International Journal of Information Science and Management
Vol. 17, No. 1, 2019, 17-31

Investigating Text Power in Predicting Semantic Similarity

Zahra Yousefi
Ph.D. candidate in Knowledge & Information Sciences, Department of Knowledge & Information Sciences, Faculty of Education & Psychology, Shiraz University,
z.yusefi@gmail.com

Hajar Sotudeh
Associate Prof. Department of Knowledge & Information Sciences, Faculty of Education & Psychology, Shiraz University,
Corresponding Author: sotudeh@shirazu.ac.ir

Mahdieh Mirzabeigi
Assistant Prof. Department of Knowledge & Information Sciences, Faculty of Education & Psychology, Shiraz University,
mmirzabeigi@gmail.com

Seyed Mostafa Fakhrahmad
Assistant Prof. Department of Computer Science & Engineering School of Electrical and Computer Engineering, Shiraz University,
fakhrahmad@shirazu.ac.ir

Alireza Nikseresht
Assistant Prof. Department of Knowledge & Information Sciences, Faculty of Education & Psychology, Shiraz University,
nikseresht@gmail.com

Mehdi Mohammadi
Associate Prof. Department of Educational Management and Planning, Faculty of Education & Psychology, Shiraz University,
m48r52@gmail.com

Abstract

This article presents an empirical evaluation to investigate the distributional semantic power of abstract, body and full-text, as different text levels, in predicting the semantic similarity using a collection of open access articles from PubMed. The semantic similarity is measured based on two criteria namely, linear MeSH terms intersection and hierarchical MeSH terms distance. As such, a random sample of 200 queries and 20000 documents are selected from a test collection built on CITREC open source code. Sim Pack Java Library is used to calculate the textual and semantic similarities. The nDCG value corresponding to two of the semantic similarity criteria is calculated at three precision points. Finally, the nDCG values are compared by using the Friedman test to determine the power of each text level in predicting the semantic similarity. The results showed the effectiveness of the text in representing the semantic similarity in such a way that texts with maximum textual similarity are also shown to be 77% and 67% semantically similar in terms of linear and hierarchical criteria, respectively. Furthermore, the text length is found to be more effective in representing the hierarchical semantic compared to the linear one. Based on the findings, it is concluded that when the subjects are homogenous in the tree of knowledge, abstracts provide effective semantic capabilities, while in heterogeneous milieus, full-texts processing or knowledge bases is needed to acquire IR effectiveness.

Keywords: Distributional Semantics, Semantic Similarity, Textual Similarity, Effectiveness, Information Retrieval, MeSH.

Introduction

Efficiency and effectiveness are the two major criteria in evaluating information retrieval (IR) system performance. The effectiveness of information retrieval is highly dependent on the accurate and complete representation of document content. Natural language processing (NLP) is among the early approaches in representing documents in automated indexing systems. In fact, it calculates distributional semantics based on the assumption that linguistic items with similar distributions have similar meanings (Gritta 2015; Harispe, Ranwez, Janaqi & Montmain 2015). NLP-based IR brings in speed and ease of indexing while removing human errors and costs (Moskovitch, Martins, Behiri, Weiss & Shahar 2007), leading to a relatively high efficiency of IR systems. Although the huge number of indexed terms endangers the efficiency of the system (Scheffler, Schumacher & March 1974). It has also been revealed to be effective in identifying relevant documents (Lu, Kim & Wilbur, 2009), meeting users' information needs (Swanson 1960; Salton 1970) even in competing with controlled vocabularies (Hersh and Hickam 1992; Hersh, Price & Donohoe 2000). However, NLP-based systems which are founded on plain lexicographic term matching, encounter serious challenges when dealing with semantics issues e.g. synonymy, polysemy, and semantic relations (Gabrilovich and Markovitch 2009). The consequence of using inaccurate and inadequate documents representations (Purcell et al. 1997) may be a high recall and a low precision (Moskovitch et al. 2007), and hence a low effectiveness in text-based IR systems. In spite of considerable improvement, advanced NLP techniques such as word embedding (Lavelli, Sebastiani & Zanoli 2004; Mikolov, Chen, Corrado & Dean 2013; Liu, Lang, Gu & Zeeshan 2017), were found to need knowledge-based techniques, such as sense graph embedding, to overcome the semantic issues (Wang, Mao, Wang & Guo 2017; Camacho-Collados & Pilehvar 2018). Therefore, NLP techniques are sometimes believed to be far from the desired situation in spite of four decades of research efforts in designing and testing the techniques (De Bellis 2009).

