

# Hybrid query expansion using linguistic resources and word embeddings

Nesrine Ksentini, Siwar Zayani, Mohamed Tmar and Faiez Gargouri

MIRACL Laboratory, City ons Sfax, University of Sfax, B.P.3023 Sfax TUNISIA

## Abstract

Since the large amount of textual data on the web, traditional information retrieval and query expansion techniques such as pseudo-relevance feedback are not always helpful to optimize the retrieval process. In this paper, we study the use of term relatedness in the context of query expansion with an hybrid approach based on linguistic resources and word embeddings such as the distributed neural language model word2vec. We perform experiments on Cystic Fibrosis Database. Obtained results are more robust than the baseline research system.

## Keywords

Semantic relationships, Word embeddings, Word2Vec, Skip-Gram, MeSH thesaurus, Information retrieval

## 1. Introduction

Information Retrieval System (IRS) has attracted an increasing research attention with the proliferation of web data. The major challenge of the IRS is to return relevant documents that meet user's need (even they do not contain the query terms) and reject irrelevant documents (even they contain the query terms). Query expansion becomes an important task to ameliorate IRS results. Query Expansion methods based on Pseudo Relevance Feedback (PRF) are widely used and rely much on the assumption that the top ranked documents in the initial search are relevant and contain good terms for query expansion [1]. This assumption is not always checked. To overcome this limitation, it will be needed to define semantic relatedness between terms either by using linguistic resources either by using statistical methods in order to select relevant terms [2, 3, 4, 5, 6, 7].

Several semantic resources have been proposed, making it possible to model domain knowledge such as dictionaries, taxonomies, ontologies and thesaurus. For statistical methods, we focus in this paper on word embeddings which is a mapping that associates each word in a document collection to a vector representation with a size is significantly lower than the size of the vocabulary of the document collection [8, 9, 10]. By adding the original query with the expansion words derived from word embeddings, we could better present the users information need in specific topic.

The remainder of this paper is organized as follows. In section 2, we present a literature review of the different works to determine semantic relationships by hybrid approaches.

---


TACC'22: Tunisian-Algerian Joint Conference on Applied Computing, December 13–14, 2022, Constantine, Algeria

✉ ksentininesrine@gmail.com (N. KSENTINI); zayani.siouar@gmail.com (S. ZAYANI);

mohamed.tmar@isimsf.rnu.tn (M. TMAR); faiez.gargouri@isims.usf.tn (F. GARGOURI)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In section 3, we focus on describing details of the proposed hybrid approach in the query expansion process.

In section 4, the evaluation process is presented and discussed. We draw at the end the conclusion and outline future works in section 5.

## 2. SEMANTIC RELATIONSHIPS BY HYBRID APPROACHES

Defining semantic relationships between terms has become a primary task in order to improve the performance of IRS [11, 12]. Several works have been proposed, can be classified into three main categories: those which are based on external semantic resources such as ontologies, thesaurus . . . , those which are based on methods based on the content of documents using statistical measures [13] and hybrid approaches combining these first two categories.

In this section, we focus on defining semantic relationships by hybrid approaches.

### **Works using hybrid approaches:**

Hybrid or mixed methods integrate both knowledge from a corpus of documents and external semantic resources are widely used to detect and evaluate semantic relationships.

In [11], the authors combine the results obtained by the method based on external semantic resources and the method based on the content of returned documents to define semantic relationships between terms.

In fact, for the first method, the system determines the terms that are related to the term of the initial query using the WordNet thesaurus which is a lexical database [14] developed by a team of experts from the cognitive science laboratory.

It is the most famous lexical resource in the English language.

The terms in this database are organized as sets of synonyms that represent concepts called synsets [15, 16]. Each synset represents a specific meaning of a word. Generally, each term can be associated with one or more synsets (polysemy). These are classified by several types of semantic relations (hyponymy, hyperonymy, meronymy and Holonymy).

