Expansion-by-Analogy: A Vector Symbolic Approach to Semantic Search

Trevor Cohen[1], Dominic Widdows[2], and Thomas Rindflesch[3]

[1] University of Texas School of Biomedical Informatics at Houston
[2] Microsoft Bing
[3] National Library of Medicine

Abstract. In this paper, we develop an approach to semantic search that utilizes high-dimensional vector representations to infer the nature of the relationship between query concepts and other concepts in relevant documents. We do so by incorporating outside knowledge drawn from tens of millions of concept-relation-concept triplets, known as *semantic predications*, extracted from the biomedical literature using a Natural Language Processing (NLP) system called SemRep. Inference is accomplished in high-dimensional space using Expansion-by-Analogy, a novel analogical approach to pseudo-relevance feedback, in which the relationships between query concepts and other concepts in documents they occur in guide the query expansion process. The semantic vector based approaches developed in this work show improvements in performance over a baseline bag-of-concepts model, and these improvements are most pronounced on queries that are not conducive to keyword-based search.

Keywords: Distributional Semantics, Information Retrieval, Vector Symbolic Architectures

1 Introduction

Within the biomedical research community, considerable effort has been invested in the development of structured knowledge resources [1]. Efforts have been made to leverage these resources to improve the performance on information retrieval tasks [2–6]. The emphasis of this work has been on the application of controlled terminologies and thesauri as a means to map between variant expressions of the same concept (which has proven especially useful in the genomics domain), at times with the utilizaton of taxonomic (ISA) relationships existing within structured knowledge resources to further elaborate upon query concepts [6]. The utilization of outside resources in order to elaborate upon a stated query is referred to as *query expansion*. In this paper, we attempt to leverage a different sort of knowledge resource for query expansion.

Specifically, we utilize SemMedDB [7], a publicly available database of concept-predicate-concept triplets (such as haloperidol TREATS schizophrenia), or *semantic predications*, that have been extracted from the biomedical literature using a Natural Language Processing system known as SemRep [8]. SemRep extracts predications from biomedical text using domain knowledge in the Unified Medical Language System [9].

For example, the predication "fluoxetine TREATS Major Depressive Disorder" (MDD) is extracted from the phrase "patients who have been successfully treated with fluoxetine for major depression." SemMedDB differs from the human-curated resources that have been utilized in previous work in several ways. Firstly, it contains a richer set of semantic relationships than the "ISA" relationships provided by a taxonomy. SemRep extracts a total of 31 predicate types, of which many relate to clinical medicine (e.g. TREATS, DIAGNOSES) and interactions between substances and biological entities (e.g. INHIBITS, STIMULATES). Therefore an inference mechanism of some sort is required in order to determine which of these possible pathways for query expansion is relevant for a particular concept. Secondly, it cannot be assumed to be perfectly accurate, on account of the difficulties inherent in the automated processing of biomedical language. In a recent evaluation of SemRep, Kilicoglu et al. report .75 precision and .64 recall (.69 f-score) [10]. Finally, it consists of a large number of assertions (more than 50 million) in predication form, and these assertions are not unique - they carry distributional information describing the number of times each predication has been extracted from the corpus of biomedical literature to which SemRep has been applied.

To model this extracted knowledge we use a method called Predication-based Semantic Indexing (PSI) [11], leveraging vector-based approaches to reasoning we have developed during the course of research documented in our prior Quantum Interaction contributions [12–14]. PSI is well suited to modeling the information contained in SemMedDB as it captures both distributional information, in the manner of conventional distributional semantic models (for recent reviews see [15] and [16]), and logical relations between concepts. Therefore, it allows for weighting of the relationships between concepts in accordance with their relative frequency. As PSI is based on the Random Indexing paradigm [17], it provides a computationally convenient way to generate a reduced-dimensional approximation of the information in SemMedDB, which can then be retained in RAM for efficient inference. In this paper, we describe a new approach to query expansion we term Expansion-by-Analogy, in which we infer the significant relationships between query concepts and other concepts in documents they occur in. These inferences are drawn from PSI concept vectors, and the document vectors derived from them, without the need to identify co-occurring concepts explicitly.