One of the widely tested solutions is the use of knowledge bases in indexing and representing documents or in the searching phase as a substitute or a supplement to the NLP techniques. Controlled vocabularies assigned by human or machine indexers are believed to be able to reduce the intrinsic ambiguity of natural language (Trieschnigg et al. 2009). Knowledge bases are advantageous in gathering together semantically-similar but lexically-different terms, overcoming linguistic discrepancies, representing concepts underlying words, controlling for word variations (Coyle 2008; Savoy 2005) and clarifying complicated relations of terms (Liu 2010). Despite the advantages, the costs, time and resources required to develop, implement and update the knowledge tools are among the primary factors affecting the efficiency of the systems (Papanikolaou, Tsoumakas, Laliotis, Markantonatos & Vlahavas 2017). This particularly holds for human indexing systems. For instance, it takes 2-3 months for a document included in Medline to be indexed manually, costing 10 dollars (Mao and Lu 2017).

Aside from the efficiency issues, the systems have not always been proven to ideally perform in retrieving relevant documents. Widespread studies aiming at comparing the

performance of the knowledge-base and NLP techniques have not been consistent in their results. For example, when contrasting NLP and MeSH searching modes, Saka, Gulkesen, Gulden & Koçgil (2005) found out that there are no significant differences between the results of the two search modes. Although some confirm superiority of the controlled vocabularies over text-based IR systems (Tenopir 1985; Svenonius 1986; Srinivasan 1996; Arellano, 2000; Chang, Heskett, & Davidson 2006; Moskovitch et al. 2007), some found that they are ineffective in improving information retrieval (Salton 1972; Savoy 2005) and even reduce the retrieval effectiveness (Hersh and Hickam 1992; 1993; Hersh et al. 2000). In searching for some solutions to increase IR effectiveness, some previous studies prescribe semantic indexing using a combination of the two methods (Peters and Kurth 1991; Hersh, Buckley, Leone & Hickam 1994; Shaw 1994; Muddamalle, 1998; Savoy 2005; Zhu, Zeng, & Mamitsuka 2009). This inconsistency of the results gives rise to the question whether the semantic tools are considerably more effective than the plain-text itself in retrieving documents. In other words, how powerful are plain-texts in predicting their semantic similarity? Furthermore, although several investigations compared the textual and semantic approaches in terms of effectively retrieving relevant documents, no studies that deal with a comparison of main parts of texts in representing semantic similarity using NLP techniques were found.

The present contribution endeavors to investigate texts power in predicting semantic similarity. To do so, it tries to verify the texts ranked by their textual similarity values in terms of their semantic similarity to a set of queries. In addition, in order to find the most powerful part in representing semantic similarity, it also tries to compare main parts of texts, including abstracts, bodies and full-texts, which have been reported to be effective in IR (Zeng, He, Chen, Ma, & Ma 2004; Rezapour, Fakhrahmad, & Sadreddini 2011), with different efficiency outcomes (Scheffler et al. 1974). As abstracts and bodies differ in the quantity of their textual component and thus in processing and memory loads they impose on IR systems, identifying the most powerful part of texts is useful not only in improving the effectiveness of IR results but also in reducing the computational costs of text processing and memory and thereby improving systems efficiency.

The plain-text similarity is measured on their lexical level. In order to measure semantic similarity, we use MeSH terms as a semantic representation of documents. Semantic similarity refers to the closeness, proximity, or nearness of two words in their meanings (Joubarne and Inkpen 2011; Inkpen and Désilets 2005). At the simplest level, two terms are believed to be semantically related if they are lexically similar. However, there exists lexically similar but semantically different words (like homographs) and vice versa. As a result, semantic similarity is measured based on the similarity or distance of their paths and hierarchies within the tree of knowledge operationally represented by hierarchical structures such as ontologies, thesauri and taxonomies e.g. Wordnet, MeSH (Névéol, Zeng, & Bodenreider 2006; Lee, Shah, Sundlass & Musen 2008; Leopold et al. 2012; Nazim Uddin, Duong, Nguyen, Qi & Jo 2013). In order to have a more realistic estimation of the plain-text similarity in predicting semantic similarity, the present study measures the semantic similarity

at two levels: the lower level is based on the intersection of the MeSH terms between two textually similar documents. This reduces the semantic similarity to a rough lexical similarity between MeSH terms. At a higher level, semantic similarity is also calculated based on their closeness within the hierarchy of the MeSH tree.

Research Questions

The present study aimed to answer the following questions:

1- Are the main parts of documents effective in representing semantic similarity?

2- Is there any significant difference between the effectiveness of the main parts of documents in representing semantic similarity at the precision points p@10, p@20 and p@50?

Methodology

In order to achieve the aforementioned aims, a test collection is built using the CITREC open source project code publicly released in 2015. Sim Pack Java Library is used to calculate a variety of textual and semantic similarity measures between queries and documents (Gipp, Meuschke & Lipinski 2015).

