MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Distilling Conceptual Connections from MeSH Co-Occurrences

Padmini Srinivasan[a], Dimitar Hristovski[b]

[a]*School of Library and Information Science, The University of Iowa, Iowa City, USA*
[b]*Institute of Biomedical Informatics, Medical Faculty, University of Ljubljana, Slovenia*
{padmini-srinivasan@uiowa.edu, dmitar.hristovski@mf.uni-lj.si}

Abstract

Our aim is to contribute to biomedical text extraction and mining research. In this paper we present exploratory research on the MeSH terms assigned to MEDLINE citations. We analyze MeSH based co-occurrences and identify the interesting ones, i.e., those that are likely to be semantically meaningful. For each selected co-occurring pair we derive a weighted vector representation that emphasizes the verb based functional aspects of the underlying semantics. Preliminary experiments exploring the potential value of these vectors gave us very good results. The larger goal of this project is to contribute to knowledge discovery research by mining the knowledge that is latent within the biomedical literature. It is also to provide a method capable of suggesting cross-disciplinary connections via the pairs derived from all of MEDLINE.

Keywords:

text mining, MEDLINE, MeSH, semantic relation extraction

Introduction

The MEDLINE database with over 11 million records is the primary access point to the literature covering biomedical research. Together with allied databases such as GenBank and SwissProt and also its associated knowledge structures: MeSH and the UMLS, MEDLINE supports the promise of text extraction and mining algorithms. Indeed as seen in recent research [1], [2], [3], [4] and [5], MEDLINE has been a key part of the substratum of numerous extraction experiments. Thus we see active research on algorithms to automatically identify objects such as genes, proteins and drugs in texts as well as their interactions. We can also observe steady progress with more complex entities such as metabolic and biochemical pathways, e.g. [1].

Our goal in this paper is to contribute to biomedical text extraction and mining research by exploring the MeSH indexing associated with MEDLINE records. We propose a method to identify the meaningful MeSH based co-occurrences. In other words we seek to filter out the coincidental pairings and retain only those that have semantic value. Our aim is also to develop a reasonable representation of each pair. The larger goal of this project is to contribute to knowledge discovery research by mining the knowledge that is latent within the biomedical literature. It is also to provide a method capable of suggesting cross-disciplinary connections via the pairs derived from all of MEDLINE.

MeSH Co-occurrences

Each MEDLINE record is manually assigned around 12 MeSH terms by trained indexers. MeSH terms, also known as MeSH concepts or MeSH headings, may stand alone or be qualified by one or more subheadings, which are also known as qualifiers. For example the MeSH concept *Hypertension* may appear by itself in a document. Alternatively it may appear as *Hypertension/treatment* and *Hypertension/etiology*. Thus although all of these documents are about hypertension, the particular aspects examined in the documents are recognizably different. Together the MeSH concepts and subheadings offer a powerful indexing tool.

A natural strategy would be to take every occurrence of a MeSH heading/subheading combination as a single unit and study co-occurrences between pairs of such units. However, there are two problems with this approach. The first is the issue of scale. The 20,742 concepts and 82 subheadings in the 2002 MeSH vocabulary, allow for a maximum of 1,700,844 concept/subheading combinations which yields almost 1.5 trillion possible pairs (although there are actually much fewer allowable combinations). A second problem, which is really the more critical one for us, is that this approach is necessarily domain dependent. Relationships between particular MeSH concepts depend upon the specifics of the discipline. In contrast we are interested in semantic relationships that stay resilient across disciplines. The difference becomes clear if we compare the pairs: *diabetes/drug therapy - insulin/therapeutic use* and *disease/drug therapy - chemical/therapeutic use*. Although the subheadings are identical, the pairs differ in their level of abstraction and their degree of domain dependency. Co-occurrences of the latter type are of interest to us. Practically speaking the higher level of abstraction allows our statistical methods to operate against the full MEDLINE database.