Figure 1 shows a graphical visualization of the word "Calcium" in WordNet <sup>1</sup>

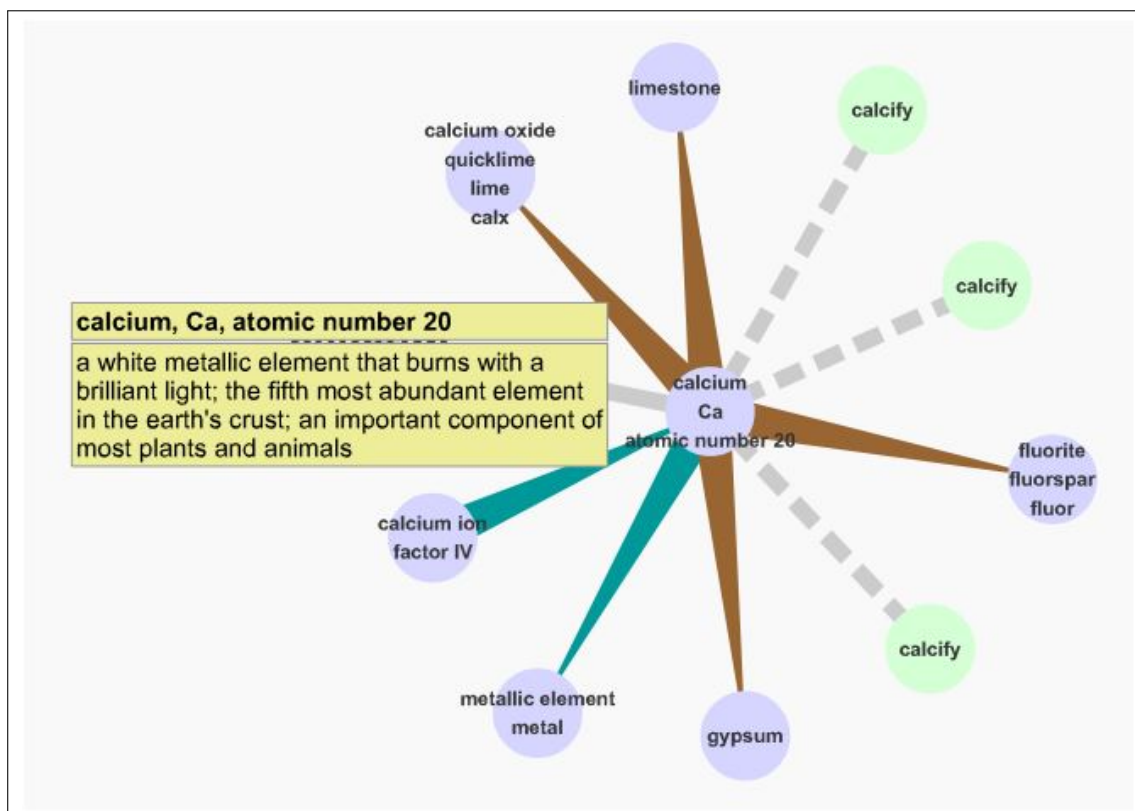
In [17], authors explore the Skip-gram model to determine the semantic relationships between medical concepts. Unlike the traditional Skip-gram model which allows the creation of distributed vector representations of words, the model proposed in this study exploits the distributed representations of UMLS concepts extracted from medical corpora, including clinical records and abstracts from medical journals.

The proposed system attempts to determine which terms are related by applying the context method. They collect snippets containing the first term searched by the search engine. Then they determine a context around these terms, replacing the term sought by the second term. Finally, they assume that the two terms are synonymous if there is no change of context.

The authors of [18] extend the CBOW model to learn distinct representations of different meanings of a term by aligning them to terms in the WordNet database. For this, a revised

---

<sup>1</sup><http://www.snappywords.com/>



**Figure 1:** Graphical visualization of the word "Calcium" in WordNet

architecture of the CBOW algorithm is proposed which allows to jointly learn in the same space both the term and the different associated candidate senses.

In [19], authors propose a hybrid approach using two different sources: the UMLS model and the word2vec word vector model.

