An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval

Bevan Koopman[1,2], Guido Zuccon[1], Peter Bruza[2], Laurianne Sitbon[2], Michael Lawley[1]
[1]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2]School of Information Systems, Queensland University of Technology, Brisbane, Australia
{b.koopman, p.bruza, l.sitbon}@qut.edu.au, {guido.zuccon, michael.lawley}@csiro.au

ABSTRACT

Measures of semantic similarity between medical concepts are central to a number of techniques in medical informatics, including query expansion in medical information retrieval. Previous work has mainly considered thesaurus-based path measures of semantic similarity and has not compared different corpus-driven approaches in depth. We evaluate the effectiveness of eight common corpus-driven measures in capturing semantic relatedness and compare these against human judged concept pairs assessed by medical professionals. Our results show that certain corpus-driven measures correlate strongly (≈ 0.8) with human judgements. An important finding is that performance was significantly affected by the choice of corpus used in priming the measure, i.e., used as evidence from which corpus-driven similarities are drawn. This paper provides guidelines for the implementation of semantic similarity measures for medical informatics and concludes with implications for medical information retrieval.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]

Keywords: Semantic Similarity, Medical Information Retrieval

1. INTRODUCTION

The advent of electronic medical records, and large corpora of medical documents (e.g. MEDLINE), create an increasing need for information retrieval (IR) systems tailored to searching medical free-text [9]. Searching medical records presents some specific challenges — vocabulary mismatch is prevalent as the same concepts can be expressed using different terms. More challenging are situations where relevance must be inferred, e.g., the presence of a certain organism in a laboratory report denoting a certain disease, even though the disease is not stated explicitly (for example, *Varicella zoster* virus → *Chicken pox*). Overcoming these challenges requires IR models capable of accurately determining *semantic similarity* between medical concepts. Semantic similarity measures are central to several techniques used in medical informatics, including: query expansion and relevance feedback [16, 7], literature-based discovery (e.g., drug discovery [1]), clustering (e.g., gene clustering [8]), and ontology construction or maintenance [6].

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.

Medical domain knowledge has been formally represented in a number of medical thesauri/ontologies such as Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS). In these symbolic representations medical concepts are organised into inheritance hierarchies and into inter-concept relationships (e.g. <organism> *causes* <disease>), thus creating a graph of concepts. Early approaches to measuring semantic similarity used thesaurus-based path measures between medical concepts[1]. An alternative to path-based measures are corpus-driven approaches, e.g., Latent Semantic Analysis (LSA), which commonly exploit co-occurrence statistics to determine similarity.

Corpus-driven approaches are used extensively in medical IR, particularly for query expansion. This is evident by the number of teams using these approaches in the TREC 2011 Medical Records track [16]. An evaluation by Pedersen et al. [12] using human judged concept pairs provided by medical professionals found that a corpus-driven approach adapted from LSA (which they call *Context Vector*) outperformed a number of existing path-based measures. Similarly, Sanchez et al. [14] showed that using the Web as a corpus also outperformed path-based measures. Finally, Trieschnigg et al. [15] proposed Cross Entropy Reduction (CER), a language modelling approach that outperformed a path baseline. These different approaches demonstrate the plethora of corpus-driven approaches available as measures of semantic similarity. To our knowledge, no previous work compares these corpus-based approaches for the purposes of similarity judgements in the medical domain. Pedersen's et al. evaluation only used a single LSA adaptation, while CER was compared against only two basic corpus-driven baselines. Additionally, important implementation issues have not been explored, like the choice of dimensionality and robustness across multiple collections, a factor that we show has significant effect on performance. This paper considers these issues by evaluating 8 different corpus-driven measures on two corpora against two separate datasets of human judged concept pairs.

2. METHODS

Evaluation of 8 corpus-driven measures was performed against two separate datasets of human judged medical concept pairs. An example of a concept pair is (*Congestive heart failure*, *Pulmonary edema*). Semantic similarity between concept pairs was computed using the following measures:

[1]Semantic similarity being inversely proportional to the length of the path between two concepts in the thesaurus.

