Similarity Metrics for Clustering PubMed Abstracts for Evidence Based Medicine

Hamed Hassanzadeh[†] **Diego Mollá**[◇] **Tudor Groza**[‡] **Anthony Nguyen**[♮] **Jane Hunter**[†]

[†]School of ITEE, The University of Queensland, Brisbane, QLD, Australia
[◇]Department of Computing, Macquarie University, Sydney, NSW, Australia
[‡]Garvan Institute of Medical Research, Darlinghurst, NSW, Australia
[♮]The Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia

h.hassanzadeh@uq.edu.au, diego.molla-aliod@mq.edu.au
t.groza@garvan.org.au, anthony.nguyen@csiro.au, jane@itee.uq.edu.au

Abstract

We present a clustering approach for documents returned by a PubMed search, which enable the organisation of evidence underpinning clinical recommendations for Evidence Based Medicine. Our approach uses a combination of document similarity metrics, which are fed to an agglomerative hierarchical clusterer. These metrics quantify the similarity of published abstracts from syntactic, semantic, and statistical perspectives. Several evaluations have been performed, including: an evaluation that uses ideal documents as selected and clustered by clinical experts; a method that maps the output of PubMed to the ideal clusters annotated by the experts; and an alternative evaluation that uses the manual clustering of abstracts. The results of using our similarity metrics approach shows an improvement over K-means and hierarchical clustering methods using TF-IDF.

1 Introduction

Evidence Based Medicine (EBM) is about individual patients care and providing the best treatments using the best available evidence. The motivation of EBM is that clinicians would be able to make more judicious decisions if they had access to up-to-date clinical evidence relevant to the case at hand. This evidence can be found in scholarly publications available in repositories such as PubMed[1]. The volume of available publications is enormous and expanding. PubMed repository, for example, indexes over 24 million abstracts. As a result, methods are required to present relevant recommendations to the clinician in a manner that highlights the clinical evidence and its quality.

The EBMSummariser corpus (Mollá and Santiago-martinez, 2011) is a collection of evidence-based recommendations published in the *Clinical Inquiries* column of the *Journal of Family Practice*[2], together with the abstracts of publications that provide evidence for the recommendations. Visual inspection of the EBMSummariser corpus suggests that a combination of information retrieval, clustering and multi-document summarisation would be useful to present the clinical recommendations and the supporting evidence to the clinician.

Figure 1 shows the title (question) and abstract (answer) associated with one recommendation (Mounsey and Henry, 2009) of the EBM-Summariser corpus. The figure shows three main recommendations for treatments to hemorrhoids. Each treatment is briefly presented, and the quality of each recommendation is graded (A, B, C) according to the Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004). Following the abstract of the three recommendations (not shown in Figure 1), the main text provides the details of the main evidence supporting each treatment, together with the references of relevant publications. A reference may be used for recommending several of the treatments listed in the recommendations. Each recommendation is treated in this study as a cluster of references for evaluation purposes, and the corpus therefore contains overlapping clusters.

It has been observed that a simple K-means clustering approach provides a very strong base-

[1] www.ncbi.nlm.nih.gov/pubmed

[2] www.jfponline.com/articles/clinical-inquiries.html

Hamed Hassanzadeh, Diego Mollá, Tudor Groza, Anthony Nguyen and Jane Hunter. 2015. Similarity Metrics for Clustering PubMed Abstracts for Evidence Based Medicine . In *Proceedings of Australasian Language Technology Association Workshop*, pages 48−56.

Which treatments work best for Hemorrhoids?
Excision is the most effective treatment for thrombosed external hemorrhoids (strength of recommendation [SOR]: B, retrospective studies). For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy (SOR: A, systematic reviews). Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence (SOR: A, systematic reviews).

Figure 1: Title and abstract of one sample (Mounsey and Henry, 2009) of the *Clinical Inquiry* section of *Journal of Family Practice*.

line for non-overlapping clustering of the EBM-Summariser corpus (Shash and Mollá, 2013; Ekbal et al., 2013). Past work was based on the clustering of the documents included in the EBMSummariser corpus. But in a more realistic scenario one would need to cluster the output from a search engine. Such output would be expected to produce much noisier data that might not be easy to cluster.

In this paper, we cluster documents retrieved from PubMed searches. We propose a hierarchical clustering method that uses custom-defined similarity metrics. We perform a couple of evaluations using the output of PubMed searches and the EBMSummariser corpus. Our results indicate that this method outperforms a K-means baseline for both the EBMSummariser corpus and PubMed's retrieved documents.