They use the natural language processing tool *cTAKES* to identify medical concepts present in a query using UMLS. They explore embedded words using the Skip-gram architecture to find the two closest terms in the original query. The size of the vector was set at 1000 and a vocabulary of 25,469 vectors were included in this model.

In [20], authors propose a method making it possible to link the terms used by patients extracted from the forum on "cancerdusein.org" to those used by professionals in the medical field in a vocabulary devoted to breast cancer developed by the 'National Cancer Institute INCa.

The originality of this approach is to use texts written by patients (PAT) collected on forums, to build a consumer health vocabulary (CHV) in French. In fact, this method is structured in six steps:

- Constitution of the corpus:  
Use of the "cancerdusein.org" forum on breast cancer, where most of its members are patients or their loved ones. This forum facilitates the sharing of information with other patients.
  
- Extraction of the candidate terms:  
Using the BioTex [21] tool to search for corpus terms that belong to the medical domain to obtain a set  $T = t_1, \dots, t_n$ .
  
- Spelling correction:  
This step corresponds to correct the spelling of all  $t_i \in T$  terms using Aspell software in order to obtain a set  $M = m_1, m_2, \dots, m_k$ , where  $k$  is the number of correction proposals for a term  $t_i$ .  
The Levenshtein distance is used to compare the term  $t_i$  and  $m_j$ , choosing only terms with a distance to  $t_i$  less than or equal to 2.
  
- Abbreviations:  
In the same way as the previous step, they seek in the whole set  $T$  those which correspond to the abbreviations by adapting the Carry padding algorithm to using a list of common suffixes used in the biomedical field.  
For a term  $t_i$  belonging to  $T$ , we obtain a set  $A = a_1, a_2, \dots, a_k$ , where  $k$  is the number of proposed abbreviations included in *INCa*.
  
- Similarity between two terms:  
In this step, the authors determine the similarity between terms by using 3 methods:
  - Consider a semantically structured resource (*Wikipedia*).
  - Consider the co-occurrences of terms from documents indexed by the Google search engine (standard Google similarity).
  - Consider co-occurrences in patient messages using the Jaccard measurement.
  
- Formalisation in SKOS:  
Finally, the authors use the relationships obtained in the previous steps to create a SKOS ontology (Simple Knowledge Organisation System). This ontology associates an *INCa* term with the different terms of the patient: preferential terms are used to define the term MeSH representing the expert term, alternative terms are used to represent abbreviations and hidden terms are used to represent spelling errors.  
In fact, this method is applied to the field of breast cancer and experienced for the French language, but it can be applied to many other areas and can be adapted to other languages.

Hybrid approaches represent a compromise between language approaches in using knowledge bases and other statistics.

The latter exploits the precision of linguistic approaches and the robustness of statistical approaches. By comparing them with statistical approaches, hybrid approaches are faster and

more independent. The use of language resources at the level of hybrid approaches, makes it possible to obtain results more satisfying the needs of users. Indeed, the intervention of linguists makes it possible to reduce the noise that can be generated by statistical approaches.

### 3. Semantic relationships in the query expansion process

IRS have evolved with the appearance of Semantic Web and aim to exploit semantic relationships between terms in order to enrich the user's initial query. To ameliorate the user's query, we integrate defined semantic relationships between terms by our proposed hybrid approach in the query expansion process based on the pseudo relevance feedback technique (PRF).

#### Hybrid definition of semantic relationships: MeSH + Word2Vec:

In this subsection, we present the combination between the linguistic definition and the statistical definition of semantic relationships. Thus, the study of defined relationships is based on the following assumption: first, we search from MeSH thesaurus, synonyms for terms in the initial query (linguistic definition). Then we define their vectors representation (see figure 2) resulting from the Skip-gram algorithm (statistical definition).