Test Collection

The test collection built consisted of three components including documents, queries, and relevancy measure.

Population: 13957 documents indexed in PubMed in 2010-2017 are downloaded. The reason for choosing PubMed was the open accessibility of its papers as well as the considerable advancement in controlled vocabularies of Life and Biomedical Sciences (Liu and Wacholder, 2017).

Queries: By Using PubMed IDs of the papers, 200 documents are randomly selected and served as queries. By measuring the similarity of each of the queries to the collection, 5115 unique documents are found to be similar to at least one of the queries. Given the matching of some of the documents to more than one query, the documents number reached 20000 at last.

Documents: Based on a "criterion technique", top 100 documents were selected from the documents similar to each query, after calculating the similarity of each query to the 13957 documents.

Relevance: relevance criterion is an essential component of test collections (Manning et al. 2008). Since the present study is focused on determining text power in predicting semantic similarity, MeSH similarity is chosen as the relevance criterion and used as the Gold Standard (Gipp et al., 2015). As Harispe et al. (2015) put it, in designing semantic measures, especially those based on domain-specific ontologies, evaluation is based on the semantic interactions between semantic entities according to the analysis of semantic proxies (texts, ontologies) which are not necessarily to mimic human relevance but to be coherent with the knowledge expressed in the considered semantic proxy.

Semantic similarity: it is measured at two levels. At a simpler definition, the semantic

similarity is just calculated based on Jaccard Coefficient which divides MeSH terms' intersection by their union. It is called "semantic similarity at the lexical level". At a deeper level, the semantic similarity is measured using the descriptors distances in the thesaurus structure. To do so, a combination of Lin's (1998) algorithm and the Information Content was proposed by Resnik (1995) as described in (Gipp et al., 2015) is used. We call it "hierarchical semantic similarity".

Textual similarity: Lucene more-Like-This function was used to measure the texts similarity at three levels based on their importance, length and thus a number of elements requiring text processing. These include abstracts, bodies, and full-texts. It should be mentioned that titles were also verified to match similar documents. However, the number of similar documents matched by their titles were found to be very limited. Previous research found that title keywords are weak in finding similar (Sotudeh & Houshyar, 2018) and relevant documents (Byrne, 1975; Gross & Taylor, 2005). This may have roots in the fact that title context is too short to contain enough elements conveying document contents.

Data Analysis

After measuring the textual and semantic similarities, the documents are ranked based on their descending values of textual similarities. The top 100 documents in terms of textual similarity are chosen to be analyzed. nDCG (normalized Discounted Cumulative Gain), which is a useful tool for graded relevance (Kekäläinen, 2005), is used to evaluate ranking quality based on the position of the documents ranked. It is yielded by dividing the DCG (Discounted Cumulative Gain) to the ideal DCG (IDCG).

DCG is calculated by:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

where "p" represents the precision point, "rel_i" is the relevance score of the i^{th} ranked documents and IDCG is measured by:

$$IDCG_p = \sum_{i=1}^{|\mathrm{REL}|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

where |*Rel*| represents the list of relevant documents (ordered by their semantic similarity scores) up to "p".

nDCG is calculated for the documents in terms of their semantic similarities at three precision points including p@10, 20, and 50. In this way, nDCG values, ranging from 0 to 1, reflect the text power in predicting semantic similarity of its MeSH terms, i.e. the semantic effectiveness of textual similarity. In order to have an insight of the text power, one may take into account the values within a continuum from very weak (0-02 meaning 0-20% of effectiveness), weak (0.2-0.4 meaning 20-40% of effectiveness), medium (0.4-0.6 meaning 40-60% of effectiveness), strong (0.6-0.8 meaning 60-80% of effectiveness) to very strong (0.8-1 meaning 80-100% of effectiveness).

Given the non-normality of the data, even after the log-normal transformation, the nDCG

values are compared using the Friedman test, which is a nonparametric statistical test alternative to the one-way ANOVA with repeated measures.

Analyses were conducted at three levels including abstracts, bodies and full-texts. The term 'body' refers to the main body of information embedded in a paper, from the introduction to the conclusion, excluding abstracts, tables, figures and references. "Full-texts" encompass "bodies" and "abstracts".

Findings

Text power in predicting semantic similarities

Figure one illustrates the text powers at the three levels of abstracts, bodies, and full-texts at P@10, 20 and 50 in terms of the mean values of nDCG. The horizontal axis of the graph is organized by descending order of efficiency level, and thus ascending order of processing load and browsing time. The vertical axis is devoted to the semantic effectiveness of textual similarity. As observed, the nDCG mean values vary from 0.55 to 0.77. The least mean value for the "semantic similarity at lexical level" is 0.67. It means that textual similarity can averagely predict 67% of the semantic similarity at the lexical level. In other words, if one wishes to reach an average semantic effectiveness of about 70%, it is sufficient to review top 10 textually-similar documents (p@10) processed for the least textual elements, i.e. abstracts.