The remainder of the paper is structured as follows. Section 2 describes related work. Section 3 provides details of the clustering approach and the evaluation approaches. Section 4 presents the results, and Section 5 concludes this paper.

2 Related Work

Document clustering is an unsupervised machine learning task that aims to discover natural groupings of data and has been used for EBM in several studies. Lin and Demner-Fushman (2007) clustered MEDLINE citations based on the occurrence of specific mentions of interventions in the document abstracts. Lin et al. (2007) used K-means clustering to group PubMed query search results based on TF-IDF. Ekbal et al. (2013) used genetic algorithms and multi-objective optimisation to cluster the abstracts referred in the EBMSummariser corpus, and in general observed that it was difficult to improve on Shash and Mollá (2013)'s K-means baseline, which uses TF-IDF similar to Lin and Demner-Fushman (2007).

It can be argued that clustering the abstracts that are cited in the EBMSummariser corpus is easier than clustering those from Pubmed search results, since the documents in the corpus have been curated by experts. As a result, all documents are relevant to the query, and they would probably cluster according to the criteria determined by the expert. However, in a more realistic scenario the documents that need to be clustered are frequently the output of a search engine. Therefore, there might be documents that are not relevant, as well as duplicates and redundant information. An uneven distribution of documents among the clusters may also result.

There are several approaches to cluster search engine results (Carpineto et al., 2009). A common approach is to cluster the documents snippets (*i.e.,* the brief summaries appearing in the search results page) instead of the entire documents (Ferragina and Gulli, 2008). Our approach for clustering search engine results is similar to this group of approaches, since we only use the abstract of publications instead of the whole articles. The abstracts of scholarly publications usually contain the key information that is reported in the document. Hence, it can be considered that there is less noise in abstracts compared to the entire document (from a document clustering perspective). A number of clustering approaches can then be employed to generate meaningful clusters of documents from search results (Zamir and Etzioni, 1998; Carpineto et al., 2009).

3 Materials and Method

In this section we describe an alternative to K-means clustering over TF-IDF data. In particular, we devise separate measures of document similarity and apply hierarchical clustering using our custom matrix of similarities.

We first introduce the proposed semantic similarity measures for quantifying the similarity of abstracts. We then describe the process of preparing and annotating appropriate data for clustering semantically similar abstracts. Finally, the experimental set up will be explained.

Prior to describing the similarity measures, a glossary of the keywords that are used in this section is introduced:

Effective words: The words that have noun, verb, and adjective Part of Speech (POS) roles.

Effective lemmas: Lemma (canonical form) of effective words of an abstract.

Skipped bigrams: The pairs of words which are created by combining two words in an abstract that are located in arbitrary positions.

3.1 Quantifying similarity of PubMed abstracts

In order to be able to group the abstracts which are related to the same answer (recommendation) for a particular question, the semantic similarity of the abstracts was examined. A number of abstract-level similarity measures were devised to quantify the semantic similarity of a pair of abstracts. Since formulating the similarity of two natural language pieces of text is a complex task, we performed a comprehensive quantification of textual semantic similarity by comparing two abstracts from different perspectives. Each of the proposed similarity measures represents a different view of the similarity of two abstracts, and therefore the sum of all of them represents a combined view of each of these perspectives. The details of these measures can be found below. Note that all the similarity measures have a normalised value between zero (lowest similarity) and one (highest similarity).

Word-level similarity: This measure calculates the number of overlapping words in two abstracts which is then normalised by the size of the longer abstract (in terms of the number of all words). The words are compared in their original forms in the abstracts (even if there were multiple occurrences). Equation (1) depicts the calculation of Word-level Similarity (WS).

$$WS(A_1, A_2) = \frac{\sum_{w_i \in A_1} \begin{cases} 1 & \text{if } w_i \text{ is in } A_2 \\ 0 & \text{Otherwise} \end{cases}}{L} \tag{1}$$

where A_1 and A_2 refer to the bags of all words in two given abstracts (including multiple occurrences of words), and L is the size of the longest abstract in the pair.

Word's lemma similarity: This measure is calculated similarly to the previous measure, but the lemma of words from a pair of abstracts are compared to each other, instead of their original display forms in the text, using WordNet (Miller, 1995). For example, for a given pair of words, such as *criteria* and *corpora*, their canonical forms (*i.e.*, *criterion* and *corpus*, respectively) are looked up in WordNet prior to performing the comparison.