Terms	Vectors Representation													
cystic	0.18599606	0.22540125	-0.046698514	0.11430553	0.36236343	-0.19119236	-0.10555536	-0.10811272	0.2254146	0.10366014	-0.010828005	0.21097365	-0.100658424	-0.19779243
fibrosi	-0.10811272	0.2254146	0.10366014	-0.010828005	0.21097365	-0.100658424	-0.19779243	0.35037935	0.11423119	0.19568369	-0.055175524	0.22044155	-0.113615505	0.038478594
cf	0.35037935	0.11423119	0.19568369	-0.055175524	0.22044155	-0.113615505	0.038478594	-0.2511047596	0.26173985	0.031057462	0.04313654	-0.12124751	0.3882619	0.20107809
normal	-0.2511047596	0.26173985	0.031057462	0.04313654	-0.12124751	0.3882619	0.20107809	0.0727003	0.08327702	-0.02995708	0.15449324	0.04240128	-0.0015069582	0.0026954552
children	0.0727003	0.08327702	-0.02995708	0.15449324	0.04240128	-0.0015069582	0.0026954552	0.4161327	-0.32635412	-0.12265513	0.2856312	0.10719238	-0.1603939	0.105236664
serum	0.4161327	-0.32635412	-0.12265513	0.2856312	0.10719238	-0.1603939	0.105236664	0.030742992	0.16926423	-0.06601683	0.12134246	0.1996248	0.13421497	-0.083949916
studi	0.030742992	0.16926423	-0.06601683	0.12134246	0.1996248	0.13421497	-0.083949916	0.47293314	-0.085278206	0.1342288	-0.18496533	0.34124145	0.4385408	0.2514374
control	0.47293314	-0.085278206	0.1342288	-0.18496533	0.34124145	0.4385408	0.2514374	0.24256262	0.075220086	0.07983794	0.2719032	-0.2058306	-0.23297745	-0.03596104
test	0.24256262	0.075220086	0.07983794	0.2719032	-0.2058306	-0.23297745	-0.03596104	0.4965191	-0.18736593	0.06834479	-0.122876234	0.3022793	0.11356426	0.009397611
active	0.4965191	-0.18736593	0.06834479	-0.122876234	0.3022793	0.11356426	0.009397611	0.34809026	0.06532959	0.0567207	-0.10079029	0.18913706	0.36804911	0.16879651
differ	0.34809026	0.06532959	0.0567207	-0.10079029	0.18913706	0.36804911	0.16879651	0.41814768	0.16404095	-0.19591479	-0.13250178	0.52890223	0.5857505	0.18373112
cell	0.41814768	0.16404095	-0.19591479	-0.13250178	0.52890223	0.5857505	0.18373112	-0.027659873	0.15454072	0.0135364225	0.2046573	0.037586667	0.15366876	-0.4013769
pancreat	-0.027659873	0.15454072	0.0135364225	0.2046573	0.037586667	0.15366876	-0.4013769	0.33730823	-0.089272745	0.053515874	-0.095208645	0.19902758	0.19564529	0.07590027
group	0.33730823	-0.089272745	0.053515874	-0.095208645	0.19902758	0.19564529	0.07590027	0.20188661	-0.028763868	0.12044476	0.24784994	0.08363915	0.19283669	0.095664345
result	0.20188661	-0.028763868	0.12044476	0.24784994	0.08363915	0.19283669	0.095664345	-0.06826684	0.17533961	-0.18171635	-0.10992673	-0.1940582	-0.06927003	-0.17068148
clinic	-0.06826684	0.17533961	-0.18171635	-0.10992673	-0.1940582	-0.06927003	-0.17068148	0.5677222	-0.31778306	-0.05139764	0.29302293	0.060964506	-0.045650695	0.030489782
level	0.5677222	-0.31778306	-0.05139764	0.29302293	0.060964506	-0.045650695	0.030489782							

Figure 2: Vectors representation

Afterwards, we measure the semantic similarity between them by calculating the cosine between their corresponding term vectors.

Finally, we take only  $k$  first terms that have the largest cosine values to define a word bag for the query expansion process.