Fortunately the UMLS classifies each MeSH concept by semantic type. Also there are only 134 semantic types, arranged as a hierarchy. Examples include *Acquired Abnormality*, *Cell Component*, and *Finding*. Each Metathesaurus concept is assigned at least one semantic type that is the most specific semantic type available in the type hierarchy.

Thus we transform each MeSH concept/subheading into a semantic type/subheading and study co-occurrences between pairs of such units. While the semantic types represent broad classes, the subheadings sharpen their semantic interpretation. Consider for example, the semantic type *Disease or Syndrome* by itself

and also *Disease or Syndrome/drug therapy* and *Disease or Syndrome/epidemiology*.

With 134 semantic types and 82 subheadings, there are a maximum of 10,988 semantic type/subheading combinations[1]. This gives us a maximum of more than 60 million possible pairs. Fortunately several combinations are not meaningful and thus will not appear. For example, the subheading *biosynthesis* applies only to chemicals. Thus we observe that only 983,784 pairs occur in the full MEDLINE database.

Methods

A standard approach for extracting interaction information is to first identify the relevant verbs and then extract their arguments from the text. For instance when extracting protein - protein interactions, verbs such as *interact*, *associate* and *bind* offer reasonable starting points [6]. Typically these interesting verbs are either hand picked from the text or independently identified by domain specialists. However, this method has serious limitations when extracting interactions based on MeSH co-occurrences. First, there is the problem of scale. Even if only 1% of the 983,784 observed co-occurrences turn out to be interesting enough to study, the set is too large to manually identify key verbs. More importantly, even when limited to a single co-occurrence, it would be very difficult to predict the verbs in the documents that typify the relationship of interest. Thus our overall approach has 2 phases. First, we automatically identify the key verbs associated with a pair. Next, we use these verbs to extract closely connected nouns and noun phrases. This paper focuses on the first phase. In particular we propose a method for automatically extracting a verb-based profile for each MeSH pair. Our plan is to focus on the noun phrases associated with these key verbs in follow-up research.

Our process begins by identifying all pairs of semantic type/subheading units that co-occur in the MEDLINE records. We then examine each to determine if it is statistically interesting and limit further study to such pairs. Our next step is to obtain a representation for each retained pair. A common strategy in information retrieval research is to represent a document (or set of documents) with a vector consisting of every nontrivial word in it. Several alternative term weighting strategies such as inverse document frequency and discrimination value may then be used to estimate the relative importance of the terms in the vector.

As motivated previously we are interested in deriving a weighted verb vector that is a reasonable representation of the functional relationship between the pair being analysed. We then test the value of this verb vector by assessing its average similarity to two sets of documents. The first is a set of documents in which the pair occurs and the second is a set of documents in which the pair does not occur. If the two similarities are not significantly different, then it indicates that the generated verb profile has little merit.

1. However, this is far fewer than the number of possible concept/subheading combinations.

Extraction of Semantic Type/Subheading Pairs

We started by processing the full MEDLINE distribution till the end of 2001 that contains about 11 million records. As the distribution is in XML format, we first extracted the relevant fields and transformed them into a relational text format (one line for each MeSH concept/qualifier occurrence in each record). Next we mapped the MeSH concepts into their corresponding semantic types using the UMLS (2002 edition). From this we built the co-occurrence file for each pair (S1/Q1, S2/Q2) where S1 and S2 are semantic types and Q1 and Q2 qualifiers. A total of 983,784 co-occurring pairs were identified and their co-occurrence frequencies noted. As expected most pairs occur very few times. In fact 30% (288,945) occur only once and 97% (952,697) occur 500 or fewer times.

As this is an exploratory study, we decided to limit our analysis to a subset of the 983,784 co-occurring pairs. The following selection criteria were used: (1) The pair should co-occur in at least 500 documents. This gives us a reasonable number of documents to use for the next steps. (2) The observed co-occurrence frequency should be greater than the expected number by at least 25%.

