by integrating a pseudo relevance feedback. The two model variations provide good but similar results. Adding a pseudo relevance feedback improves the results providing the best MAP results for 2009 CLEF campaign. We also made an experimentation on French queries where we use the combination method to solve the 'lack' of French terms in UMLS, this results show that combination methods can also be used on various methods of concepts detection.

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] Müller H., Kalpathy-Cramer J., Kahn Jr. C., Hatt W., Bedrick S., and Hersh W. Overview of the ImageCLEFmed 2008 medical image retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Aarhus, Denmark, September 2008.

[3] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 1999.

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

[5] Eric Gaussier Loic Maisonnasse and Jean Pierre Chevallet. Model fusion in conceptual language modeling. In *ECIR 2008*, 2008.

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

[7] 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.

[8] Henning Mller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Sad Radhouani, Brian Bakke, Charles Kahn Jr., and William Hersh. Overview of the clef 2009 medical image retrieval track. In *Working Notes of the 2009 CLEF Workshop*, Corfu, Greece, September 2009.

[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] George K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley (Reading MA), 1949.