### 3.0.1. Semantic relationships with Mesh

Define semantic relationships between terms is paramount to improve user's queries and search quality. In our case, we try to find synonymy relations between terms of the initial query and MeSH thesaurus concepts with three methods:

- scopeNote method: Method based on concepts descriptions extracted from the MeSH thesaurus. Indeed, for each term in the initial query, we select its description which represents the medical definition of term.
- termList method: Method is based on the list of associated terms (TermList). This method try to select synonymous terms that are semantically linked to terms of the initial query.
  - If a term of the initial query is a MeSH concept, we take the list of synonymous terms linked to this concept.
  - If a term of the initial query is a MeSH term, we take its parent concept.
- fusion method: This third method is to mix the two previous methods. Indeed, we choose to use this method to add more semantically related terms to the context of initial query.

### 3.0.2. Semantic relationships with Word2Vec

The Word2vec model proposed by [8], is based on an neural network to learn vectors terms representations and to detect synonymous terms or suggest additional terms [9].

Word2vec is a group of related models that are used to produce word embeddings.

This model is based on a simple neural architecture and computational simplifications through mathematical expressions, allowing the exploitation of a very large amount of textual data to learn them. Indeed, Word2vec takes as its input a large corpus of text and produces a vector space for each unique word in the corpus.

This model has different parameters, the most important of which are:

- The choice of the learning model (1 for the Skip-Gram model and 0 for the CBOW model)[8].
- The dimension of the vector space to be constructed: it represents the number of numerical descriptors used to describe terms in the corpus.[8].
- The size of the context window of a term: it represents the number of terms surrounding the word in question (authors in [8] suggest using contexts of size 10 with the Skip-Gram architecture and 5 with CBOW architecture).

In our case, we have trained word2vec model using the gensim library [22]and the Skip-Gram architecture. The basic idea of the Skip-Gram architecture is to use the current word in order to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words.

For example, if we have a vocabulary represented by this set of words (pseudomonas, aeruginosa, infection, cystic, fibrosis), and the target word is "infection".

The Skip-Gram architecture is as follows (see figure 3):

As input layer, we find the target word "infection" with its binary representation  $X$  whose length is equal to 5 (the size of the vocabulary in this example).

As output layer, we find four binary vectors corresponding to the words of the context: “pseudomonas”, “aeruginosa”, “cystic”, “fibrosis”.

The projection layer (hidden layer)  $h$  is represented by the weight matrix  $W$  which rows present words in a vocabulary and columns present hidden neurons. Before training step, the  $W$  matrix is initialized with small random values.

We can calculate a score for each word of the vocabulary which represent a correspondence measure between the context window and the target word. This score is calculated by the scalar product between the predicted representation and the target word representation.

Subsequently, we use the hierarchical Softmax activation function to determine which words are similar to the target word. This prediction is then corrected using backpropagation for each words in the context window.

Indeed, we use backpropagation to find the optimal weights of a neural network. These weights make it possible to minimize the loss function by applying the gradient descent algorithm.

This backpropagation makes it possible to correct the global matrix by bringing words of their respective contexts. Finally, vectors resulting from the learning step are used to define semantic relationships.

## 4. Evaluation

We explored the effectiveness of our proposed method on the standard ad-hoc task using a *CysticFibrosis*<sup>2</sup> database of around 1000 Documents. The evaluation model of search system is based on the evaluation model of the Cranfield project.

### 4.1. Cystic Fibrosis Database

The Cystic Fibrosis Database (CF) is composed of 1239 documents discussing Cystic Fibrosis Aspects, and a set of 100 queries with the respective relevant documents as answers. Documents in this collection focus on cystic fibrosis disease. They present the symptoms, diagnoses, and treatments of this disease.

### 4.2. Results

We have proposed an hybrid approach to define semantic relationships between terms in order to improve search results. Our approach based on the combination between the linguistic definition and the statistical definition of semantic relationships. Since, relationships are defined, we try to they exploit in the expansion query process.

We present results for retrieval experiments in Table 1 for both methods of semantic relationships definition.

We find from these results that better results are obtained when to use scopeNote and termList to define semantic relations for linguistic method, and the Skip-Gram model using the Softmax activation function for statistical method.