2 Expansion-by-Analogy

In previous research, we have developed methods to draw inference from SemMedDB in order to recover held-out therapeutic relationships using a process of analogical reasoning [12, 13, 18, 14]. This process occurs in a high-dimensional space in which concepts are represented as vectors that encode the nature and distribution of the relationships they occur in. Sets of predicates that link one concept to another can be inferred from their vector representations by reversing the vector transformations used during encoding. Once inferred, vector representations of these predicate pathways can be used to find concepts that relate to some other concept in a similar way. As distributional models can derive document representations from concept vector representations, it seems reasonable that one might infer the predicates that connect a query concept to related concepts in a document from a document vector in a similar manner. Fig-

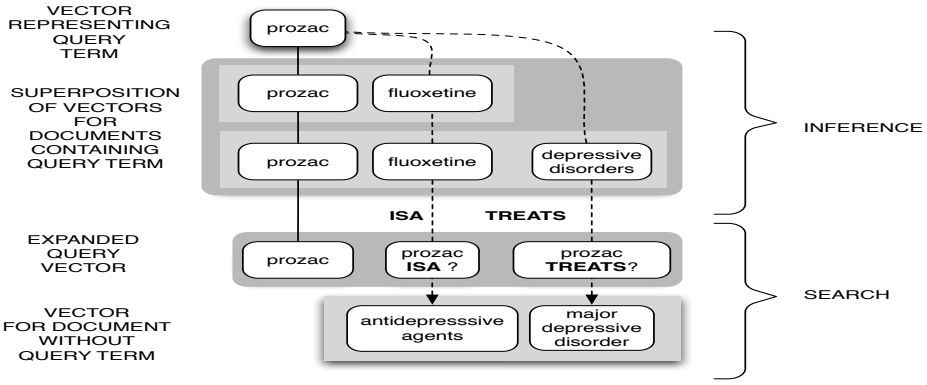

Fig. 1. Inferring predicates from related documents. The solid line indicates direct co-occurrence. The dashed lines indicate inferred predicate paths, which are used to expand the term "prozac" so as to retrieve related documents that do not contain it directly.

ure 1 illustrates this process schematically. Documents containing the query concept "prozac" are retrieved, and the types of relationships (or predicates) between prozac and other concepts in these documents are inferred. It is not necessary to decompose the vector representations of the concepts or the documents to draw these inferences: the connecting predicates are inferred from these representations directly. Vector representations of these predicates can then be used to generate an expanded query vector, which will be similar to vectors representing documents containing concepts that relate to prozac in accordance with the inferred predicates. We refer to this process as Expansion-by-Analogy as it involves applying a relational structure inferred from one set of documents to retrieve others. Such alignment of relational structure is a defining characteristic of analogical reasoning [19]. In the section that follows we will describe the methods through which the vectors and inferences concerned are generated.

3 Mathematical Structure and Methods

3.1 Circular Holographic Reduced Representations

To accomplish the encoding of predicate types within a vector space representation of concepts, we draw upon the capabilities of a family of representational approaches collectively known as Vector Symbolic Architectures (VSAs) [20]. In our experiments the VSA we will employ is Plate's Circular Holographic Reduced Representation (CHRR) [21], which uses complex vectors each of whose coordinates is a number on the unit circle in the complex plane, generated using the implementation developed in [22]. We will refer to such a complex vector as a *circular vector*. The use of complex vectors is standard in physics, particularly quantum theory, but remains comparatively unexplored in artificial intelligence and machine learning [23]. However, the approach we have developed is readily applicable to other VSAs, such as the Binary Spatter Code (BSC) [24]. Though the BSC offers more storage capacity on a bit-for-bit basis [25], this is not

required for the modestly sized document collection we employ here. Furthermore, the simplicity of the circular binding operation (which unlike real-valued HRR, involves simple addition of phase angles) and superposition (which unlike the BSC, requires no random tie-breaking) make the complex vectors more agile for experimentation.

VSAs share with other distributed vector representations the ability to generate composite vector representations of terms or concepts, which we will refer to as *semantic vectors*, by superposing randomly generated *elemental vectors*. For example, the semantic vector representation of a term might be composed from the elemental vectors of terms that surround it. In this way, two terms surrounded by similar other terms will obtain similar vector representations, providing a convenient way to estimate semantic relatedness [26]. In addition to the standard superposition operator (+), VSAs introduce a compositional operator known as *binding*, which we will represent with the symbol $\otimes$. Binding is a multiplication-like operator through which two vectors are combined to form a third vector C that is *dissimilar from* either of its component vectors A and B. Binding has an inverse, which we will represent with the symbol $\oslash$. If C = A $\otimes$ B, then A $\oslash$ C = A $\oslash$ (A $\otimes$ B) $\approx$ B. This recovery may be approximate, but the robust nature of the representation guarantees that A $\oslash$ C is similar enough to B that B can be recognized as the best candidate for A $\oslash$ C in the original set of concepts. Thus the invertible nature of this operator facilitates retrieval of the information it encodes.

