EMSE at TREC 2015 Clinical Decision Support Track

Bissan AUDEH and Michel BEIGBEDER

Mines Saint-Étienne
158, cours Fauriel
CS 62362
F 42023 Saint-Étienne Cedex 2 - France
http://www.mines-stetienne.fr/

Abstract. This paper describes the participation of the EMSE team at the clinical decision support track of TREC 2015 (Task A). Our team submitted three automatic runs based only on the summary field. The baseline run uses the summary field as a query and the query likelihood retrieval model to match articles. Other runs explore different approaches to expand the summary field: RM3, LSI with pseudo relevant documents, semantic ressources of UMLS, and a hybrid approach called SMERA that combines LSI and UMLS based approaches. Only three of our runs were considered for the 2015 campaign: RM3, LSI and SMERA.

Keywords: Query Expansion, Latent semantic analysis, Ontology-based query expansion

1 Introduction

Browsing the state of the art of query expansion reveals an overwhelming amount of theories [1]. If the retrieval model is precise enough to detect relevant documents at high ranks, approaches based on pseudo relevance feedback perform quite well without user intervention. On the other hand, most of these approaches depend on word-based statistical calculation, which makes them unable to explicitly introduce phrases or multi-word named entities (assuming a word-based indexation). This issue can be addressed by ontology based techniques. Using an external resource provides the system with semantic information which leads to valuable expansion terms that can not necessarily be obtained by feedback documents.

Our participation in the clinical decision track aims to evaluate a Semantic Mixed Expansion and Reformulation Approach (SMERA) in the medical context. This approach to query expansion uses ontologies (UMLS) and a local approach based on pseudo relevant feedback documents using LSI. A brief description of our submitted runs is given in section 2. A detailed explanation about our proposed approaches is given in section 3 for the LSI approach, and section 4 for the hybrid approach SMERA.

2 Our runs

We submitted three runs to the task A in the clinical decision track of TREC 2015:

EMSE_SumRM3 : Query expansion using pseudo relevance feedback with a language model [2];

EMSE_LSI : Query expansion with pseudo relevance feedback documents using LSI (cf. Sect. 3);

SMERA : A mixed query expansion and reformulation approach that uses a combination of LSI and an ontology based query expansion approaches (cf. Sect. 4).

Our query reformulation method considers the final query to be a linear combination of the user's original query terms and the representations of the expansion sets obtained in the expansion step. The relevance score value can thus be expressed by equation 1:

$$p(Q|d) = \lambda \prod_{q} p(q|d) + (1-\lambda) \prod_{i=1}^{k} b(r_i)^{w_i} \tag{1}$$

where $p(q|d)$ is the query likelihood probability for the original query term q and a document d, r_i corresponds to an expansion set that is associated to at least one original query term, $b(r_i)$ is the belief calculated for this expansion set according to the Metzler's approach [3], finally w_i is the weight of the estimated belief of the representation r_i. In this current participation, expansion sets are considered to be equally important to the query so w_i was set to one for all i.

3 EMSE-LSI approach

Several approaches exist to extract concepts from a set of documents (like LDA, ESA or LSI). In this study we chose to apply LSI on pseudo relevant feedback documents. It was argued that LSI can detect high level co-occurrence relationships between terms. This means that two terms that do not occur together in the studied set of documents but frequently co-occur with a third term will be considered as semantically related by LSI. The idea is to do singular value decomposition on a matrix A of m lines (corresponding to m terms) and n columns (corresponding to n feedback documents) that contains the frequencies tf of the terms in the feedback documents. The result of this step are the three matrixes presented in equation 2:

$$A_{\{m,n\}} = U_{\{m,m\}} S_{\{m,n\}} V^T_{\{n,n\}} \tag{2}$$

where S is the diagonal matrix that contains the singular values of A. The theory of LSI is that reducing the dimension of the three resulting matrixes gives an approximation of the original matrix A while reducing the noise (equation 3).