In order to improve the results obtained, we combined the linguistic method (fusion) with the

---

<sup>2</sup><https://people.ischool.berkeley.edu/~hearst/irbook/cfc.html>

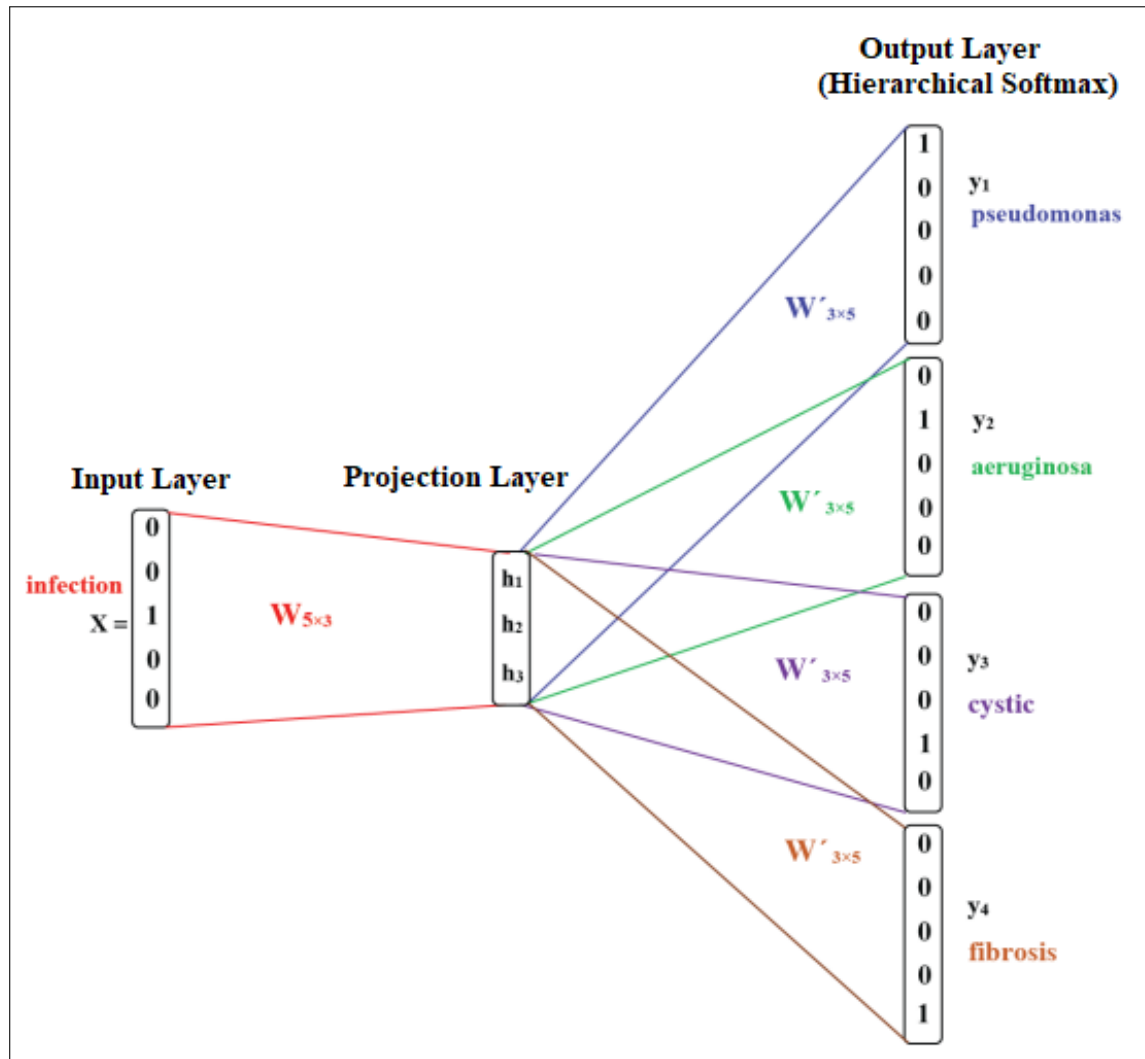


Figure 3: skip-Gram architecture

statistical method (Skip-Gram) to define semantic relationships and subsequently integrate them into the expansion process. We add only related terms to the initial query based on the first 5 returned documents by adopting the Pseudo Return of Relevance (PRF) technique.