In CHRR, binding through circular convolution is accomplished by pairwise multiplication: $X \otimes Y = \{X_1Y_1, X_2Y_2, ...X_{n-1}Y_{n-1}, X_nY_n\}$, which is equivalent to addition of the phase angles of the circular vectors concerned. Binding is inverted by binding to the inverse of the vector concerned: $X \oslash Y = X \otimes Y^{-1}$, where the inverse of a vector is its complex conjugate. Elemental vectors are initialized by randomly assigning a phase angle to each dimension (dimensionality is user-defined). Superposition is accomplished by pairwise addition of the unit circle vectors, and normalization of the result for each circular component. In the implementation used in our experiments, normalization occurs after training concludes, so the sequence in which superposition occurs is not relevant. Also, the "random" initiation of elemental vectors is rendered deterministic by seeding the random number generator with a hash value derived from a string or character of interest following the approach developed in [25], ensuring that incidental overlap between elemental vectors is consistent across experiments.

3.2 Predication-based Semantic Indexing (PSI)

PSI derives vector representations of concepts by superposing obound products representing concept-predicate pairs. Elemental vectors are generated for each concept $E(\text{concept})$, and each relation type $E(\text{PREDICATE})$ and its inverse $E(\text{PREDICATE-IN})$. Semantic vectors are learned gradually by superposing the bound products of elemental vectors representing related items: thus, to encode a predication xRy, the semantic vector for x, written $S(x)$, is incremented by the bound product $E(R) \otimes E(y)$. The same process is applied in reverse to $S(y)$. For example, encoding a single instance of the predication "prozac ISA fluoxetine" is accomplished as follows:

$$S(\text{prozac}) \mathrel{+}= E(\text{ISA}) \otimes E(\text{fluoxetine})$$
$$S(\text{fluoxetine}) \mathrel{+}= E(\text{ISA-INV}) \otimes E(\text{prozac})$$

Thus, the semantic vector for prozac encodes the assertion that it is (a trade name for) fluoxetine, and the semantic vector for fluoxetine encodes the assertion that it has the hyponym prozac. As the same predication may be extracted from many documents, it is advantageous to apply weighting metrics to temper the effect of repeated predications, and increase the influence of infrequently occurring concepts. In our experiments we applied local (LW) and global weighting (GW) metrics as follows:

$$\begin{aligned} S(\text{concept}_1) \mathrel{+}= E(\text{PREDICATE}) \otimes E(\text{concept}_2) \times \text{LW} \times \text{GW} \\ \text{LW} &= \log(1 + \text{total occurrences of predication}) \\ \text{GW} &= IDF(\text{concept}_2) \\ IDF(\text{concept}_2) &= \log \frac{\text{number of documents}}{\text{documents containing concept}_2} \end{aligned}$$

The net result is a set of concept vectors derived from the set of predications each concept occurs in. On account of the reversible nature of the binding operator, this information can be retrieved. So one would anticipate:

$$\begin{aligned} S(\text{fluoxetine}) \oslash E(\text{ISA-INV}) &\approx E(\text{prozac}) \\ S(\text{fluoxetine}) \oslash E(\text{prozac}) &\approx E(\text{ISA-INV}) \end{aligned}$$

This process results in three sets of vector representations, collectively containing a semantic and elemental vector for each concept, as well as an elemental predicate vector for each predicate and its inverse.