The first criterion left us with 31,087 pairs. For the second criterion we assume that the two members of the pair are independent of each other and then estimate their probability of co-occurrence. Thus for a pair A-B if A occurs in $n1$ documents in a database of size N and B occurs in $n2$ documents, then their co-occurrence probability is

$$\left(\frac{n_1}{N} \times \frac{n_2}{N}\right)$$

and their expected co-occurrence frequency is .

$$\left(\frac{n_1 \times n_2}{N}\right)$$

The deviation calculated for criterion 2 is then

$$\left(\frac{\exp(ectedfrequency - observedfrequency)}{\exp ectedfrequency} \times 100\right)$$

This value should be greater than 25%. The application of this criterion left us with 21,975 pairs. Out of these, we randomly selected 250 pairs for further analysis. Unfortunately, due to unforeseen problems, we were able to download documents reliably for only 228 pairs. Our further analysis is based on these 228 pairs.

Background Dataset

Our statistical procedure requires a background dataset. Thus we extracted a random set of 100,000 records from the 11 million MEDLINE records. The title and abstract of each document is tagged using the Brill part of speech tagger [7]. Words tagged as verbs are extracted and used to generate a vector. For each unique verb in the set we calculate the inverse document frequency (IDF) as: $\log_2 N/n$

where n is the number of documents in which the verb occurs and N is the number of documents in the set (100,000). This vector of IDF weighted verbs is referred to henceforth as the background vector or background profile (BV).

Given variations in verb forms due to differences in tense we use the UMLS to transform them to the infinitive form before calculating weights. Thus for example, variants such as *increasing*, *increased* and *increases* are all transformed into instances of *increase* before frequencies are counted. This transformation is done for all the verb vectors derived in this study.

Analysis for a Single Pair

We first identify all the documents in which the pair occurs. This set is randomly split into two portions: two-thirds are placed in a training set while the remaining one-third in combination with an equal number of randomly selected documents from the rest of MEDLINE forms the test set. Thus the training and test set sizes depend upon the number of documents in which the pair occurs.

Creation of Verb Profile Representing the Pair

The documents (titles and abstracts) in the training set of each pair are processed by the Brill tagger and the verbs are extracted. The criteria for inclusion of a verb in the profile vector are: (1) it should occur in at least 5 documents and (2) if the verb occurs in the background profile BV, then its occurrence in the dataset for the pair should be significantly different from that in the background data. We test this by computing theχ^2 statistic using the frequencies shown in Table 1.

Table 1. χ^2 Statistic

	Verb Present	Verb Absent
Observed Number of documents	*o*1	*o*2
Expected Number of documents as per background distribution	*e*1	*e*2

If the χ^2 statistic is greater than 3.84[1], then we can be about 95% percent confident that the difference between the observed and expected pattern of frequencies for the verb does not result from mere random variability. The occurrence of the verb in the documents corresponding to the pair is then significantly different from the occurrence in the background dataset.

We include such verbs in the profile vector (PV). Again we calculate a weight for each term denoting its relative importance. For a particular verb *t*, the alternative weights computed are:

1. augmented term frequency (AugTF),

$$w_{AugTF}(t) = \frac{TF_t}{\sum_{k \in V} TF_k}$$

where TF_j is the number of times verb j occurs in the dataset and V is the set of verbs in the vocabulary.

2. augmented term frequency * IDF (AugTFIDF),

$$w_{AugTFIDF}(t)) = w_{AugTF}(t) \times IDF_t$$

3. term frequency * IDF (TFIDF),

$$w_{TFIDF}(t) = TF_t \times IDF_t$$

4. IDF,

$$w_{IDF}(t) = IDF_t$$

In other words we generate four different profile vectors for each pair that differ only in the weights associated with the component verbs. Our intent is to compare these alternative weighting schemes. It should be noted that IDF weights in each case refer to the weights computed from the background dataset.