Set intersection of effective lemmas: The sets of lemmas of effective words of abstract pairs are compared. The number of overlapping words (or the intersection of two sets) is normalised by the size of the smaller abstract. In contrast to the previous measure, only unique effective lemmas participate in the calculation of this measure. This measure is calculated as follows:

$$SEL(A_1, A_2) = \frac{|A_1^{set} \cap A_2^{set}|}{S} \tag{2}$$

In Equation (2), A_1^{set} and A_2^{set} are the sets of effective lemmas of two abstracts, and S is the size of the smallest abstract in a pair.

Sequence of words overlap: We generate sliding windows of different sizes of words, from a window of two words up to the size of the longest sentence in a pair of abstracts. We compute the number of equal sequences of words of two abstracts (irrespective of length). Also, we keep the size of the longest equal sequence of words that the two abstracts share together. Hence, this results in two similarity measures; (*i*) the number of shared sequences of different sizes, and (*ii*) the size of the longest shared sequence. Due to the variety of sizes of sentences / abstracts and therefore varying sizes and number of sequences, we normalise each of these measures to reach a value between zero and one. In addition, following the same rationale, sequence-based measures are calculated by only considering effective words in abstracts, and alternatively, from a grammatical perspective, by only considering POS tags of the constituent words of abstracts. The number of shared sequences (or Shared Sequence Frequency — *SSF*) for two given abstracts (*i.e.*, A_1 and A_2) is calculated as follows:

$$SSF(A_1, A_2) = \frac{\sum_{l=2}^{M} \frac{\sum_{S_l \in A_1} \begin{cases} 1 & \text{if } S_l \in A_2 \\ 0 & \text{Otherwise} \end{cases}}{N}}{M} \tag{3}$$

In Equation (3), M is the size of the longest sentence in both abstracts and N is the number of available sequences (*i.e.*, S in formula) with size l.

POS tags sequence alignment: For this similarity measure, a sequence of the POS tags of words in an abstract is generated. The Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) was employed for aligning two sequences of POS tags from a pair of abstracts to find their similarity ratio. The Needleman-Wunsch algorithm is an efficient approach for finding the best alignment between two sequences, and has been successfully applied, in particular in bioinformatics, to measure regions of similarity in DNA, RNA or protein sequences.

Jaccard Similarity: An abstract can be considered as a bag of words. To incorporate this perspective, we calculate the Jaccard similarity coefficient of a pair of abstracts. We also calculate the Jaccard similarity of sets of effective lemmas of abstract pairs. The former similarity measure shows a very precise matching of the occurrences of words in exactly the same form (singular / plural, noun / adjective / adverb, and so on), while the latter measure considers the existence of words in their canonical forms.

Abstract lengths: Comparing two abstracts from a word-level perspective, the relative length of two abstracts in terms of their words (length of smaller abstracts over the longer one) provides a simple measure of similarity. Although this can be considered as a naive attribute of a pair of abstracts, it has been observed that this measure can be useful when combined with other more powerful measures (Hassanzadeh et al., 2015).

Cosine similarity of effective lemmas: In order to calculate the cosine similarity of the effective lemmas of a pair of abstracts, we map the string vector of the sequence of effective lemmas to its corresponding numerical vector. The numerical vector, with the dimension equal to the number of all unique effective lemmas of both abstracts, contains the frequency of occurrences of each lemma in the pair. For example, for the two sequences $[A, B, A, C, B]$ and $[C, A, D, B, A]$ the numerical vectors of the frequencies of the terms A, B, C and D for the sequences are $[2, 2, 1, 0]$ and $[2, 1, 1, 1]$, respectively. Equation (4) depicts the way the cosine similarity is calculated for two given abstracts A_1 and A_2.

$$Cosine(A_1, A_2) = \frac{V_1.V_2}{||V_1||||V_2||} \tag{4}$$

where V_1 and V_2 are the vector of lemmas of the effective words of two abstracts in a pair, and $V_1.V_2$ denotes the dot product of two vectors which is then divided by the product of their norms (*i.e.* $||V_1||||V_2||$).

Skipped bigram similarities: The set of the skipped bigrams of two abstracts can be used as a basis for similarity computation. We create the skipped bigrams of the effective words and then calculate the intersection of each set of these bigrams with the corresponding set from the other abstract in a pair.