3.3 Document vector construction

Document vectors are constructed by superposition of the PSI semantic vectors representing concepts (C_1 to C_n) extracted from this document, as follows:

$$\begin{aligned} S(\text{document D}) = \sum_{i=1}^{n} S(\text{C}_\text{i}) \times TF(\text{C}_\text{i}) \times IDF(\text{C}_\text{i}) \\ TF(\text{C}) &= \text{frequency concept C in document D} \\ IDF(\text{C}) &= \log \frac{\text{number of documents}}{\text{documents containing C}} \end{aligned}$$

3.4 Pseudo-relevance Feedback

Pseudo-relevance feedback is an automated technique based on the assumption that the nearest neighboring documents retrieved using standard methods are relevant, and that their contents can therefore be used to expand the original query. In our experiments we implement a form of pseudo-relevance feedback as follows. For each concept that was extracted from a query, we retrieve ten related documents by finding the ten nearest neighboring semantic document vectors to the semantic vector for this concept. These semantic document vectors representations are then superposed, and the predicates that connect them to the concept in question are inferred by finding the nearest neighboring predicate vectors to the composite query $S(\text{superposed document vectors}) \oslash$

Table 1. Nearest Neighboring Predicate Vectors to $S(\text{prozac}_{\text{NN}}) \oslash E(\text{prozac})$

score	predicate	STD above mean
0.072	ISA-INV	3.83
0.043	LOCATION_OF-INV	2.24
0.040	PREVENTS-INV	2.11
0.039	NEG_STIMULATES	2.06
0.038	SAME_AS	1.99

$E(\text{concept})$, as illustrated schematically earlier in Figure 1. Consider for example the concept "prozac", which was extracted from the query "relationship between prozac and liver disease". First we find the ten-nearest neighboring *semantic* document vectors to the semantic vector for the concept "prozac". Then we superpose those vectors to generate the vector $S(\text{prozac}_{\text{NN}})$, and retrieve the nearest neighboring predicate vectors to the bound product $S(\text{prozac}_{\text{NN}}) \oslash E(\text{prozac})$, which are shown in Table 1.

Predicates with a similarity to $S(\text{prozac}_{\text{NN}}) \oslash E(\text{prozac})$ of more than 2.5 standard deviations above the mean across all predicates are retained. In our case, this applies to ISA-INV. For the concept "liver_disease" (LD) only COEXISTS_WITH met this threshold, and no predicate met the threshold for concept "relationships". So the query vector for "relationship between prozac and liver disease" is constructed as follows:

$$\begin{aligned} QV(\text{prozac}_{\text{ISA-INV}}) &= S(\text{prozac}) + E(\text{ISA-INV}) \otimes E(\text{prozac}) \\ QV(\text{LD}_{\text{COEXISTS\_WITH}}) &= S(\text{LD}) + E(\text{COEXISTS\_WITH}) \otimes E(\text{liver\_disease}) \\ QV(\text{entire query}) &= QV(\text{prozac}) \times IDF(\text{prozac}) + QV(\text{LD}) \times IDF(\text{LD}) \\ &\quad + S(\text{relationships}) \times IDF(\text{relationships}) \end{aligned}$$

Documents are then ranked in order of the relatedness between their vector representations and this composite query.

4 Evaluation

4.1 Methods and Materials

We evaluate Expansion-by-Analogy (EbA) using OHSUMED, a widely-used information retrieval evaluation set [27]. OHSUMED consists of 348,566 clinically-oriented abstracts and titles extracted from 270 medical journals over a five year period, and 106 clinically-oriented queries, with background information. For each query, a set of documents have been annotated as probably relevant, definitely relevant or irrelevant. This annotation is not exhaustive, but does include all relevant articles discovered by a set of human annotators and a baseline information retrieval system. For the purpose of our evaluation, we consider any document annotated as probably or definitely relevant to a query to be relevant. Background information was not utilized - we restricted our evaluation to the query text only. Two queries were excluded from the evaluation - query 8 as no documents are annotated as possibly relevant, and query 68 as this maps to a

single concept, "mesenteric_vasculitis", which was not extracted from any document in the OHSUMED corpus resulting in an empty query vector in concept-based models. Both the queries and documents were processed by SemRep. Rather than attempting to extract predications from these documents, SemRep was configured to extract and normalize concepts recognized in the text. This step would usually precede the extraction of predications, and is accomplished within SemRep by the widely-used MetaMap concept extraction and normalization system [28]. Concepts occurring in more than 100,000 documents were excluded. The concepts extracted from queries and documents are then used in place of the original terms following [29, 30], an approach that has been referred to as "bag-of-concepts" (BoC). Our PSI space was derived from the June 2013 release of SemMedDB [7], which contains 65,465,536 predications extracted from 13,537,476 MEDLINE citations. From this, we created a 2000-dimensional complex-valued PSI space using the open source Semantic Vectors package [31, 23]. Concepts occurring in more than 500,000 predications were excluded, in order to eliminate uninformative frequently-occurring concepts. The purpose of our evalution was to determine whether query expansion improved the performance of the BoC approach, and the extent to which both of these approaches were able to address queries that are difficult to address using a conventional keyword-based, or "bag-of-words" (BoW), approach. To do so, we evaluate the performance of six models, summarized in Table 2.

