Analysis combination and pseudo relevance feedback in conceptual language model

Loïc Maisonnasse, Farah Harrathi, Catherine Roussey, Sylvie Calabretto

Author-produced version of the paper published in Lecture Notes in Computer Science, 2010, vol.6242, pages 203-210.
Multilingual Information Access Evaluation II. Multimedia Experiments
10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers
Original Publication available at http://www.springerlink.com/content/p04856228257
doi:10.1007/978-3-642-15751-6

Analysis Combination and Pseudo Relevance Feedback in Conceptual Language Model

LIRIS participation at ImageCLEFMed

Loïc Maisonnasse[1], Farah Harrathi[1], Catherine Roussey[1,2], and Sylvie Calabretto[1]

[1] Université de Lyon, CNRS, INSA de Lyon, Université Lyon 1, LIRIS UMR5205,
Villeurbanne, France
firstname.lastname@liris.cnrs.fr
[2] Cemagref, 24 Av. des Landais, BP 50085, 63172 Aubiére, France

Abstract. This paper presents the LIRIS contribution to the CLEF 2009 medical retrieval task (i.e. 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 and we proposed a language modeling approach. 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. 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.

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 that handle different conceptual representations of documents or queries. In this paper we extend this approach 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. One of them is a new method of concept detection using only statistical methods. 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 extensions 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 of 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:

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

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

$$\begin{aligned} RSV_{kld}(q,d) = -\mathcal{D}\left(M_q \| M_d\right) &= \sum_{c_i \in C} P(c_i|M_q) \log\left(\frac{P(c_i|M_q)}{P(c_i|M_d)}\right) \\ &= \sum_{c_i \in C} P(c_i|M_q) * \log(P(c_i|M_d)) - \sum_{c_i \in C} P(c_i|M_q) * \log(P(c_i|M_q)) \end{aligned} \tag{2}$$

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

$$RSV_{kld}(q,d) \propto \sum_{c_i \in C} P(c_i|M_q) * \log(P(c_i|M_d)) \tag{3}$$

where $P(c_i|M_d)$ is estimated as previously. $P(c_i|M_q)$ 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} uses the log-probability and that the second RSV_{kld} uses a divergence, the combination of the rsv is computed through a sum over the different queries:

$$RSV(Q, d) \propto \sum_{C_q \in Q} RSV(C_q, d) \tag{4}$$

where $RSV(C_q, d)$ is either RSV_{log} (equation 1) or RSV_{kld} (equation 3). With this fusion,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 d 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:

$$RSV(Q, D) = argmax_{d \in D} RSV(Q, d) \tag{5}$$

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 4), 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.

$$PRF(Q_{fd}, d) = (1 - \lambda_{prf})RSV(Q, d) + (\lambda_{prf})RSV(Q_{fd}, d) \tag{6}$$

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 these results using equation 5.

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 Linguistic Detection Process

The detection of concepts based on linguistic analysis of document from a thesaurus is a relatively well established process. It consists of four major steps (refer to [5] for details on these 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.

3.2 Statistical Detection Process

We develop a statitical method of concept detection that could be apply on several languages without any linguistic analysis. This method replace the morphosyntactic analysis (step 1 and 2 of previous section) by statistical method. Our method is composed of four main steps:

1. Empty Word and Simple Term Extraction based on corpus analysis.
2. Compound Term Extraction.
3. Concept Detection
4. Concept Filtering

The last two steps (3 and 4) are similar to the linguistic detection process, thus we will not describe them in the next paragraphs.

Empty Word and Simple Term Extraction Empty words are words that have no discriminate power to identify a specific document over a corpus, because they have a linear distribution over all the documents. They can be stop words or general word like the day of the week and so one. In order to extract the empty word of the document we use two corpora: The indexing corpus and the support corpus. The support corpus should have the same languages than the indexing

corpus but should deals with another domain. For example in our experiment the indexing corpus is about medicine, the support corpus is about laws (the European Parliament collection[3]). We define empty word as a word belonging to the indexing corpus and the support corpus and its frequency inside the two corpora should be above a threshold fixed by experience. Simple terms are the words of the indexing corpus which are not detected as empty word.