3.2 Combining similarities

In order to assign an overall similarity score to any two given abstracts, the (non-weighted) average of all of the metrics listed above is calculated and is considered as the final similarity score. These metrics compare the abstracts from different perspectives, and hence, the combination of all of them results in a comprehensive quantification of the similarity of abstracts. This averaging technique has been shown to provide good estimation of the similarity of sentences when compared to human assessments both in general English and Biomedical domain corpora (Hassanzadeh et al., 2015).

3.3 Data set preparation and evaluation methods

In order to prepare a realistic testbed, we generated a corpus of PubMed abstracts. The abstracts are retrieved and serialised from the PubMed repository using E-utilities URLs[3]. PubMed is queried by using the 465 medical questions, unmodified, from the EBMSummariser corpus (Mollá and Santiago-martinez, 2011). The maximum number of search results is set to 20,000 (if any) and the results are sorted based on relevance using

[3] `www.ncbi.nlm.nih.gov/books/NBK25497/`

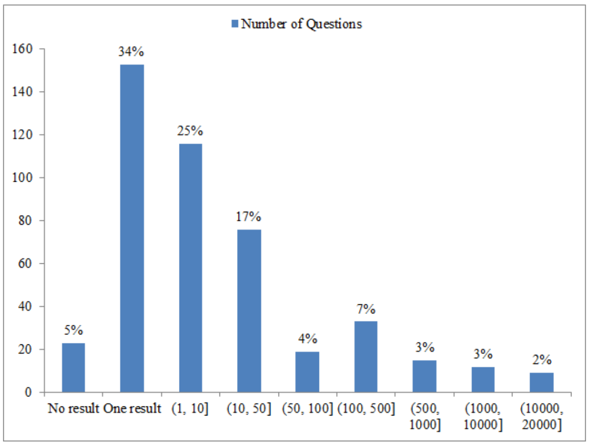

Figure 2: Statistics on the queried questions and their retrieved documents.

PubMed's internal relevance criteria.[4] In total, 212,393 abstracts were retrieved and serialised. The distributions of the retrieved abstracts per question were very imbalanced. There are a considerable number of questions with only one or no results from the PubMed search engine (39% of the questions). Figure 2 shows the frequency of the retrieved results and the number of questions with a given number and/or range of search results.

Some types of published studies may contain better quality of evidence than others, and some, such as opinion studies, provide very little evidence, if any at all. In addition, it is common to have a large number of search results for a given query. Hence, in order to find EBM-related publications as well as to ensure the quality and higher relevance of the abstracts, the retrieved abstracts were filtered based on their publication types. The types of publications are provided in the metadata returned by the PubMed abstracts. To determine the filters, we performed statistical analysis over available corpora in the EBM domain, in particular, EBMsummariser corpus (includes 2,658 abstracts), NICTA-PIBOSO corpus (includes 1,000 abstracts) (Kim et al., 2011), and our retrieved PubMed documents (includes 212,393 abstracts) — more details about the corpora can be found in Malmasi et al. (2015). Table 1 shows the frequency of the most frequent publication types in these EBM corpora. There are 72 different types of publications in PubMed[5], but we limited the retrieved abstracts to the seven more frequently occurring publication types in the EBM domain. Whenever we needed to reduce the number of retrieved abstracts from PubMed search results, we filter the results and only keep the abstracts with the mentioned publication types in Table 1. Note that each PubMed abstract can have more than one publication type. For example, a "Clinical Trial" abstract can also be a "Case Report" and so on. Hence, the sum of the percentages in Table 1 may exceed 100%. We assume that all the documents are informative when the number of returned search results is less than 50, and hence, no filtering was applied in these cases.

After retrieving the documents, in order to be able to evaluate the automatically-generated clusters of retrieved abstracts we devised two scenarios for generating gold standard clusters: Semantic Similarity Mapping and Manual Clustering.