1. Random Indexing [13] (RI): a technique that constructs an approximation of the full term-document matrix by assigning each term a unique *index* vector. The index vector is of fixed length and sparsely consists of randomly assigned -1s, 0 and 1s. Similarity was measured as the cosine angle between two concepts' index vectors. Random Indexing was evaluated using 50, 150, 300, and 500 dimensions; results were averaged over 10 runs for each dimensional setting.
2. Latent Semantic Analysis (LSA): evaluated on 50, 150, 300, and 500 dimensions. Similarity was computed as the cosine angle between reduced concept vectors.[2]
3. Hyperspace Analogue to Language [11] (HAL): constructs a full term-term co-occurrence matrix with context window of size 5[3]. Similarity was calculated as the cosine of the angle between the two HAL based concept vectors.
4. Document Vector Cosine Similarity (DocCosine): cosine angle between concepts represented by document vectors; weighted with tf-idf.
5. Positive Pointwise Mutual Information [4] (+PMI): variation of PMI where negative values are substituted by zero-values. Bullinaria and Levy [4] found negative PMI values, which correspond to less than expected number of co-occurrence, indicate a poor coverage of the concepts in the corpus. This is often the case in the medical domain due to infrequently appearing concepts referring to specific diseases or rare conditions. In preliminary experiments +PMI significantly outperformed PMI.
6. Cross Entropy Reduction [15] (CER): Trieschnigg et al. [15] distance between two concept's unigram language models. A concept language model θ_c is defined as a distribution over concepts based on the concatenation of all documents containing concept c; background smoothing using Jelinek-Mercer.
7. Language Model + Jensen-Shannon divergence (LM JSD): unigram concept language model (constructed in the same manner as CER) but comparison was performed using standard Jensen-Shannon divergence.
8. Latent Dirichlet Allocation (LDA): topic model evaluated using 50, 150, 300 and 500 topics. Similarity between two concepts was determined by comparing their topic distributions $P(\text{topic}|c)$ using Jensen-Shannon divergence.

3. EXPERIMENTAL SETUP

Two separate datasets of human judged concept pairs were used for evaluation. The first dataset consists of twenty-nine[4] UMLS medical concept pairs, as developed by Pedersen et al. [12], involving 3 physician and 9 clinical terminologists; inter-coder correlation was reported to be 0.85. A concept pair example is (*Brain tumor*, *Intracranial hemorrhage*), judged as having a similarity of 2.0 on a scale of 1.0 (unrelated) to 4.0 (synonymous). We refer to this dataset as Ped. The second dataset, from Cavides and Cimino [5], contains forty-five MeSH/UMLS concept pairs[5] judged by three physicians on a scale of 1 to 10; Cavides and Cimino report "consensus" amongst judges, but no precise value was reported. This dataset is referred to as Cav.

Two separate corpora were used as data to prime each corpus-driven method. The first corpus was MedTrack, a collection of 100,866 clinical record documents used in the TREC 2011 Medical Records Track. Documents belonging to a single patient's admission were treated as sub-documents and were concatenated together into a single document called a patient *visit* document. The corpus then contained 17,198 patient visit documents. This was done to encapsulate the closely related content of different reports (e.g. pathology report and surgical report) belonging to the same patient admission[6]. The second corpus used was OHSUMED, a MEDLINE subset consisting of 348,566 medical journal abstracts, as used in TREC 2000 Filtering Track. Statistics for each corpus are provided in Table 1.

Corpus	#Docs	Avg. doc. len.	#Vocab.
MedTrack	17,198*	932	54,546
OHSUMED	293,856	100	55,390

*100,866 original reports collapsed to 17,198 patient *visit* documents.

Table 1: Collection statistics of the test corpora: MedTrack, collection of clinical patient records; and OHSUMED, MEDLINE abstracts.