All models use Term-frequency Inverse Document Frequency (TF-IDF) weighting for the generation of both query and document vector representations, and terms occurring in more than 100,000 documents were excluded from term-based models. In addition to representing the full document-by-term matrix (BoC$_L$), we generate a reduced-dimensional approximation of this space by deriving document vectors from the elemental vector representations of concepts they contain (BoC$_E$). We do so in order to evaluate the extent to which information loss on account of dimension reduction affects performance. We also generate document vectors using the PSI semantic vectors for concepts (BoC_S), so we can distinguish between improvements in performance on account of the enriched nature of these vector representations, and improvements due to inference by analogy. Finally, we evaluate the performance of two bag-of-words based models, one using the full document-by-term matrix (BoW$_L$), and the other using a reduced-dimensional approximation of this space derived from elemental term vector representations, also in an effort to evaluate the effects of information loss during dimension reduction. For each of these models we report the Mean Average Precision (MAP) and the precision at k=10 and 100 (P$^{k=10|100}$), estimated using trec_eval [33].

Table 2. Evaluated models

BoC$_L$	Bag-of-concepts implemented using Apache Lucene [32]
BoC$_E$	Vector space implementation of bag-of-concepts, using $E(\text{concept})$
BoC$_S$	Semantically enriched bag-of-concepts, using $S(\text{concept})$
EbA	Expansion-by-Analogy
BoW$_L$	Bag-of-terms implemented using Apache Lucene
BoW$_E$	Vector space implementation of bag-of-terms, using $E(\text{term})$

Table 3. Cumulative results. Best in class (BoC vs. BoW) and overall are shown in boldface.

	BoC_L	BoC_E	BoC_S	EbA	BoW_L	BoW_E	μSIM
MAP	0.1212	0.1261	0.1530	**0.1574**	**0.1748**	0.1456	**0.1996**
$\text{P}^{k=10}$	**0.2615**	0.2394	0.2587	0.2529	**0.3212**	0.2394	**0.3250**
$\text{P}^{k=100}$	0.1019	0.1158	0.1347	**0.1388**	**0.1397**	0.1232	**0.1589**

4.2 Results and Discussion

It is apparent upon review of the results in Table 3 that unlike the case with other test sets such as the TREC Medical Records collection (see for example [34]), concept extraction has a detrimental effect on overall performance, as compared with BoW approaches. Nonetheless, the baseline performance of BoC is improved considerably by the application of semantic vector based approaches for all metrics shown other than $\text{P}^k = 10$. For example, there is a 26% and 30% improvement in MAP over BoC_L for BoC_S and EbA respectively. These improvements are still not adequate to improve performance beyond a term-based baseline (BoW_L), unless the document-by-term matrix is subjected to the same representational constraints as the reduced-dimensional vector representations (BoW_E). In addition to the results for individual models, we report those obtained by combining the best-performing models in each category (concept-based: EbA, keyword-based: BoW_L) by assigning the mean of the scores from these models to each query document pair (μSIM). These results exceed those obtained by any individual model, as the performance gains of semantic vector based approaches often occur on queries where keyword-based approaches perform poorly.