It is also shown in the figure, that textual similarity is considerably powerful in predicting the hierarchical semantic similarity, although at a relatively lower level. As seen, the abstracts are capable of predicting about 55% of this kind of semantic similarity. Users should review 50 top-ranked textually-similar and fully-processed documents to reach the highest mean semantic effectiveness that can be achieved by textual similarity, i.e. 0.67%.

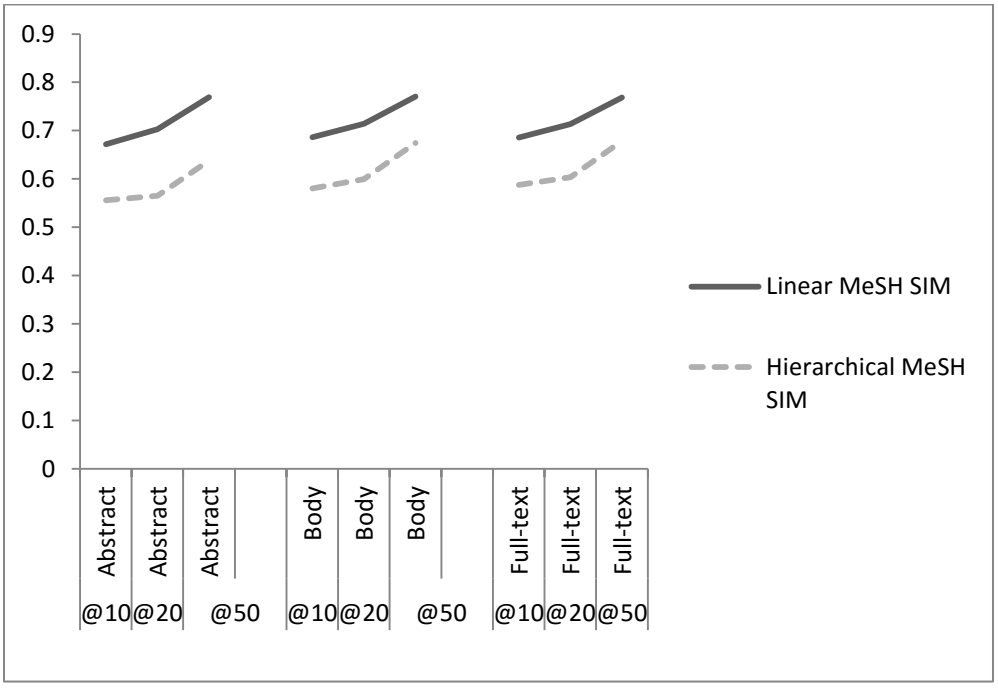

Figure 1: The nDCG mean values for the semantic similarities predicted by textual similarities

Comparison of text powers in achieving semantic effectiveness

Using the Friedman test, the main parts of documents are compared in terms of their nDCG values, which reflect their powers in achieving semantic effectiveness at p@k=10,20,50.

Table 1
Comparison of main parts of texts powers in predicting semantic similarities

Semantic SIM Measure	Textual Similarity K	Text	Chi-Square	df	Asymp. Sig	nDCG Mean Rank	nDCG Mean
Linear MeSH SIM	@10	Body	4.3884	2	0.1114	2.050	0.686
		Fulltext				2.048	0.685
		Abstract				1.903	0.672
	@20	Body	8.365	2	0.0152	2.083	0.714
		Fulltext				2.063	0.714
		Abstract				1.855	0.703
	@50	Body	0.9837	2	0.6114	2.053	0.770
		Abstract				1.970	0.769
		Fulltext				1.978	0.769
Hierarchical MeSH SIM	@10	Fulltext	9.592	2	0.008	2.159	0.588
		Body				1.975	0.581
		Abstract				1.866	0.556
	@20	Fulltext	22.535	2	0.000	2.227	0.603
		Body				2.015	0.599
		Abstract				1.758	0.565
	@50	Fulltext	53.750	2	0.000	2.323	0.678
		Body				2.075	0.675
		Abstract				1.601	0.638

The results are illustrated in Table 1. As seen, the results reveal that the main parts of texts do not significantly differ in their powers of achieving lexical semantic effectiveness for the p@10 and 50. However, it is significant for p@20, where the abstract is the weakest among the textual parts.