Compound Terms Extraction We assume that compound term (term composed of more than one word) is a kind of word collocation. According to [1] we can detect words involved in a collocation by following two assumptions. (1) The words must appear together significantly more often than expected by chance. (2) The words should appear in a relatively rigid way because of syntactic constraints. [3] uses the Mutual Information (MI) measure to extract a collocation of two words. Unfortunately the MI measure is not able to extract compound terms composed of empty words and it is not adapted to extract compound terms of more than two words. Thus we propose to adapt the Mutual Information measure to avoid these two drawbacks. Considering two words m_1 and m_2, our formula of the $AdaptedMutualInformation(AMI)$ is:

$$AMI(m_1, m_2) = \begin{cases} \log_2\left(\frac{P(m_1,m_2)}{P(m_1)^2}\right) & if\ m_2\ is\ an\ empty\ word \\ \log_2\left(\frac{P(m_1,m_2)}{P(m_1)*P(m_2)}\right) & otherwise \end{cases} \quad (7)$$

Where $P(m_1)$ is estimated by counting the number of observations of m_1 in the collection and normalizing by N, the size of the collection. $P(m_1, m_2)$ is estimated by counting the number of times that m_1 is followed by m_2 and normalizing by N.

The term extraction process is iterative and incremental process. The compound terms of the iteration $i+1$ (that is to say that their length is $i+1$ words) are built from the term of the iteration i (their length is i words). The extraction process starts from the simple terms composed of 1 word. For each couple of words (simple term + another word of the indexing corpus), we compute the AMI. If its AMI is above a threshold, this new compound term is added in the starting list of the next iteration of this process. In our experiment we fixed the AMI threshold to 15. The iterations carry one as far as a new compound term is extracted.

3.3 Linguistic Detection versus Statistical Detection

We test our new statistical concept detection process on the collection CLEFmed 2007 using UMLS. We compare the statistical detection from those obtained by [4] using linguistic techniques with the similar collection and UMLS. In [4] , three linguistic analyzers are used prior to the concept detection: MetaMap *MM*, MiniPar *MP* and TreeTagger *TT*. The results obtained by these various analyzers as those obtain by our statistical method, that we named *FA*, are

[3] http://www.statmt.org/europarl/

given in Table 1. We note the linguistic methods perform slightly better for MAP, and the statistic method perform better for P@5. Thus we can conclude that our statistical method of concept detection has similar results than those using linguistic techniques.

Table 1. comparison of statistical versus linguistic concept detection using CLEFMed 2007

Method	analysis	MAP	P@5	Δ MAP	Δ P@5
Linguistic	MM	0.246	0.357	-0.81%	19.05%
	MP	0.246	0.424	-0.81%	0.24%
	TT	0.258	0.462	-5.43%	-8.01%
Statistical	FA	0.244	0.425		

3.4 Our Four Detection Processes

Due to the previous results we use the fourth analyses in our experiments. ¿From these analyses, we use the MP and TT ones 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[4] and we extract concepts from this English translation with the MP and the TT analysis.

4 Evaluation

We train our methods on the corpus CLEFmed 2008 and we run the best parameters obtained on CLEFmed 2009 corpus[2]. 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 3).
- (PRF.kld) that combine the conceptual unigram model with a pseudo relevance feedback (as define in 6).

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

4.1 Results

Table 2. Results for different query analysis combination, for the two unigram models

	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

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 2. 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 3. 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).

Table 3. Results for different size of pseudo relevance feedback with the Kullback-Leiber divergence and with different query analysis

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

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] Smadja F.A. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[2] Müller H., Kalpathy-Cramer J., Eggel I., Bedrick S., Radhouani S., Bakke B., Kahn C.Jr., and Hersh W. Overview of the clef 2009 medical image retrieval track. In *Working Notes of the 2009 CLEF Workshop*, Corfu, Greece, September 2009.

[3] Church K.W. and Hanks P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[4] Chevallet J.P. Maisonnasse L., Gaussier E. Combinaison d'analyses sémantiques pour la recherche d'information médicale. In INFORSID, editor, *Atelier RISE (Recherche d'Information SEmantique) dans le cadre de la conférence INFORSID'2009*, May 2009.

[5] Gaussier E. Maisonnasse L. and Chevallet J.P. Model fusion in conceptual language modeling. In *ECIR 2008*, 2008.

[6] Gaussier E. Maisonnasse L. and Chevallet J.P. Multiplying concept sources for graph modeling. In *CLEF 2007, LNCS 5152 proceedings*, 2008.