Testing the Verb Profiles

We test the value of each derived verb profile using the test sets described in section 3. It may be observed that half of each test set contains the pair being examined, while the other half does not. In essence we extract the verbs from each test document and create a vector using the AugTF scheme (described before). Thus each document is represented by a weighted vector of its verbs. We then compute similarity between a document verb vector (D) and a verb profile vector (PV) as their dot product:

$$\left(Similarity(D, PV) = \sum_{k \in V} (wt_{D_k}, wt_{p_k}) \right) \quad (1)$$

where *V* is the universe of verbs[2], Wt_{Dk} and wt_{PVk} are the weights of verb *k* in D and PV respectively.

We then compute the mean similarity for each of the two subsets of documents in the test set. The first consists of documents in which the pair occurs. The second set contains documents in which the pair does not occur. The idea is that if these similarities are too close, then the verb profile is not sufficiently interesting for further study. If the means are significantly different, then the profile has potential and will be explored further.

Results

Verb Profiles

A total of 7,689 verbs appear in the background dataset. Almost 60% of these occur in only 1 document while 97% of them occur in 500 or fewer documents. We use a small stoplist containing: *was*, *were*, *is*, *have*, *be*, *are*, *do*, *been*, and *use*. These were the highest frequency verbs in the background vector. There were an average of 140 verbs per profile with a standard deviation of 100. The highest is 956 while the lowest 16.

Table 2 presents the top five verbs as ranked by the *AugTFIDF* scheme for eight pairs that were chosen at random. There are interesting similarities and differences between the lists. For instance *tolerate* occurs only in the first pair, whereas *upregulate* occurs in three of them. On the whole these top ranked verbs appear reasonable, although some very trivial verbs such as *show* and *utilize* tend to appear quite often, suggesting that we may need to enlarge our stopword list. Also, rather unusual verbs such as *well* indicate problems with the tagging.

Following are a few example sentences from the document set corresponding to the pair *Disease or Syndrome/drug therapy* and

1. 3.84 is the critical value for df = 1 at 0.05 level of significance

2. Note that the product of weights will be zero for any verb that is not in the overlap between D and PV.

Lipid/adverse effects that have the verb *withdrawal* in them, along with their relevant MeSH terms. The first sentence states an adverse consequence of a withdrawal. The second poses a question on the consequences (adverse or otherwise) of the withdrawal of some drugs. The third shows the benefits of the withdrawal of a substance in terms of the reduction of adverse effects. In phase 2 of our research we will address the challenge of extracting the key noun(s) and noun phrases associated with these verbs from such sentences. It is acknowledged that this is a challenging problem. However, the advantage of the verb profiles is that they provide important constraints to this second phase extraction problem.

Table 1: Sample of Top 5 Verbs from Profiles

Pair	Verbs
disease or synd./drug & lipid/adv. effects	withdraw, warrant, undertake, treat, tolerate
Pharm. Subs./diag & Neoplastic Proc./diag	visualize, undergo, sustain, show
Bacterium/metab & Genetic Func./drug	utilize, trigger, transform synthesize, suppress
Tissue/cytol & Immun. Factor/pharm	well, upregulate, understand, trigger, treat
Cell/immunol & Genome/immunol	well, vacinate, utilize, understand, trigger
Cell/drug & Genetic Func./genet	utlize, upregulate, understand, truncate, trigger
Gene/genet & Peptide or Protein/chem.	yield, wound, utilize upregulate, express
Cell Func./rad & Neoplastic Proc./path	yeild, warrant, vary utilize, underlie

1. The onset of NCSE was temporally related to the withdrawal of sodium valproate
   Status Epilepticus/drug therapy
   Valproic Acid/adverse effects
2. Does withdrawal of different antiepileptic drugs have different effects on seizure recurrence?
   Epilepsy/drug therapy
   Anticonvulsants/adverse effects
3. A complete resolution of the hematologic damage was observed after valproate withdrawal
   Anticonvulsants/adverse effects
   Epilepsy, Tonic-Clonic/drug therapy
   Valproic Acid/adverse effects