Moreover, all analyses at all precision levels showed significant differences among the textual parts for the hierarchical semantic effectiveness in favor of longer parts, i.e. bodies and full-texts (Table 1). It should be mentioned that further analyses of the variables revealed that the significant difference is between abstracts on the one hand and bodies and full-texts on the other. The longer parts, however, equal in their semantic effectiveness.

As a result, the main parts of documents, being different in their potential to contain content and subject clues, vary in representing the meaning underlying the texts. Based on the finding, in achieving lexical semantic effectiveness, the shortest part of documents, i.e. abstracts, are more or less as powerful as the longer parts including bodies and full-texts. However, for the hierarchical semantic similarity, the length of texts is highly determining in

the achievement, so that the longer the text, the more powerful it can be in predicting this kind of semantic similarity. Consequently, processing of full-texts (including abstracts and bodies) of documents is seen to be inevitable to get to the highest power of semantic prediction.

Discussion & Conclusion

NLP-based information retrieval is built on the assumption that texts have the elements necessary to identify the concepts and subjects they carry and hence "all necessary information needed to retrieve them" (Hjørland, 2008). It measures distributional semantics based on the notion that similar meanings underlie similar distributions of words. Although the method has been found to successfully perform in retrieving relevant documents and satisfying users' information needs (Swanson, 1960; Salton, 1970 & Lu, et al. 2009), it was criticized for its ignorance of related and synonym concepts which are not lexically similar (Petrakis, Varelas, Hliaoutakis & Raftopoulou, 2006). However, as far as our literature review goes, no studies have been found to tackle the problem by comparing the distributional semantics versus controlled hierarchical semantics. There are neither found investigations contrasting the main parts of documents in terms of their powers in predicting semantic similarity using NLP techniques.

Contrary to the previous literature which questions the effectiveness of distributional similarity in reflecting semantics (Mihalcea, Corley & Strapparava, 2006), the results of the present communication reveal that textual similarity can predict semantic similarity at two levels of lexical and hierarchical semantic similarity between MeSH terms. Although texts are averagely more powerful in predicting the former (up to 0.77), they are also shown to have considerable effectiveness in predicting the latter (up to 0.67). This means that semantically similar texts have linguistic elements in common and are more or less equally powerful for predicting semantic similarity either in the hierarchical structure of knowledge or at the reduced level of lexical similarity. The prediction of hierarchical semantic similarity also implies the effectiveness of the distributional similarity in word sense disambiguation e.g. for lexically similar but semantically different words (e.g. Homographs) that have already been tested and approved in several studies (Han, Giles, Zha, Li & Tsioutsiouliklis, 2004; Tang, Fong, Wang & Zhang, 2012; Sotudeh and Houshyar, 2018).

As seen in figure 1, browsing a higher number of documents by users and processing higher parts of documents by systems, does not necessarily yield higher lexical semantic effectiveness. The more or less equal effectiveness of abstracts, compared to their full-texts, is due to the fact that abstracts are not only consistent with their mother articles, but also more subject-intensive, as they cover central rather than peripheral features of the articles (Strang, 1997). It may also have roots in the fact that keywords used by authors have a lot in common with controlled terms used in knowledge tools (Ansari, 2005; Gil-Leiva & Alonso-Arroyo, 2007). This may also happen - either consciously or subconsciously - when writing abstracts. The finding is consistent with Byrne's (1975), confirming the robustness of abstracts alone in retrieving relevant papers. It is also in line with Shin, Han & Gelbukh (2004) who found that taking Medline abstracts into consideration, along with MeSH terms, significantly improves

the retrieval results. However, it is not consistent with theirs, in that they revealed that the abstracts alone do not provide enough information for search. Scheffler et al. (1974) also discovered that excluding document bodies from indexing terms would considerably increase efficiency, causing no significant decrease in effectiveness compared to other textual elements (including titles, abstracts, table of contents and figures) that achieve the optimum retrieval. The fact is highlighted by the present study, too. As seen in Figure 1, the distance between the lowest and highest levels of semantic effectiveness is about 10 and 12% for the lexical and hierarchical semantic measurement, respectively. The improvement is gained by minimizing the efficiency of the system, i.e. highest processing loads possible and a relatively longer browsing time for users. This means that maximizing the processing load would yield just a small improvement in the semantic effectiveness of textual similarity (10-12%). The improvement does not seem such considerable to justify the costs imposed on NLP and indexing systems.

The effectiveness of abstracts in reflecting the similarity of MeSH terms implies that databases founded just on abstract processing, not supported by knowledge bases, realize a considerable level of lexical and an acceptable level of hierarchical semantic similarity. In more heterogeneous milieus, with more hierarchical subject relations, texts can be effective provided that longer parts of documents are processed and users are encouraged to browse more top-ranked documents returned.