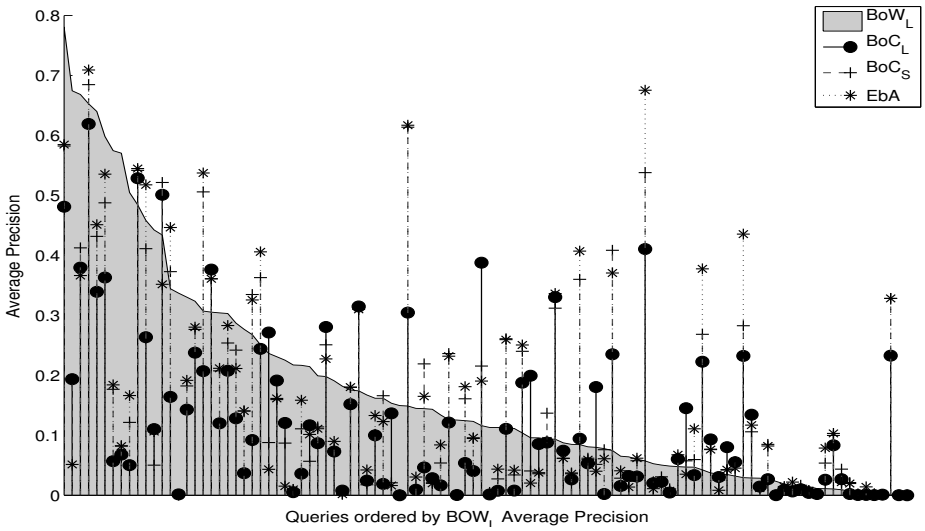

Fig. 2. Comparison of concept-based methods with bag-of-words.

Figure 2 shows the average precision for each query, with queries ordered in accordance with the performance of BoW_L, represented by the grey shaded area of the graph. When BoC_L (—•), BoC_S (- -+) or EbA ($\cdots *$) perform better than the baseline, their respective demarcator appears above the shaded area. It is evident from this figure that in many cases the semantic vector approaches lead to considerable improvements on queries in which the average precision of the term-based baseline was relatively poor (≤ 0.2). In many of these cases, concept extraction alone (—•) had less effect than retrieval based on PSI semantic vector representations (- -+), with further improvements obtained when using EbA to accomplish inference ($\cdots *$) on a number of queries. These queries provide insight into where EbA offers advantages over conventional approaches. Table 4 shows those queries in which the AP of EbA was at least double that of BoW_L. Two characteristics of these queries stand out, though these are not universal. Firstly, many of the queries concern rare clinical entities. However, we were not able to identify a consistent pattern relating the document frequency of query concepts to EbA performance. Secondly, the degree to which many of these queries were expanded (P↑) is often greater than the average across all queries (μ=5.2, σ=3.4) suggesting that identification of further pathways for expansion may be advantageous.

Figure 3 shows all predicates that were used more than five times to expand queries in the set, with counts of the number of times they were employed. These counts are aggregated with respect to direction (such that counts of ISA and ISA-INV are aggregated) and include negated forms of the predicates (e.g. NEG_TREATS), which made up approximately 13% of the 384 expansions that occurred. It is apparent from this figure that EbA uses a much broader range of semantic relations than the ISA relationships that predominate in taxonomy-based query expansion. In fact "ISA" expansions made up a small proportion (< 2 %) of the total number only, and were not utilized for expansion of any of the queries in Table 4. This may be an artifact of our method. EbA is likely to infer paths for expansion from documents that contain an exact match for the query concept concerned. So inferring an ISA pathway would require that both this concept and its taxonomic relative appear in the same document, which may not always be the case. Nonetheless, it is clear that EbA makes extensive use of a wide range of the predicate types represented in SemMedDB.

Table 4. Queries with EbA $>$ 100% improvement in AP over BoW_L and AP $>$ MAP(BoW). P↑ = no. predicates added. %↑ = % improvement.

Query	P↑	%↑
"review article on cholesterol emboli"	1	315
"spontaneous unilateral galactorrhea differential diagnosis and workup"	5	1386
"keratoconus treatment options"	3	260
"indications for and success of pericardial windows and pericardectomies"	15	1081
"diverticulitis differential diagnosis and management"	10	383
"surgery vs percutaneous drainage for lung abscess"	12	141
"infiltrative small bowel processes information about small bowel lymphoma and heavy alpha chain disease"	10	793

