Evaluation of Vector Space Models for Medical Disorders Information Retrieval

Yaoyun Zhang, Trevor Cohen, Min Jiang, Buzhou Tang and Hua Xu*

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
{yaoyun.zhang, Trevor.Cohen, Min.Jiang, Buzhou Tang, hua.xu}@uth.tmc.edu

Abstract. Nowadays, consumers often search online to seek medical and health care information that they need. To improve this access, the ShARe/CLEF eHealth Evaluation Lab (SHEL) organized a shared task on information retrieval for Medical Disorders in 2013. This paper describes our participation in this task. In order to detect latent semantic relevance between queries and webpages about disorders, a semantic vector model based on distributional semantics is used as the information retrieval model. Specifically, variants of random indexing are employed to generate document and term representations. In addition, to reduce the lexical lap between different clinical expressions of the same concept, query expansion is also conducted using the UMLS. A baseline information retrieval method using the vector space model (VSM) and semantic vector models with different random indexing building procedures were developed and evaluated with or without query expansion in the shared task. The best performance was achieved by VSM, with MAP of 0.1480, P@10 of 0.3700 and nDCG@10 of 0.3363. Experimental results indicate that VSM and semantic vector model are complementary, and suggest combining these methods may further improve performance.

Keywords: medical disorder, information retrieval, vector space model, semantic vector model, query expansion.

1 Introduction

Nowadays consumers increasingly access electronically available medical and health care information. The rapid development and wide use of the Web has significantly altered the way people find medical information. Nearly 6% of Internet users on an average day search for medical information on the Web [1]. However, existing web search engines often fail to retrieve relevant results for medical queries [2].

Previous work has been attempting to solve this problem in multiple ways. Health information search behavior, information needs and contexts were analyzed based on query logs and social communities [3-4]. The expert system technology was integrated into the search engine to build a consumer-centric intelligent medical

search system [5]. Consumer queries reformulation and recommendation with professional terminologies were employed to reduce lexical gaps between queries and webpages [6-8]. Adaptive user model was built for performance evaluation from both the user and system perspective [9]. Context sensitive information retrieval considering negations in medical data is conducted to improve the retrieval precision [10]. Semantic resources like mesh and the UMLS were applied for query expansion [11-12]. What's more, unsupervised semantic relation measurements based on distributional semantics were employed and brought performance improvement for information retrieval on biomedical and clinical texts [13-14].

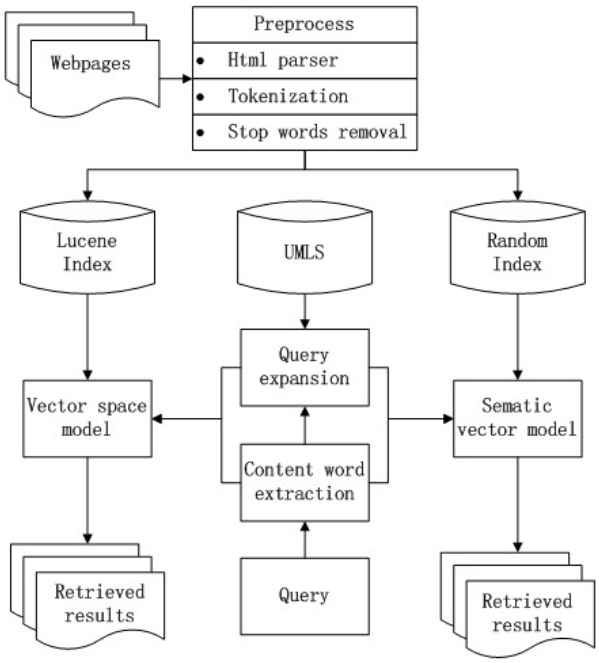

Figure 1. The Process of Information Retrieval for Medical Disorders

The ShARe/CLEF eHealth Evaluation Lab (SHEL) takes an initiative to organize a shared task on information retrieval for Medical Disorders in 2013 [15]. This paper describes our participation in this task. In order to detect latent semantic relations between queries about disorders and webpages (i.e. relations between queries and webpages that are relevant but do not contain the terms in the query), a semantic vector model based on distributional semantics [16] is used as the information retrieval model. Specifically, Reflective Random Indexing [17], an iterative variant of Random Indexing [18-19] that is better able to capture implicit relations, is employed for index building. In addition, to reduce the lexical lap between different clinical term expressions of the same concept, query expansion is also conducted using the UMLS. A baseline information retrieval method using the vector space model (VSM), semantic vector models with different random indexing building procedures, and the influence of query expansion are evaluated and compared using query datasets in the

shared task.