Table 2 shows obtained results before and after query expansion.

In order to check and validate the influence of using defined semantic relations on the performance of our IRS, we use student test tool [23] whose p-value is less than or equal to 0.05. This tool allows to compare the means of two groups of samples. This validation is indicated by (\*) for obtained results in table 2. We note from obtained results in table 2, an improvement of recall and precision with a relevance rate equal to 11.23%. (According to [24], from 5%



**Table 1**  
Obtained Results

	MAP (Mean Average Precision)	Recall
Linguistic Method		
ScopeNote	0.0867	0.3268
TermList	0.0947	0.3396
<b>Fusion (ScopeNote+ TermList)</b>	<b>0.1008</b>	<b>0.3409</b>
Statistical Method		
<b>Skip-Gram + Softmax</b>	<b>0.0990</b>	<b>0.3706</b>
Skip-Gram + negative sampling	0.073	0.3365
CBOW + Softmax	0.0686	0.3220
CBOW + negative sampling	0.0685	0.3154

**Table 2**  
Obtained results

	System	system with expansion
Retrieved documents	4786	4786
Pertinent Retrieved documents	1637	1738
MAP	0.1406	<b>0.1564*</b>
Recall	0.3420	0.3631

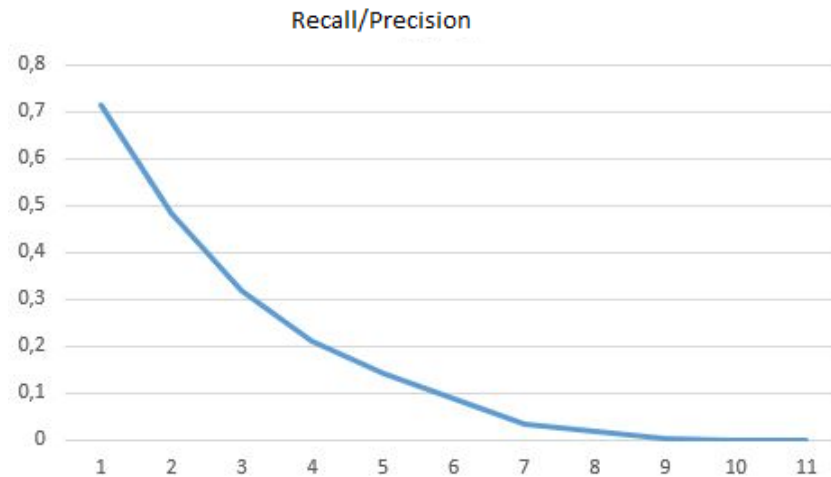
improvement or relevance rate, we can consider that the system with expansion is better than the base system).

The appearance of the recall/precision curve for the 100 queries from the Cystic Fibrosis Database is presented in figure 4, which corresponds to the precision at 11 recall points. To better check the performance of our search system, we used student's t-test [23] and we obtained a significant result with  $p < 0.004 < 0.0$

## 5. Conclusion and future works

We present in this paper an hybrid query expansion using linguistic resources and word embeddings. Indeed, we try to define semantic relationships between terms based on the combination between the linguistic definition (MeSH) and the statistical definition (Skip-Gram). We look for synonymy relations between terms in the initial query and concepts of MeSH thesaurus. Then, we apply an artificial neural network to learn a continuous vectors representation of words which will be able to capture semantic relations.

Experiments performed on Cystic Fibrosis Database show that the query expansion process



**Figure 4:** recall/precision curve

improves retrieval results. As future work, we will try to perform experiments on large databases.