One should bear in mind that the effectiveness of textual similarity in predicting hierarchical semantic similarity is at the utmost 0.67 for the top 50 ranked documents. Consequently, it does not eradicate the need for knowledge bases in order to achieve a 100% performance. This is in line with previous literature which emphasized the need for continuing the use of controlled vocabularies along with texts to maximize the effectiveness of searching (Gross and Taylor, 2005; Garrett, 2007; McCutcheon, 2009; Strader, 2011; Gross, Taylor & Joudrey, 2015).

Given the challenges of developing, maintaining, and updating knowledge bases, it is desired to have more ready-made and efficient measures helping to improve the semantic effectiveness. In our ongoing studies, we are, therefore, trying to use textual elements enriched with bibliometric and altmetric evidence to enhance the effectiveness of retrieving medical papers.

Acknowledgment

We used CITREC source code, courtesy of the University of Konstanz. We would like to cordially thank all, especially Norman Meuschke for all his valuable advice and help.

References

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review*, 54(7), 410-414. https://www.emeraldinsight.com/doi/abs/10.1108/00242530510611901.

Arellano, F. F. M. (2000). Subject searching in online catalogs including Spanish and English

material. *Cataloging & classification quarterly*, 28(2), 45-56. Retrieved from: https://www.tandfonline .com/doi/pdf/10.1300/J104v28n02_04

Byrne, J. R. (1975). Relative effectiveness of titles, abstracts, and subject headings for machine retrieval from the COMPENDEX services. *Journal of the Association for Information Science and Technology*, 26(4), 223-229. Retrieved from: https://onlinelibrary.wiley.com/ doi/pdf/10.1002/asi.4630260405

Camacho-Collados, J., & Pilehvar, T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *arXiv:1805.04032*. Retrieved from: https://arxiv.org/pdf/1805.04032.pdf

Chang, A. A., Heskett, K. M., & Davidson, T. M. (2006). Searching the literature using medical subject headings versus text word with PubMed. *The Laryngoscope*, 116(2), 336-340. Retrieved from: https://onlinelibrary.wiley.com/doi/epdf/10.1097/01.mlg.0000195371.72887.a2

Coyle, K. (2008). Machine Indexing. *The Journal of Academic Librarianship*, 34(6), 530-531. Retrieved from https://kcoyle.net/jal_34_6.html

De Bellis, N. (2009). *Bibliometrics and citation analysis from the science citation index to cybermetrics*. Lanham, Md: Scarecrow Press.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34, 443-498. Retrieved from https://www.aaai.org/Papers/JAIR/Vol34/JAIR-3413.pdf

Garrett, J. (2007). Subject headings in full-text environments: the ECCO experiment. *College & Research Libraries*, 68(1), 69-81. Retrieved from https://kopernio.com/viewer?doi=10.5860/crl.68.1.69&route=6

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American society for information science and technology*, 58(8), 1175-1187. Retrieved from https://onlinelibrary.wiley.com/doi/epdf /10.1002/asi.20595

Gipp, B., Meuschke, N., & Lipinski, M. (2015). CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central. *Proceedings of the iConference*. Newport Beach: iSchools.

Gritta, M. (2015). Distributional Semantics and Authorship Differences (Doctoral dissertation, University of Cambridge).

Gross, T., & Taylor, A. G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, 66(3), 212-230. Retrieved from http://marklindner.info/presentations/Gross_Taylor_Pres/GrossTaylor .pdf

Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a lot to lose: the role of controlled vocabulary in keyword searching. *Cataloging & classification quarterly*, 53(1), 1-39. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01639374.2014.917447?needAccess=true

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning

approaches for name disambiguation in author citations. *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries.* New York, NY: ACM.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1), 1-254. Retrieved from https://www.researchgate.net/profile/ Sebastien_Harispe/ publication/277328095_Semantic_Similarity_from_Natural_Language_and_Ontology_Analysis/links/58f5e11a458515ff23b6307d/Semantic-Similarity-from-Natural-Language-and-Ontology-Analysis.pdf

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York, NY: ACM.

Hersh, W., Price, S., & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proceedings of the AMIA Symposium.* San Francisco, CA: AMIA.

Hersh, W. R., & Hickam, D. H. (1992). A comparison of retrieval effectiveness for three methods of indexing medical literature. *The American Journal of the Medical Sciences*, 303(5), 292-300. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0002962915357013?via%3Dihub

Hersh, W. R., & Hickam, D. H. (1993). A comparison of two methods for indexing and retrieval from a full-text medical database. *Medical Decision Making*, 13(3), 220-226. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0272989X9301300308

Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge organization*, 35(2,3), 86-101. Retrieved from https://www.researchgate.net/publication/277803483_What_is_Knowledge_Organization_KO

Inkpen, D., & Désilets, A. (2005). Semantic similarity for detecting recognition errors in automatic speech transcripts. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. New York, NY: ACM.