Semantic Similarity Mapping scenario: We generated the gold standard clusters automatically using the cluster information from the EBMSummariser corpus. The answers for each question is known according to this corpus; each answer forms a cluster and citations associated with that answer are assigned to the respective cluster. In order to extend the gold standard to include all the retrieved PubMed abstracts, each abstract was assigned to one of these clusters. To assign an abstract to a cluster, we compute the similarity between the abstract and each of the cited abstracts for the question. To achieve this, we used our proposed combination of similarity measures. The abstract is assigned to the cluster with the highest average similarity. For example, suppose that for a given question there are three clusters of abstracts from the EBMSummariser corpus. By following this scenario, we assign each of the retrieved documents to one of these three clusters. We first calculate the average similarity of a given retrieved document to the documents in the three clusters. The cluster label (*i.e.,* 1, 2, or 3 in our example) for this given retrieved abstract is then adopted from the cluster with which it has the highest average similarity. This process is iterated to assign cluster labels to all the retrieved abstracts. However, it could occur that some clusters may not have any abstracts assigned to them. For the mentioned example, this will result when the retrieved documents would be assigned only to two of the three clusters. When that happens, the question is ignored to avoid a possible bias due to cluster

[4] www.nlm.nih.gov/pubs/techbull/so13/so13_pm_relevance.html

[5] www.ncbi.nlm.nih.gov/books/NBK3827/

Table 1: Statistics over the more common publication types in EBM domain corpora.

Publication Type	EBMSummariser	NICTA-PIBOSO	Retrieved
Clinical Trial	834 (31%)	115 (12%)	12,437 (6%)
Randomized Controlled Trial	763 (29%)	79 (8%)	13,849 (7%)
Review	620 (23%)	220 (22%)	26,162 (12%)
Comparative Study	523 (20%)	159 (16%)	19,521 (9%)
Meta-Analysis	251 (9%)	22 (2%)	2,067 (1%)
Controlled Clinical Trial	61 (2%)	9 (1%)	1,753 (1%)
Case Reports	37 (1%)	82 (8%)	8,599 (4%)

incompleteness. Following this scenario, we were able to create proper clusters for retrieved abstracts of 129 questions out of the initial 465.

Manual Clustering scenario: This scenario is based on the Pooling approach used in the evaluation of Information Retrieval systems (Manning et al., 2008). In this scenario, a subset of the top k retrieved documents is selected for annotation. To select the top k documents we use the above clusters automatically generated by our system. In order to be able to evaluate these automatically generated clusters, for each of them we determine its central document. A document is considered the central document of a cluster if it has the highest average similarity to all other documents in the same cluster. We then select the k documents that are most similar to the central document. The intuition is that if a document is close to the centre of a cluster, it should be a good representation of the cluster and it would less likely be noise. Two annotators (authors of this paper) manually re-clustered the selected top k documents following an annotation guideline. The annotators are not restricted to group the documents to a specific number of clusters (*e.g.*, to the same number of clusters as the EBMSummariser corpus). These manually generated clusters are then used as the gold standard clusters for the Manual Clustering evaluation scenario. The system is then asked to cluster the output of the search engine. Then, the documents from the subset that represents the pool of documents are evaluated against the manually curated clusters. The value of k in our experiment was set to two per cluster. In total, 10 queries (with different numbers of original clusters, from 2 to 5 clusters) were assessed for a total of 62 PubMed abstracts.

3.4 Experimental setup

We employed a Hierarchical Clustering (HC) algorithm in order to cluster the retrieved abstracts (Manning et al., 2008). HC methods construct clusters by recursively partitioning the instances in either a top-down or a bottom-up fashion (Maimon and Rokach, 2005). A hierarchical algorithm, such as Hierarchical Agglomerative Clustering (HAC), can use as input any similarity matrix, and is therefore suitable for our approach in which we calculate the similarity of documents from different perspectives.

As a baseline approach, we use K-means clustering (KM) with the same pre-processing as reported by Shash and Mollá (2013), namely we used the whole XML files output by PubMed and removed punctuation and numerical characters. We then calculated the TF-IDF of the abstracts, normalised each TF-IDF vector by dividing it by its Euclidean norm, and applied K-means clustering over this information. We employed the HC and KM implementations available in the *R* package (R Core Team, 2015).

We use the Rand Index metric to report the performance of the clustering approaches. Rand Index (RI) is a standard measure for comparing clusterings. It measures the percentage of clustering decisions on pairs of documents that are correct (Manning et al., 2008). Eq. 5 depicts the calculation of RI.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \qquad (5)$$

A true positive (*TP*) refers to assigning two similar documents to the same cluster, while a true negative (*TN*) is a decision of assigning two dissimilar documents to different clusters. A false positive (*FP*) occurs when two dissimilar docu-

Table 2: Clustering results over 129 questions of the EBMSummariser corpus.