The latter sections are arranged as follows: Section 2 describes the information retrieval methods for medical disorders in detail. Section 3 presents the experiments and results. Section 4 discusses the experimental results and Section 5 is the conclusion.

2 Methods

2.1 Overview

Figure 1 shows the process of our proposed method for medical disorders information retrieval. The raw webpages are preprocessed to extract main content by the Html parser Tika[1]. Tokenization and stop words removal are then conducted, based on which indexes used for query retrieval are built. Each query in the training and test sets contains three parts, namely the title, description and narrative. We extract content words (i.e., nouns, verbs, and adjectives) from these three parts as the final query for information retrieval. Query expansion is also conducted using the UMLS [20]. Results are retrieved by two different information retrieval models, namely, the vector space model and the semantic vector model. Details of the two models are described as follows.

2.2 Information Retrieval Model

This paper employs two different information retrieval models: one is the vector space model, as one of the state-of-the-art benchmarks in information retrieval [21]. Another is the semantic vector model [16], which has been attracting research attention for effectively revealing latent semantic relations between terms and documents (unlike the VSM, which will only retrieve documents containing at least one of the query terms).

2.2.1 Vector Space Model

Represent the document $\boldsymbol{d_j}$ and the query $\boldsymbol{q}$ as vectors, $\overrightarrow{\boldsymbol{d_j}} = \langle \boldsymbol{w_{1,j}}, \ldots, \boldsymbol{w_{n,j}} \rangle$ and $\vec{\boldsymbol{q}} = \langle \boldsymbol{w_{1,q}}, \ldots, \boldsymbol{w_{n,q}} \rangle$ in an *n*-dimensional vector space [22]. Each dimension corresponds to a unigram, with tf.idf as the value. The cosine similarity between $\boldsymbol{d_j}$ and $\boldsymbol{q}$ is used for relevance ranking of documents. It is defined as:

$$\cos\left(\overrightarrow{\boldsymbol{d_j}}, \vec{\boldsymbol{q}}\right) = \frac{\overrightarrow{\boldsymbol{d_j}} \cdot \vec{\boldsymbol{q}}}{\|\boldsymbol{d_j}\| \cdot \|\boldsymbol{q}\|} \quad (1)$$

[1] http://tika.apache.org/

2.2.2 Semantic Vector Model

Methods of distributional semantics [4-5] assume that words and concepts with similar contextual distributions have similar or related meanings. In the semantic vector model we employ, words and concepts are represented by high-dimensional vectors in a mathematical space. Two vectors with a close distance in that space are considered to have high semantic similarity or relevance [23].

One key issue of semantic vector model is to reduce dimensions to improve processing performance, and in some cases measures of semantic relatedness. Random Indexing [18-19] offers an efficient and effective method for reducing the dimensionality of the semantic space. The process of random indexing is as follows:

1) Generate an *m*-dimensional index vector for each term. The term index vectors are generated using random projection [24], which projects the *n*-dimensional (*m* << *n*) term vector in to a lower dimensional subspace. Each index vector is a sparse vector with a small number of +1 and -1 values like $\langle 0, 0, 0, 1, 0 \ldots -1, 0, -1, 0, 0 \rangle$. Two arbitrary vectors are nearly orthogonal to each other, so that the distance between original vectors can be preserved [25]. Terms that co-occur directly will have similar vector representations
2) Sum index vectors of terms contained in a document to generate the document index vector.

Reflection:

3) Sum index vectors of documents containing a term to re-generate the term index vector.
4) Iterate between 2) and 3) to reflectively generate index vectors, so that to reveal higher-order relations between terms.

2.3 Query Expansion Using the UMLS

Previous work of query expansion [26] using the Unified Medical Language System (UMLS) [20] has achieved performance enhancement for information retrieval. We used content words extracted from queries as input and use UMLS API [27] to obtain possible Concept Unique Identifiers (CUI). Terms belonging to the top ranked CUI of the content word were used as expanded queries.