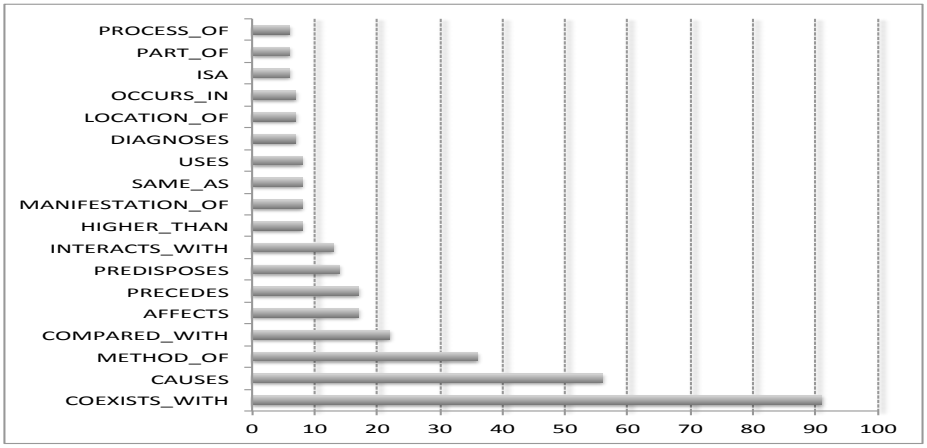

Fig. 3. Frequently-utilized predicates. The X axis shows the number of times this predicate was utilized for expansion across all queries.

Though these results do suggest that EbA may be complementary to BoW and BoC approaches, our evaluation has limitations - we have evaluated our method on a single test set only, and have made no attempt to optimize parameters such as statistical weighting metrics, dimensionality or underlying VSA. Nonetheless, the current evaluation suggests several directions for future research. These include combining EbA with term-based approaches, and extending the length of inferred predicate pathways which has improved performance in other applications [35]. Inferring directions for expansion from a subset of the relevant results may also lead to more pertinent predicates than those identified through the pseudo-relevance based approach we have developed here.

5 Conclusion

This paper describes EbA, an approach to query expansion that utilizes as a knowledge source a reduced-dimensional vector space approximation of tens of millions of semantic predications extracted from the biomedical literature. In addition to document vector representations, expanded vector representations of query concepts are derived from the vectors in this space using a vector-symbolic model of analogical reasoning, and used to construct query vectors. Evaluation on a standard information retrieval test set shows improvements over the aggregate performance of bag-of-concepts vector space approaches, and that the method performs well on a number of queries that are not conducive to standard keyword-based approaches. To do so, EbA utilizes a broader range of semantic relations than is possible with taxonomy-based approaches.

Acknowledgments: This research was supported by US National Library of Medicine grants R21 LM010826 and R01 LM011563. It was also supported in part by the Intramural Research Program of the US National Institutes of Health, National Library of Medicine. We would like to thank Lance DeVine, for contributing the CHRR implementation that was used in this research.

References

1. O. Bodenreider and R. Stevens, "Bio-ontologies: current trends and future directions," *Briefings in bioinformatics*, vol. 7, pp. 256–274, Sept. 2006. PMID: 16899495 PMCID: PMC1847325.
2. W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong, "Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 655–662, ACM, 2007.
3. W. R. Hersh, "Report on the TREC 2004 genomics track," in *ACM SIGIR Forum*, vol. 39, pp. 21–24, ACM, 2005.
4. W. R. Hersh, A. M. Cohen, P. M. Roberts, and H. K. Rekapalli, "TREC 2006 genomics track overview.," in *TREC*, 2006.
5. B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley, "An evaluation of corpus-driven measures of medical concept similarity for information retrieval," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2439–2442, ACM, 2012.
6. G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, "Exploiting medical hierarchies for concept-based information retrieval," in *Proceedings of the Seventeenth Australasian Document Computing Symposium*, pp. 111–114, ACM, 2012.
7. H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: a PubMed-scale repository of biomedical semantic predications," *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.
8. T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of Biomedical Informatics*, vol. 36, pp. 462–477, 2003.
9. O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. Database Issue, p. D267, 2004.
10. H. Kilicoglu, M. Fiszman, G. Rosemblat, S. Marimpietri, and T. C. Rindflesch, "Arguments of nominals in semantic interpretation of biomedical text," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp. 46–54, 2010.
11. T. Cohen, R. Schvaneveldt, and T. Rindflesch, "Predication-based semantic indexing: Permutations as a means to encode predications in semantic space," *AMIA Annu Symp Proc.*, pp. 114–8, 2009.
12. T. Cohen, D. Widdows, R. Schvaneveldt, and T. Rindflesch, "Finding schizophrenia's prozac: Emergent relational similarity in predication space," in *Proc 5th International Symposium on Quantum Interactions. Aberdeen, Scotland. Springer-Verlag Berlin, Heidelberg.*, 2011.
13. T. Cohen, D. Widdows, R. W. Schvaneveldt, and T. C. Rindflesch, "Logical leaps and quantum connectives: Forging paths through predication space," in *Proc AAAI Fall Symp on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pp. 11–13, 2010.
14. T. Cohen, D. Widdows, L. D. Vine, R. Schvaneveldt, and T. C. Rindflesch, "Many Paths Lead to Discovery: Analogical Retrieval of Cancer Therapies," in *Quantum Interaction* (J. R. Busemeyer, F. Dubois, A. Lambert-Mogiliansky, and M. Melucci, eds.), no. 7620 in Lecture Notes in Computer Science, pp. 90–101, Springer Berlin Heidelberg, Jan. 2012.
15. T. Cohen and D. Widdows, "Empirical distributional semantics: methods and biomedical applications," *Journal of Biomedical Informatics*, vol. 42, pp. 390–405, Apr. 2009.
16. P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.
17. P. Kanerva, J. Kristofersson, and A. Holst, "Random indexing of text samples for latent semantic analysis," *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, vol. 1036, 2000.