$$A'_{\{m,n\}} = U_{\{m,k\}} S_{\{k,k\}} V^T_{\{k,n\}} \tag{3}$$

For our case of query expansion, we are interested in the matrix $U_{\{m,k\}}$. This matrix contains the m vectors of terms appearing in pseudo relevance feedback documents. These vectors belong to the semantic space of k dimensions created by LSI (cf. Fig. 1). To find the expansion set of a query term q, we measure its

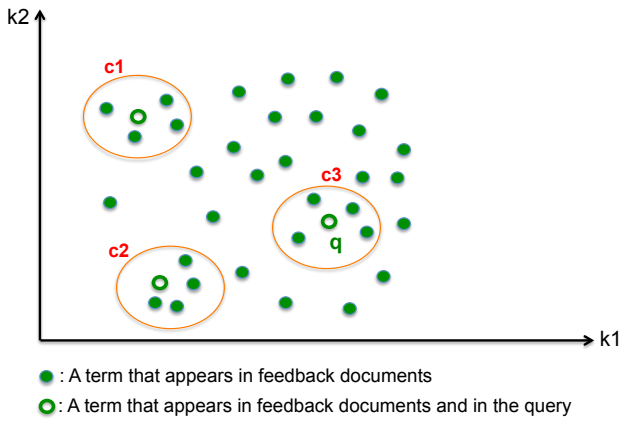

Fig. 1. Terms of feedback documents in the semantic space of LSI (example for the case of 2 dimensions k1 et k2)

similarity with the m terms that appear in the feedback documents based on the euclidean distance. We then suppose that the terms that are the most similar to q belong to the same implicit concept, as presented in Fig. 1. In some cases, two original query terms are strongly related to each other. These two terms will have the same statistics in the feedback documents, and obtain identical vectors in the semantic space generated by LSI (cf. Fig. 2). In this case, we consider that these original query terms belong to the same implicit concept (c2 in Fig. 2) and will both correspond to one expansion set in the reformulated query.

In our run in TREC 2015 we used query likelihood language model [4] to retrieve pseudo relevance documents. Twenty documents were used to construct matrix A. For LSI, 10 dimensions were considered. λ was tuned to 0.5 (cf. Equation 1) and sets of three expansion terms were built.

4 EMSE-SMERA approach

SMERA is a mixed approach that combines both the LSI method with pseudo relevance feedback documents, and a semantic method based on UMLS concepts. The LSI-based method was used only to expand summary terms that can't be matched to UMLS concepts. Medical terms are disambiguated using MetaMap, which results in finding unique concepts in the UMLS semantic ressources. Concepts names and "preferred names" are then used as expansion terms and added

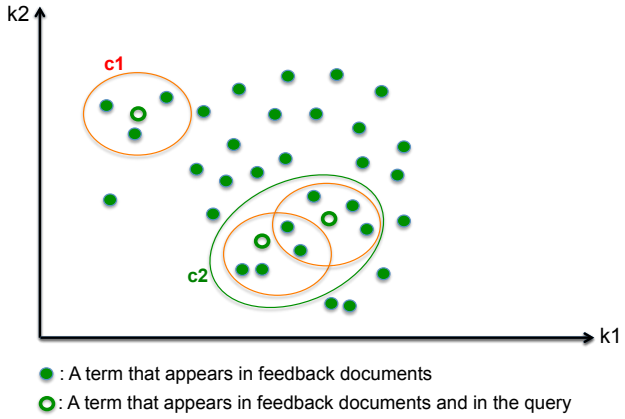

Fig. 2. The merging of expansion sets in the case of query terms that are semantically close in LSI semantic space

to the reformulated query. Temporal concepts were not explicitly eliminated with this approach. Parameters of this run were fixed as followed : for the LSI part of the approach we used 20 documents, 5 dimensions and 3 expansion terms; for the UMLS part we used the matched concept name (retrieved by MetaMap) and the preferred concept name as the expansion term, λ was also set to 0.5.

References

1. Carpineto, C., Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys 44(1), 1–50 (Jan 2012)
2. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 120–127. ACM Press, NY, USA (2001)
3. Metzler, D., Croft, W.B.: Combining the language model and inference network approaches to retrieval. Inf. Process. Manage. 40, 735–750 (Sep 2004)
4. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 275–281. ACM (1998)