Joubarne, C., & Inkpen, D. (2011). Comparison of semantic similarity for different languages using the Google N-gram corpus and second-order co-occurrence measures. *Proceeding of the Advances in Artificial Intelligence, Lecture Notes in Computer Science.* Berlin: Springer.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information processing & management*, 41(5), 1019-1033. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.7722&rep =rep1&type=pdf

Lavelli, A., Sebastiani, F, & Zanoli, R. (2004). Distributional term representations: an experimental comparison. *Proceedings of the 13th ACM International Conference on Information and Knowledge Management.* New York, NY: ACM.

Lee, W. N., Shah, N., Sundlass, K., & Musen, M. (2008). Comparison of ontology-based

semantic-similarity measures. *Proceedings of the AMIA annual symposium.* San Francisco, CA: AMIA.

Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R., & Stuckenschmidt, H. (2012). Probabilistic optimization of semantic process model matching. International Conference on Business Process Management. *Lecture Notes in Computer Science* (7481). Berlin, Heidelberg: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-32885-5_25

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc. Retrieved from http://l2r.cs.uiuc.edu/~danr/Teaching/CS598-05/Papers/Lin-Sim.pdf

Liu, Y.-H. (2010, August 18-21). On the Potential Search Effectiveness of MeSH (Medical Subject Headings) Terms. *Proceedings of the third symposium on Information interaction in context.* New York, NY: ACM.

Liu, Y.-H., & Wacholder, N. (2017). Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management*, 53(4), 851-870. Retrieved from https://www.researchgate.net/publication/315665728_Evaluating_the_Impact_of_MeSH_Medical_Subject_Headings_Terms_on_Different_Types_of_Searchers

Liu, M., Lang, B., Gu, Z., & Zeeshan, A. (2017). Measuring similarity of academic articles with semantic profile and joint word embedding. *Tsinghua Science and Technology*, 22(6), 619-632. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8195345

Lu, Z., Kim, W., & Wilbur, W. J. (2009). Evaluating relevance ranking strategies for MEDLINE retrieval. *Journal of the American Medical Informatics Association*, 16(1), 32-36. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605593/pdf/32.S1067502708001916.main.pdf

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press.

Mao, Y., & Lu, Z. (2017). MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of Biomedical Semantics*, 8(15), 1-9. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5392968/pdf/13326_2017_Article_123.pdf

McCutcheon, S. (2009). Keyword vs controlled vocabulary searching: the one with the most tools wins. *The Indexer*, 27(2), 62-65. Retrieved from https://www.researchgate.net/publication/233506550_Keyword_vs_Controlled_Vocabulary_Searching_The_One_with_the_Most_Tools_Wins

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st national conference on Artificial intelligence.* Boston, Massachusetts: AAAI Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word

representations in vector space. *arXiv:1301.3781*. Retrieved from https://arxiv.org/pdf/1301.3781.pdf

Moskovitch, R., Martins, S. B., Behiri, E., Weiss, A., & Shahar, Y. (2007). A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. *Journal of American Medical Information Association*, 14(2), 164–174. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2213470/pdf/164.S1067502706002726.main.pdf

Muddamalle, M. R. (1998). Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. *Journal of the American Society for Information Science*, 49(10), 881-887. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199808%2949%3A10%3C881%3A%3AAID-ASI4%3E3.0.CO%3B2-M

Névéol, A., Zeng, K., & Bodenreider, O. (2006). Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *AMIA Annual Symposium Proceedings.* Washington, DC: AMIA.

Papanikolaou, Y., Tsoumakas, G., Laliotis, M., Markantonatos, N., & Vlahavas, I. (2017). Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models. *Journal of Biomedical Semantics*, 8(1), 43. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5610407/pdf/13326_2017_Article_150.pdf

Peters, T. A., & Kurth, M. (1991). Controlled and uncontrolled vocabulary subject searching in an academic library online catalog. *Information technology and libraries*, 10(3), 201-211. Retrieved from https://www.researchgate.net/publication/234764158 _Controlled_and_Uncontrolled_Vocabulary_Subject_Searching_in_an_Academic_Library_Online_Catalog

Petrakis, E. G., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. (2006). Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. 4th Workshop on Multimedia Semantics (WMS'06). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.3681&rep=rep1&type=pdf

Purcell, G. P., Rennels, G. D., & Shortli, E. H. (1997). Development and evaluation of a context-based document representation for searching the medical literature. *International Journal of Digital Libraries*, 1(3), 288-296. Retrieved from https://link.springer.com/article/10.1007/s007990050023

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th international joint conference on Artificial intelligence (1)*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Rezapour, A., Fakhrahmad, S., & Sadreddini, M. (2011). Applying weighted KNN to word sense disambiguation. *Proceedings of the world congress on engineering.* London: Newswood Limited.