## 6. Citations and Bibliographies

### References

- [1] Y. Wang, H. Huang, C. Feng, Query expansion with local conceptual word embeddings in microblog retrieval, *IEEE Transactions on Knowledge and Data Engineering* 33 (2019) 1737–1749.
- [2] N. Ksentini, M. Tmar, F. Gargouri, Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms, in: *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, IEEE, 2015, pp. 134–139.
- [3] N. Ksentini, M. Tmar, M. Boughanem, F. Gargouri, Miracl at clef 2015: User-centred health information retrieval task., in: *CLEF (Working Notes)*, 2015.
- [4] N. Ksentini, M. Tmar, F. Gargouri, The impact of term statistical relationships on rocchio's model parameters for pseudo relevance feedback, *International Journal of Computer Information Systems and Industrial Management Applications* 8 (2016) 135–44.
- [5] N. Ksentini, T. Mohamed, F. Gargouri, Towards automatic improvement of patient queries in health retrieval systems, *Applied Medical Informatics* 38 (2016) 73–80.
- [6] N. Ksentini, M. Tmar, F. Gargouri, Towards a contextual and semantic information retrieval system based on non-negative matrix factorization technique, in: *International Conference on Intelligent Systems Design and Applications*, Springer, 2017, pp. 892–902.
- [7] W. Shalaby, W. Zadrozny, Measuring semantic relatedness using mined semantic analysis, *arXiv preprint arXiv:1512.03465* (2015).

- [8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
- [9] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.
- [10] D. Roy, D. Paul, M. Mitra, U. Garain, Using word embeddings for automatic query expansion, arXiv preprint arXiv:1606.07608 (2016).
- [11] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa, A study on similarity and relatedness using distributional and wordnet-based approaches (2009).
- [12] N. Ksentini, M. Tmar, F. Gargouri, Detection of semantic relationships between terms with a new statistical method., in: WEBIST (2), 2014, pp. 340–343.
- [13] M. Sahami, T. D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, in: Proceedings of the 15th international conference on World Wide Web, 2006, pp. 377–386.
- [14] G. A. Miller, WordNet: An electronic lexical database, MIT press, 1998.
- [15] S. Liu, F. Liu, C. Yu, W. Meng, An effective approach to document retrieval via utilizing wordnet and recognizing phrases, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 266–272.
- [16] G. Feki, R. Fakhfakh, A. B. Ammar, C. B. Amar, Knowledge structures: Which one to use for the query disambiguation?, in: 2015 15th international conference on intelligent systems design and applications (ISDA), IEEE, 2015, pp. 499–504.
- [17] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, P. Bruza, Medical semantic similarity with a neural language model, in: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, 2014, pp. 1819–1822.
- [18] M. Mancini, J. Camacho-Collados, I. Iacobacci, R. Navigli, Embedding words and senses together via joint knowledge-enhanced training, arXiv preprint arXiv:1612.02703 (2016).
- [19] H. Yang, T. Gonçalves, Improving personalized consumer health search (2018).
- [20] M. D. Tapi Nzali, J. Azé, S. Bringay, C. Lavergne, C. Mollevi, T. Optiz, Reconciliation of patient/doctor vocabulary in a structured resource, Health Informatics Journal 25 (2019) 1219–1231.
- [21] J. A. Lossio-Ventura, C. Jonquet, M. Roche, M. Teisseire, Biotex: A system for biomedical terminology extraction, ranking, and validation, in: ISWC: International Semantic Web Conference, 1272, 2014, pp. 157–160.
- [22] R. Rehurek, P. Sojka, Software framework for topic modelling with large corpora, in: In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Citeseer, 2010.
- [23] S. Yue, P. Pilon, A comparison of the power of the t test, mann-kendall and bootstrap tests for trend detection/une comparaison de la puissance des tests t de student, de mann-kendall et du bootstrap pour la détection de tendance, Hydrological Sciences Journal 49 (2004) 21–37.
- [24] K. Sauvagnat, M. Boughanem, A la recherche de noeuds informatifs dans des corpus de documents xml., in: CORIA, 2005, pp. 119–134.