Method	Rand Index
KM + TF-IDF	0.5261
HC + TF-IDF	0.5242
HC + Similarity Metrics	**0.6036***

* Statistically significant (p-value< 0.05) when compared with second best method.

ments are grouped into the same cluster. A false negative (*FN*) decision assigns two similar documents to different clusters.

4 Experimental Results

In this section, the results from applying our similarity metrics in order to cluster abstracts in the EBM domain are presented. We first introduce our experiments on clustering the abstracts from the EBMSummariser corpus and then we report the results over the retrieved abstracts from PubMed.

4.1 Results on EBMSummariser corpus

In order to evaluate our clustering approach using our similarity metrics, we first employ the EBMSummariser corpus. As previously mentioned, this corpus contains a number of clinical inquiries and their answers. In each of these answers, which are provided by medical experts, one or more citations to published works are provided with their PubMed IDs. We apply our clustering approach to group all the citations mentioned for a question and then compare the system generated clusters with those of the human experts. Table 2 shows the results of using Hierarchical Clustering (HC) and K-means clustering (KM) using the proposed similarity measures and TF-IDF information. In order to have a consistent testbed with our experiments over retrieved documents, the reported results of the corpus are over a subset of the available questions of the EBMSummariser corpus, that is, those 129 questions which were found valid for evaluation in the Semantic similarity mapping scenario in Section 3.3.

Note the improvement of the Rand Index against the TF-IDF methods, *i.e.*, 0.0775. This difference between HC using our similarity metrics and the next best approach, namely KM clustering using TF-IDF, is statistically significant (Wilcoxon signed rank test with continuity correction; p-value = 0.01092).

Our implementation of KM used 100 random starts. It should also be noted that KM can not be used over our similarity metrics, because the final representation of these metrics are the quantification of the similarity of a *pair* of documents and not a representation of a single document (*i.e.*, the appropriate input for KM clustering).

4.2 Results on PubMed documents

As mentioned in Section 3.3, we devised two methods for evaluating the system's generated clusters: the manual scenario, and the semantic similarity mapping scenario. The results of the clustering approach are reported for these two scenarios in Table 3 and Table 4, respectively.

Table 3 shows the results for the manual evaluation. It reports the comparison of the system's results against the manually clustered abstracts from the two annotators. This evaluation scenario shows that, in most cases, the HC approach that employs our similarity metrics produced the best Rand Index. The only exception occurs over the Annotator 1 clusters, where KM using TF-IDF gained better results (*i.e.*, 0.4038 RI). However, for this exception, it is noticed that this difference between the HC approach that uses our similarity metrics and KM using TF-IDF is not statistically significant (p-value=0.5).

Table 3 also shows that the results are similar for two of the three approaches on each annotator, which suggests close agreement among annotators. Note, incidentally, that the annotations were of *clusters*, and not of *labels*, and therefore standard inter-annotator agreements like Cohen's Kappa cannot be computed.

Table 4 shows the results of the methods by using the semantic similarity mapping evaluation approach. It can be observed that, similar to the manual evaluation scenario, HC clustering with the similarity metrics gained the best Rand Index. Finally, although the absolute values of Rand Index are much higher than that from the manual clustering evaluations, the difference between HC on our similarity metrics and the HC and KM methods on TF-IDF information is not statistically significant (p-value=0.1873).

To compare with the results reported in the literature, we computed the weighted mean cluster Entropy for the entire set of 456 questions. Ta-

Table 3: Clustering results over retrieved PubMed documents with Manual Clustering evaluation scenario (Rand Index) for 129 questions from the EBMSummariser corpus.

Methods	**Annotator 1 clusters**	**Annotator 2 clusters**	**Average**
KM + TF-IDF	0.4038	0.3095	0.3566
HC + TF-IDF	0.2877	0.2898	0.2887
HC + Similarity Metrics	0.3825	0.3926	**0.3875**

Table 4: Clustering results over retrieved PubMed documents with Semantic Similarity Mapping evaluation scenario for 129 questions from the EBMSummariser corpus.

Method	**Rand Index**
KM + TF-IDF	0.5481
HC + TF-IDF	0.5463
HC + Similarity Metrics	**0.5912**

Table 5: Clustering results over the entire EBM-Summariser corpus.