3 Experiments

3.1 Experimental Setup

The open source toolkit Lucene[2] is used for index building and information retrieval by VSM. Another open source toolkit, Semantic Vectors[3], is used for semantic vector index building and information retrieval [28-29]. Based on the manual annotations provided by the task organizer, performance on the training set is evaluated by MAP and P@10. Performance on the test set is evaluated by MAP, P@10 and nDCG@10.

[2] http://lucene.apache.org/core/

[3] https://code.google.com/p/semanticvectors/

MAP, as in equation 2, is abbreviated for mean average precision, which is the mean of the average precision scores for each query q in a query set Q. P@10 is the number of relevant results on the top ten search results. DCG, as in equation 3, uses a graded relevance scale rel_i to evaluate the usefulness of a document based on its position in the result list. DCG assumes highly relevant documents appearing lower in a search result list should be penalized. NDCG (Normalized Cumulative Gain) is the normalization of DCG value of the ideal ranking at rank n.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AvePrecision}(q)}{Q} \quad (2)$$

$$\text{DCG}_Q = rel_1 + \sum_{i=2}^{Q} \frac{rel_i}{log_2 i} \quad (3)$$

The following methods are compared in our experiment:

- **VSM** (UTHealth_CCB.1.3.noadd): Results are retrieved from Lucene using VSM, with content words from title and description as queries.
- **SemVec** (UTHealth_CCB.5.3.noadd): Results are retrieved from 4000-dimensional semantic vector based index, with content words from title, description as queries. The index is built without reflection.
- **VSM&UMLS** (UTHealth_CCB.6.3.noadd): Results are retrieved from Lucene using VSM, with content words from title, description, narrative, and expanded terms from the UMLS as queries.
- **SemVec&UMLS** (UTHealth_CCB.7.3.noadd): Results are retrieved from 2000-dimensional semantic vector based index, with content words from title, description, narrative, and terms from the UMLS as queries. The index is built with one turn of reflection.

3.2 Results

Table 1 shows the performance of the employed methods on the training queries for the ShARe/CLEF eHealth 2013 shared task 3. VSM and SemVec obtained comparable results. Query expansion using UMLS achieved enhancements both for VSM and SemVec, especially in P@10, with an improvement of 100% for VSM, and an improvement of 75% for SemVec. The MAP also increased about 50% for VSM.
Table 2 displays the performance of our methods on the test queries. The best performance was achieved by the baseline method, i.e., VSM, with MAP of 0.1480, P@10 of 0.3700 and nDCG@10 of 0.3363. Nevertheless, performance of the other three methods dropped severely on all the three evaluation criteria compared with VSM.

Table 1. Performance on Training Queries

Methods	MAP	P@10
VSM	0.0706	0.0800
SemVec	0.0672	0.0800
VSM&UMLS	0.1061	0.1600
SemVec&UMLS	0.0764	0.1400

Table 2. Performance on Test Queries

Methods	MAP	P@10	nDCG@10
VSM	0.1480	0.3700	0.3340
SemVec	0.0862	0.2440	0.2338
VSM&UMLS	0.1104	0.2520	0.2270
SemVec&UMLS	0.0539	0.1420	0.1337

4 Discussion

As demonstrated in Table 1 and 2, performance on the training and test queries differed in several aspects. All the results enhanced significantly on test queries compared to the training set, except for SemVec with query expansion using the UMLS. Besides, the performance of SemVec was lower than that of VSM on the test set, instead of a comparable performance between the two on the training set. Furthermore, in contrast to a performance increase using query expansion on the training set, the test set illustrated a performance decrease. One possible reason of these differences could be the different pooling sets used for evaluation between the two sets. The pooling set for training was built from results from two information retrieval models, VSM and Okapi BM25 [30]. In contrast, the pooling set for test was built from the results submitted from all participants. Another reason could be the different information needs contained in the two sets. In addition to questions asking for disorder definition and treatment in the training set, relational questions asking for connections between other entities and disorders account for a large proportion in the test set, such as "<desc> can chest pain hinder the transplantation of liver? </desc>" and ""<desc> what is the connection between acidosis and metastasic adeno carcinoma</desc>". The employed methods in our experiment may not be suitable for retrieving relevant information for such kind of information needs.