For both corpora, the original textual documents were translated into UMLS concept identifiers using MetaMap, the biomedical concept identification system [2]. After processing, the individual documents contained only UMLS concept ids, for example the phrase *Congestive heart failure* in the original document will be replaced with `C0018802` in the new document; more details of this approach are provided in [10]. Both test datasets, Ped and Cav, contained UMLS concept pairs (which may actually represent term phrases rather than single terms); converting the test corpora to concepts thus allows direct comparison of the single concept pairs contained in the two datasets.

Each of the 8 models outlined in the Methods section provide a representation of a concept, for example, in DocCosine a concept is a vector based on the documents the given concept appears in. Similarity can be determined by comparing two concepts' representations. For each similarity measure, comparison was made against human judges for each dataset (Ped and Cav) using Pearson's correlation coefficient.

4. RESULTS & DISCUSSION

Results showing the correlation coefficient against human judges for each corpus-driven method are reported in Figure 1. The x-axis is ordered by decreasing correlation averaged across all datasets/corpora[7].

[2]Both RI and LSA were implemented using the SemanticVectors software package:
`http://code.google.com/p/semanticvectors`

[3]Lund & Burgess [11] found HAL was most effective with small context windows in this range.

[4]One concept pair (*Lymphoid hyperplasia*) was removed from Pedersen's original 30 as it was not found in our test collections.

[5]10 pairs containing the concept `C0030631`, not present in the test corpus, were removed.

[6]Collapsing reports to patient visits was a common practise among many TREC MedTrack participants [16].

[7]LDA (avg.) is the average for LDA across 50, 150, 300, 500 topics, all of which exhibit almost equivalent results.

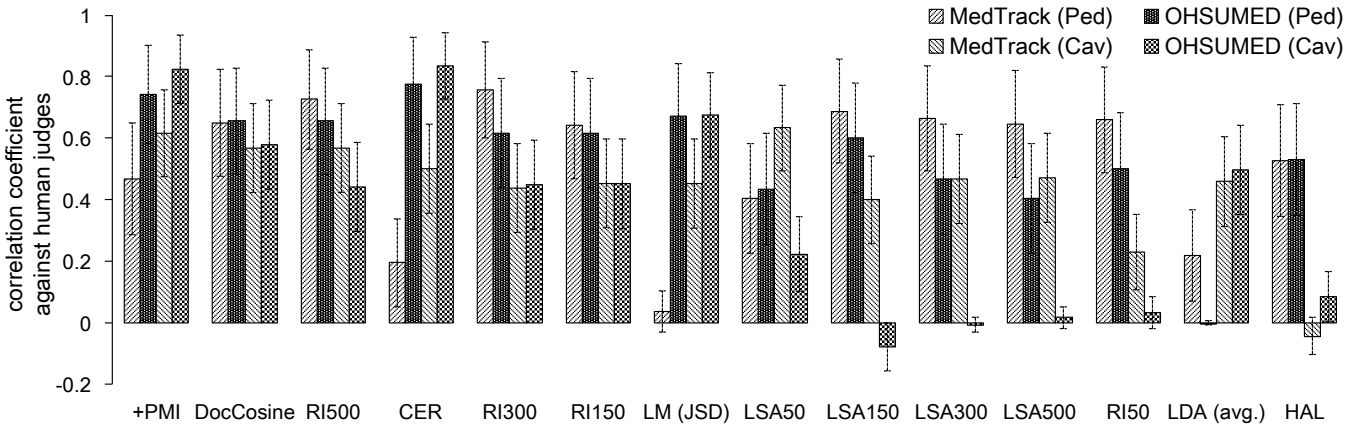

Figure 1: Correlation coefficient against human judged similarity for each corpus-driven semantic similarity measure. Judgements made against two gold standard datasets (Ped & Cav) using two corpora (MedTrack & OHSUMED). x-axis ordered by decreasing correlation averaged across all datasets/corpora; error bars signify confidence interval at 95%.

The first observation we make is that similar types of measures demonstrate similar results: the three probabilistic language model measures, +PMI, CER and LM (JSD) exhibit comparable performance profiles across datasets / corpora. Similarly, the vector-based measures (RI and LSA and DocCosine) exhibit similar profiles between each other and across different dimensions.