18. T. Cohen, D. Widdows, R. Schvaneveldt, P. Davies, and T. Rindflesch, "Discovering discovery patterns with predication-based semantic indexing," *Journal of Biomedical Informatics*, vol. 45, pp. 1049–65, Dec 2012.
19. D. Gentner and A. B. Markman, "Structure mapping in analogy and similarity.," *American psychologist*, vol. 52, no. 1, p. 45, 1997.
20. R. W. Gayler, "Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience," in *In Peter Slezak (Ed.), ICCS/ASCS International Conference on Cognitive Science*, (Sydney, Australia. University of New South Wales.), pp. 133–138, 2004.
21. T. A. Plate, *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. Stanford, CA.: CSLI Publications, 2003.
22. L. De Vine and P. Bruza, "Semantic oscillations: Encoding context and structure in complex valued holographic vectors," *Proc AAAI Fall Symp on Quantum Informatics for Cognitive, Social, and Semantic Processes*, 2010.
23. D. Widdows, T. Cohen, and L. De Vine, "Real, complex, and binary semantic vectors," in *Proc Sixth Intl Symp on Quantum Interactions, Paris, France.*, 2012.
24. P. Kanerva, "Binary spatter-coding of ordered k-tuples," *Artificial Neural Networks—ICANN 96*, pp. 869–873, 1996.
25. M. Wahle, D. Widdows, J. R. Herskovic, E. V. Bernstam, and T. Cohen, "Deterministic Binary Vectors for Efficient Automated Indexing of MEDLINE/PubMed Abstracts," *AMIA Annual Symposium Proceedings*, vol. 2012, pp. 940–949, Nov. 2012.
26. J. Karlgren and M. Sahlgren, "From words to understanding," *Foundations of Real-World Intelligence*, pp. 294–308, 2001.
27. W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "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*, pp. 192–201, 1994.
28. A. R. Aronson and F. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17:, pp. 229 –236, May 2010.
29. W. R. Hersh, D. H. Hickam, R. B. Haynes, and K. A. McKibbon, "A performance and failure analysis of SAPHIRE with a MEDLINE test collection," *Journal of the American Medical Informatics Association*, vol. 1, pp. 51–60, Jan. 1994. PMID: 7719787.
30. A. R. Aronson, T. C. Rindflesch, and A. C. Browne, "Exploiting a large thesaurus for information retrieval.," in *RIAO*, vol. 94, pp. 197–216, 1994.
31. D. Widdows and T. Cohen, "The semantic vectors package: New algorithms and public tools for distributional semantics," in *Fourth IEEE International Conference on Semantic Computing (ICSC)*, 2010.
32. "Apache lucene (https : //lucene.apache.org),"
33. "trec-eval (http : //trec.nist.gov/trec_eval/),"
34. B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley, "Graph-based concept weighting for medical information retrieval," in *Proceedings of the Seventeenth Australasian Document Computing Symposium*, ADCS '12, (New York, NY, USA), pp. 80–87, ACM, 2012.
35. T. Cohen, Widdows, Dominic, R. Schvaneveldt, and T. Rindflesch, "Discovery at a distance: Farther journey's in predication space.," in *Proceedings of the First International Workshop on the role of Semantic Web in Literature-Based Discovery (SWLBD2012), The IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012); Oct 4-7 2012; Philadelphia, PA, USA.*, 2012.