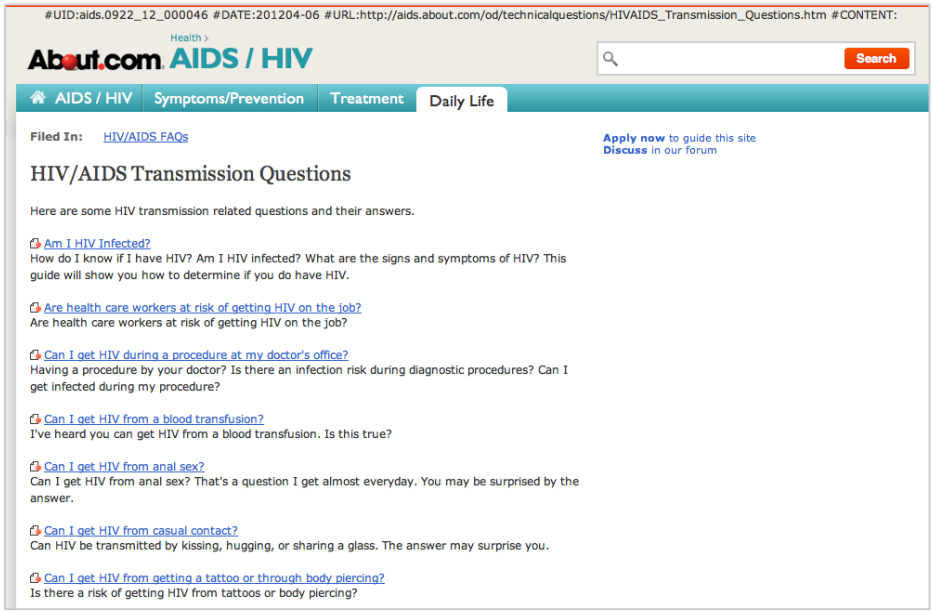

Figure 2. A Navigational Webpage of HIV retrieved by SemVec.

As illustrated by the P@10 plots provided by the task organizer, both VSM and SemVec contributed best P@10 for several different queries. Looking into the results of VSM and SemVec, it was found that they retrieved different relevant webpages from the index. Take test query 12 "<desc>is clots in jugular in connection with HIV</desc>" as an example, VSM mainly retrieved webpages informative of HIV. Nevertheless, SemVec obtained more navigational webpages containing links to diverse aspects of HIV, with each link leading to an informative webpage focusing a specific aspect of HIV. Figure 2 shows a navigational webpage of HIV retrieved by SemVec.

Furthermore, relevant webpages with different topics of the query were retrieved. As an example, for test query 10 "<desc>is there a connection between multiple sclerosis and dysplasia in oesophagus</desc>", VSM obtained three relevant webpages about "dysplasia in oesophagus" and one about "multiple sclerosis", while SemVec found eight relevant webpages about "multiple sclerosis". Those observations demonstrate that these two methods are complementary to each other and may be combined to produce more diverse and relevant results.

5 Conclusions

This paper describes our participation in the task 3 in ShARe/CLEF eHealth 2013 challenge. Different information retrieval models, namely vector space model and semantic vector models based on the distributional semantics theory were employed. The experimental results demonstrate that both models are complementary to each

other. The next step in our future work would be exploiting the combination of semantic vector models with other information retrieval models for further performance improvement.