Considering the best performing measures, Table 2 provides a breakdown of the top 3 semantic similarity measures for each dataset / corpus.

	Dataset	
Corpus	Ped	Cav
MedTrack	RI300, LSA150, DC	LSA50, +PMI, DC
OHSUMED	CER, +PMI, LM/DC	CER, +PMI, LM

Table 2: Top 3 semantic similarity measures for each corpus and dataset. DocCosine abbreviated to DC.

Consensus is observed between the two datasets, Ped and Cav. However, the best measure differs significantly between the two corpora. In general, vector-based measures perform best when primed with the MedTrack corpus, while probabilistic measures are most effective primed with OHSUMED. This may be explained by the different characteristics of the two corpora: MedTrack contains detailed clinical notes from patient encounters, whereas OHSUMED contains MEDLINE article abstracts. As a result, the *scope* of concepts found in a documents differs between the two collections. Clinical notes relating to a patient's admission may cover a wide range of different concepts, especially if they have been admitted with multiple conditions or for a lengthy period. In contrast, journal abstracts are descriptions of a particular topic and are therefore typically narrower in scope. The probabilistic measures use the whole document as the "context window" for determining co-occurrence, OHSUMED's documents of narrower scope therefore offer more precise context windows, whereas the wider scoped MedTrack documents may contain more noise. In addition to the nature of the documents found in each corpus, the average *document length* differs considerably — MedTrack documents are about an order of magnitude larger (Table 1). Intuitively, longer documents will, in general, cover more topics and be wider in scope. The vector-based measures benefit from the addition context found in the longer documents, which is in contrast to the probabilistic measures.

The nature of the *language* also differs between the two corpora. MEDLINE abstracts contain precise descriptions of a particular topic, whereas clinical records are often terse narratives with considerable jargon and shorthand — and in some cases typographic errors.

Given the differences in *scope*, *document length* and *language* of the two corpora we could hypothesis that OHSUMED appears a higher quality corpus for similarity judgements, and that measures primed with MedTrack would exhibit degraded performance. However, the results do not affirm this hypothesis. Probabilistic measures primed with OHSUMED display excellent results, however, the longer, less consistent documents found in MedTrack still provide good evidence for similarity judgements when used with vector-based methods.

Table 2 also highlights the robustness of +PMI and DocCosine, which both occupy three out of four cells. The traditional IR measure of DocCosine, although not producing the best results on a single test, is particularly stable across both corpora and datasets. Both +PMI and DocCosine are simple and computationally efficient, making them more attractive than more computationally intensive measures such as LSA and language model-based measures. Certain measures may perform well on one particular collection / dataset, but have poor performance on others — LM (JSD), LDA and HAL all exhibit this behaviour.

More generally, the results reaffirm the findings of Pedersen et al. that corpus-driven approaches outperform path-based measures, which failed to yield a correlation greater than 0.5[8]. Additionally, our findings using *vector-based* measures are in line with Petersen et al. who reported a 0.69 correlation obtained using their *Context Vector* measure on

[8]Path-based measures are *corpus independent*, based on the UMLS network, as such Pedersen's results can be used for a direct comparison in our study.

the Mayo Clinic Corpus of Clinical Notes; our vector-based measure results using MedTrack were ≈ 0.7. MedTrack and the Mayo Clinic Corpus are of similar size and nature (both being clinical records)[9].

An outcome of this study are a set of guidelines for the implementation of corpus-based semantic similarity measures for medical text:

1. The choice of corpus used to prime the similarity measure is an important consideration that may significantly affect the performance of the particular measure.
2. More specifically, the characteristics of individual documents should be considered. Do documents cover a range of topics, in which case vector-based measures are preferable, or are they smaller in scope, probabilistic methods are then preferred. Average document length can be an indicator of scope — large documents typically cover more topics. Additionally, the type of language (e.g., clinical notes vs. medical literature) should be taken into consideration.
3. +PMI and DocCosine are robust across collections and datasets and have the added advantage of being computationally efficient. Other measures may perform well on certain collection / datasets, but can perform extremely poorly in certain cases, it may be best to avoid these measures.
4. When implementing a semantic similarity on a particular corpus the two datasets can be used to find a measure most appropriate to the nature of the corpus documents. Both Ped and Cav are publicly available.