Method	**Entropy**
KM + TF-IDF (as in Shash and Mollá (2013))	0.260
KM + TF-IDF (our replication)	0.3959
HC + Similarity metrics	**0.3548***

* Statistically significant (p-value$<$ 0.05) when compared with preceding method.

ble 5 shows our results and the results reported by Shash and Mollá (2013). The entropy generated by the HC system using our similarity metrics was a small improvement (lower entropy values are better) on the KM baseline (our replication of K-means using TF-IDF), which is statistically significant (p-value=0.00276). However, we observe that our KM baseline obtains a higher entropy than that reported in Shash and Mollá (2013), even though our replication would have the same settings as their system. Investigation into the reason for the difference is beyond the scope of this paper.

5 Conclusion

In this paper we have presented a clustering approach for documents retrieved via a set of PubMed searches. Our approach uses hierarchical clustering with a combination of similarity metrics and it reveals a significant improvement over a K-means baseline with TF-IDF reported in the literature (Shash and Mollá, 2013; Ekbal et al., 2013).

We have also proposed two possible ways to evaluate the clustering of documents retrieved by PubMed. In the semantic similarity mapping evaluation, we automatically mapped each retrieved document to a cluster provided by the corpus. In the manual clustering evaluation, we selected the top k documents and manually clustered them to form the annotated clusters.

Our experiments show that using semantic similarity of abstracts can help gain better clusters of related published studies, and hence, can provide an appropriate platform to summarise multiple similar documents. Further research will focus on employing domain-specific concepts in similarity metrics calculation as well as using tailored NLP tools in biomedical domain, such as BioLemmatizer (Liu et al., 2012). Further investigations can also be performed in order to track the effects and contribution of each of the proposed similarity measures on formulating the abstract similarities, and hence, on their clustering. In addition, in order to have more precise quantification of the similarity of abstracts, their sentences can be firstly classified using EBM related scientific artefact modeling approaches (Hassanzadeh et al., 2014). Knowing the types of sentences, the similarity measures can then be narrowed to sentence-level metrics by only comparing sentences of the same type. These investigations can be coupled with the exploration of overlapping clustering methods for allowing the inclusion of a document in several clusters.

Acknowledgments

This research is funded by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) – DE120100508. It is also partially supported by CSIRO Postgraduate Studentship and a visit of the first author to Macquarie University.

References

Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. 2009. A survey of web clustering engines. *ACM Comput. Surv.*, 41(3):17:1–17:38.

Mark H Ebell, Jay Siwek, Barry D Weiss, Steven H Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of Recommendation Taxonomy (SORT): a Patient-Centered Approach to Grading Evidence in the Medical Literature. *The Journal of the American Board of Family Practice / American Board of Family Practice*, 17(1):59–67.

Asif Ekbal, Sriparna Saha, Diego Mollá, and K. Ravikumar. 2013. Multiobjective Optimization for Clustering of Medical Publications. In *Proceedings ALTA 2013*.

P. Ferragina and A. Gulli. 2008. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159–170.

Hamed Hassanzadeh, Tudor Groza, Anthony Nguyen, and Jane Hunter. 2015. A supervised approach to quantifying sentence similarity: With application to evidence based medicine. *PLoS ONE*, 10(6):e0129392, 06.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, 13(Suppl 2):S5.

Jimmy J. Lin and Dina Demner-Fushman. 2007. Semantic clustering of answers to clinical questions. In *AMIA Annual Symposium Proceedings*, volume 33, pages 63–103.

Yongjing Lin, Wenyuan Li, Keke Chen, and Ying Liu. 2007. A Document Clustering and Ranking System for Exploring {MEDLINE} Citations. *Journal of the American Medical Informatics Association*, 14(5):651–661.

Haibin Liu, Tom Christiansen, and William A. Baumgartner Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(3).

Oded Maimon and Lior Rokach. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Shervin Malmasi, Hamed Hassanzadeh, and Mark Dras. 2015. Clinical Information Extraction Using Word Representations. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

George A. Miller. 1995. Wordnet – a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Diego Mollá and Maria Elena Santiago-martinez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*.

Anne L Mounsey and Susan L Henry. 2009. Clinical inquiries. Which treatments work best for hemorrhoids? *The Journal of family practice*, 58(9):492–3, September.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SaraFaisal Shash and Diego Mollá. 2013. Clustering of medical publications for evidence based medicine summarisation. In Niels Peek, Roque Marn Morales, and Mor Peleg, editors, *Artificial Intelligence in Medicine*, volume 7885 of *Lecture Notes in Computer Science*, pages 305–309. Springer Berlin Heidelberg.

Oren Zamir and Oren Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 46–54. ACM.