References

1. C. Sherman: Curing medical information disorder. (2005)
2. G. Luo, C. Tang, H. Yang, and X. Wei: MedSearch: a specialized search engine for medical information retrieval. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 143–152. (2008)
3. Y. Zhang, P. Wang, A. Heaton, and H. Winkler: Health information searching behavior in MedlinePlus and the impact of tasks. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, USA, pp. 641–650. New York, USA (2012)
4. Y. Zhang: Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 210–219. New York, USA (2010)
5. G. Luo: Lessons learned from building the iMED intelligent medical search engine. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC 2009, pp. 5138–5142. (2009)
6. S. P. Crain, S.-H. Yang, H. Zha, and Y. Jiao: Dialect topic modeling for improved consumer medical search. In: AMIA Annu Symp Proc, vol. 2010, pp. 132–136. (2010)
7. R. M. Plovnick and Q. T. Zeng: Reformulation of consumer health queries with professional terminology: a pilot study. Journal of medical Internet research, vol. 6, no. 3. (2004)
8. Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo, and E. Dibble: Assisting consumer health information retrieval with query recommendations. Journal of the American Medical Informatics Association, vol. 13, no. 1, pp. 80–90. (2006)
9. E. Santos, H. Nguyen, Q. Zhao, and E. Pukinskis: Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application. In: User Modeling 2003, P. Brusilovsky, A. Corbett, and F. de Rosis, Eds. Springer Berlin Heidelberg, pp. 292–296. (2003)
10. M. Fieschi: Context-sensitive medical information retrieval. In: Medinfo 2004: Proceedings of the 11th World Conference on Medical Informatics, vol. 107, pp. 282. San Francisco (2004)
11. A. L. Houston, H. Chen, B. R. Schatz, S. M. Hubbard, R. R. Sewell, and T. D. Ng: Exploring the use of concept spaces to improve medical information retrieval. Decision Support Systems, vol. 30, no. 2, pp. 171–186. (2000)
12. M. C. Díaz-Galiano, M. Á. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ureña-López: Integrating MeSH Ontology to Improve Medical Information Retrieval. in Advances in Multilingual and Multimodal Information Retrieval, C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard,

A. Peñas, V. Petras, and D. Santos, pp. 601–606. Springer Berlin Heidelberg (2008)
13. W. R. Hersh, Information retrieval: a health and biomedical perspective. Springer (2009)
14. J. Lin and W. J. Wilbur: PubMed related articles: a probabilistic topic-based model for content similarity. BMC bioinformatics, vol. 8, no. 1, pp. 423. (2007)
15. Hanna Suominen, Sanna Salanter"a, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Danielle Mowery, Lorraine Goeuriot, Liadh Kelly, David Martinez, and Guido Zuccon: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. (2013)
16. T. Cohen and D. Widdows: Empirical distributional semantics: Methods and biomedical applications. Journal of Biomedical Informatics, vol. 42, no. 2, pp. 390. (2009)
17. T. Cohen, R. Schvaneveldt, and D. Widdows: Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics, vol. 43, no. 2, pp. 240–256. (2010)
18. P. Kanerva, J. Kristofersson, and A. Holst: Random indexing of text samples for latent semantic analysis. In: Proceedings of the 22nd annual conference of the cognitive science society, vol. 1036. (2000)
19. M. Sahlgren: Vector-based semantic analysis: Representing word meanings based on random labels. In: ESSLI Workshop on Semantic Knowledge Acquistion and Categorization. (2001)
20. “Unified Medical Language System (UMLS) - Home.” [Online]. Available: http://www.nlm.nih.gov/research/umls/. [Accessed: 22-May-2013].
21. C. D. Manning, P. Raghavan, and H. Schütze: Introduction to information retrieval, vol. 1. Cambridge University Press Cambridge (2008)
22. G. Salton, A. Wong, and C.-S. Yang: A vector space model for automatic indexing. Communications of the ACM, vol. 18, no. 11, pp. 613–620. (1975)
23. M. Sahlgren: An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE, vol. 5. (2005)
24. E. Bingham and H. Mannila: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245–250. (2001)
25. P. Frankl and H. Maehara: The Johnson-Lindenstrauss lemma and the sphericity of some graphs. Journal of Combinatorial Theory, Series B, vol. 44, no. 3, pp. 355–362. (1988)
26. A. R. Aronson and T. C. Rindflesch: Query expansion using the UMLS Metathesaurus. In: Proceedings of the AMIA Annual Fall Symposium, pp. 485. (1997)
27. B. T. McInnes, T. Pedersen, and S. V. Pakhomov: UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In: AMIA Annual Symposium Proceedings, vol. 2009, pp. 431. (2009)

28. D. Widdows and K. Ferraro: Semantic vectors: a scalable open source package and online technology management application. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp. 1183–1190. (2008)
29. D. Widdows and T. Cohen: The semantic vectors package: New algorithms and public tools for distributional semantics. In: IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 9–15. (2010)
30. S. E. Robertson, S. Walker, and M. Beaulieu: Experimentation as a way of life: Okapi at TREC. Information Processing & Management, vol. 36, no. 1, pp. 95–108. (2000)