Saka, O., Gulkesen, K., Gulden, B., & Koçgil, O. D. (2005). Evaluation of Two Search Methods in PubMed; the Regular Search and Search by MeSH Terms. *Acta Informatica*

Medica, 13(4), 180-183.

Salton, G. (1970). Automatic text analysis. *Science*, 168(3929), 335-343. Retrieved from http://science.sciencemag.org/content/168/3929/335/tab-pdf

Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the Association for Information Science and Technology*, 23(2), 75-84. Retrieved from https://onlinelibrary.wiley.com/doi /epdf/10.1002/asi.4630230202

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing and Management*, 41(4), 873–890. Retrieved from https://kopernio.com/viewer?doi=10.1016/j.ipm.2004.01.004&route=1

Schnase, J. L., & Cunnius, E. L. (Eds.). (1995). Proceedings from CSCL '95: The First International Conference on Computer Support for Collaborative Learning. Mahwah, NJ: Erlbaum.

Scheffler, F., Schumacher, H., & March, J. (1974). The significance of titles, abstracts, and other portions of technical documents for information retrieval. IEEE Transactions on Professional Communication, 17 (1), 1-8. Retrieved from https://ieeexplore.ieee.org/ document/6592970

Shaw Jr, W. M. (1994). Retrieval expectations, cluster-based effectiveness, and performance standards in the CF database. *Information Processing & Management*, 30(5), 711-723. Retrieved from:
https://www.sciencedirect.com/science/article/abs/pii/ 0306457394900795

Shin, K., Han, S.-Y., & Gelbukh, A. (2004). Balancing manual and automatic indexing for retrieval of paper abstracts. *Lecture Notes in Computer Science(3206).* Berlin, Heidelberg: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-540-30120-2_26

Sotudeh, H., & Houshyar, M. (2018). Comparing discrimination powers of text and citation-based context types. *Scientometrics*, 114(1), 229-251. Retrieved from https://link.springer.com/article/10.1007/s11192-017-2566-9

Srinivasan, P. (1996). Optimal document-indexing vocabulary for MEDLINE. *Information Processing & Management*, 32(5), 503-514. Retrieved from https://www.sciencedirect .com/science/article/abs/pii/0306457396000258

Strader, C. R. (2011). Author-assigned keywords versus Library of Congress subject headings. *Library resources & technical services*, 53(4), 243-250. Retrieved from https://journals.ala.org/index.php/lrts/article/view/5183/6292

Strang, D. (1997). Cheap talk: Managerial discourse on quality circles as an organizational innovation. Presented at the annual meetings of the American Sociological Association, Toronto. Retrieved from:
https://pdfs.semanticscholar.org/0d38/ac8bc7abac3e3f70466e64f84810870d0842.pdf?_ga =2.118267657.1814913894.1544304451-1532394667.1538580728

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5), 331-340. Retrieved from:

https://pdfs.semanticscholar.org/5e4f/0ffb76dcdfb267ed310105b502c369b9418a.pdf

Swanson, D. R. (1960). Searching natural language text by computer. *Science*, 132(3434), 1099-1104. Retrieved from: http://science.sciencemag.org/content/132/3434/1099/tab-pdf

Tang, J., Fong, A. C., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 975-987. Retrieved from: http://keg.cs.tsinghua.edu.cn/jietang/publications/TKDE12-Tang-Name-Disambiguation.pdf

Tenopir, C. (1985). Full text database retrieval performance. *Online Review*, 9(2), 149-164. Retrieved from https://www.emeraldinsight.com/doi/abs/10.1108/eb024180

Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11), 1412-1418. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682526/pdf/btp249.pdf

Nazim Uddin, M., Duong, T. H., Nguyen, N. T., Qi, X.-M., & Jo, G. S. (2013). Semantic similarity measures for enhancing information retrieval in folksonomies. *Expert Systems with Applications*, 40(5), 1645-1653. Retrieved from https://dspace.inha.ac.kr/bitstream/10505/33507/1/35644.pdf

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724-2743. Retrieved from https://www.researchgate.net/publication/319947524_Knowledge_Graph_Embedding_A_Survey_of_Approaches_and_Applications

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., & Ma, J. (2004). Learning to cluster web search results. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY: ACM.

Zhu, S., Zeng, J., & Mamitsuka, H. (2009). Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. *Bioinformatics*, 25(15), 1944-1951. Retrieved from: https://kopernio.com/viewer?doi=10.1093/bioinformatics/btp338&route=6.