The reported findings may have important impacts for medical information retrieval, specifically for systems making significant use of query expansion and relevance feedback, as was the case with participants of TREC MedTrack. Firstly, the effective of corpus-based query expansion varied significantly between participants of TREC MedTrack — some techniques showed gains, while others degraded performance. Although a number of factors affect query expansion performance, a poor semantic similarity measure could certainly be a major contributor. The most appropriate similarly measure, based on the findings of this study, should be considered when employing corpus-based query expansion.

Finally, having highlighted the choice of corpus as an important consideration, we conjecture that in some cases it may be advantageous to prime the similarity measure with a separate corpus from the one being used for retrieval. For example, when searching medical literature (e.g. OHSUMED), priming with clinical records (e.g. those found in MedTrack) may increase effectiveness. In the literature there is evidence supporting the use of Wikipedia as a background priming corpus [3]. An in-depth evaluation of this aspect is left to future work.

5. CONCLUSION

In this paper we evaluate eight different corpus-driven approaches to determining the semantic similarity between medical concepts. Corpus-driven approaches exhibit strong correlations (up to ≈ 0.8) with human judged concept pairs provided by medical professionals. Our findings show that the choice of corpus used to prime the similarity measure can significantly affect performance. We provide a number of guidelines for the use of semantic similarity measures that include consideration of document scope, length and language. Simple measures such as +PMI and DocCosine demonstrate effective and robustness across evaluations. This work provides an in depth review of corpus-driven semantic similarity measures, a technique central to medical informatics.

[9]Note that the Mayo Clinic Corpus of Clinical Notes corpus is not publicly available.

Acknowledgements Dolf Trieschnigg kindly provided the data from his experiments on CER [15] and was used to validate our CER implementation.

6. REFERENCES

[1] P. Agarwal and D. B. Searls. Can literature analysis identify innovation drivers in drug discovery? *Nature reviews. Drug discovery*, 8(11):865–78, Nov. 2009.

[2] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3):229–236, 2010.

[3] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *SIGIR '11*, pages 605–614, Beijing, China, July 2011.

[4] J. Bullinaria and J. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510, 2007.

[5] J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Journal of biomedical informatics*, 37(2):77–85, Apr. 2004.

[6] S. Cederberg and D. Widdows. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proc of CoNLL'03*, pages 111–118, NJ, USA, 2003.

[7] T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.

[8] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. D. Moor. Evaluation Of The Vector Space Representation In Text-Based Gene Clustering. In *Proc Pacific Symposium of Biocomputing*, pages 391–402, Kauai, Hawaii, 2003.

[9] W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Verlag, New York, 3rd edition, 2009.

[10] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information Retrieval. *Australasian Medical Journal*, In Press, 2012.

[11] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods*, 28(2):203–208, 1996.

[12] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

[13] M. Sahlgren. An introduction to random indexing. In *Proc of TKE'05*, pages 1–9, Leipzig, Germany, 2005.

[14] D. Sánchez, M. Batet, and A. Valls. Computing Knowledge-Based Semantic Similarity from the Web: An Application to the Biomedical Domain. In *Proc of Knowledge Science, Engineering and Management*, KSEM'09, pages 17–28, Berlin, Germany, 2009.

[15] D. Trieschnigg, E. Meij, M. de Rijke, and W. Kraaij. Measuring concept relatedness using language models. In *Proc of SIGIR'08*, pages 823–824, NY, USA, 2008.

[16] E. Voorhees and R. Tong. Overview of the TREC Medical Records Track. In *Proc of TREC'11*, MD, USA, 2011.