Test Set Results

Figure 1 shows the results on the test datasets for the augmented TF weighting scheme. Each data point in the graph represents a pair. The value on the x-axis represents mean similarity in the topical subset (the subset in which the pair occurs) and the y-axis value is the mean similarity for the random subset. If these two means are identical then the data point would fall on the diagonal. Distance away from the diagonal indicates greater differences. The graph also shows standard error bars for the means. Thus each data point is actually an intersection of two lines where the lines, representing the standard errors on the 2 axes, specify the edges of a box. If any region of a box intersects with the diagonal, then the mean similarities for that pair are not statistically significant. It may be observed that all pairs show significant differences in mean similarity.

In summary, for all cases, the verb profile generates significantly different similarity scores with topical documents than with random documents. Moreover, given that the points are below the diagonal, these similarities are higher for the topical subset. These results give us confidence that the verb profiles are in the right direction. Thus we have now set the stage for the next phase in our research. We will now study methods that may be used to extract the key nouns that are associated with the verbs in the profiles. We will also explore the use of the NLM MetaMap program to map the noun phrases to UMLS concepts.

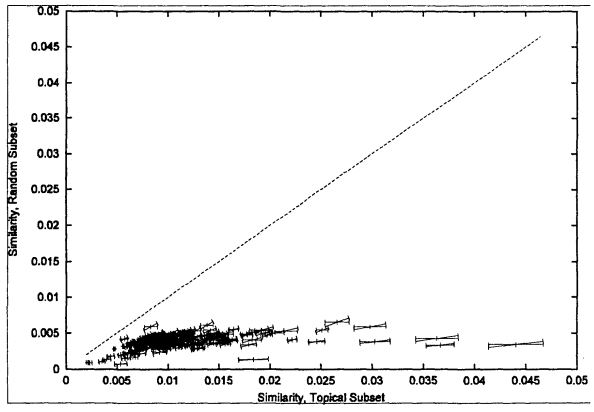

Figure 1 - Augmented TF.

Related Research

We provide a brief review of two key dimensions related to our work: (1) the importance of verbs and (2) research on MeSH co-occurrences with emphasis on the subheadings.

It is well acknowledged that verbs have a central role in expressing interactions between objects. Therefore, it is of no surprise to observe that research on the extraction of interactions such as protein-protein interactions pay close attention to verbs. For example, in [4] text fragments containing a verb (from a fixed list of 14 pre-specified verbs indicating actions related to protein interactions) and at least two protein names are analysed. Special rules considering word order and the verb are used to extract occurrences of protein interactions. [6] hand analyse the occurrence of about 30 verbs in 200 MEDLINE abstracts. Based on their analysis they decided to focus on 3 verbs: interact (with), associate (with) and bind (to) that indicate protein interactions with high confidence.

A key paper that is centered on the role of verbs is [8] in which the authors examine several of the most frequently seen verbs in a MEDLINE subset corresponding to certain nuclear proteins. Frequent verbs such as activate, bind, and interact were found to vary in the distribution of their conjugational forms. The authors identify the subject and object noun phrases for the verbs using an algorithm that applies different syntactic rules depending on whether the verb is in active or passive voice.

The research that is most closely related to ours is [9], [10]. In [9] the authors map the MeSH concepts into a set of 5 classes (such as Chemicals and Drugs; Diseases and Procedures). They then examine the semantics underlying co-occurring pairs of semantic class/subheading units. In [10] they map the MeSH concepts to the UMLS semantic types and examine co-occurring semantic types. Their motivation is similar to ours which is to build knowledge bases automatically from the semantic relationships that lie latent in the MeSH indexing. In both papers the authors limit their analysis to a particular discipline - cardiovascular diseases. In contrast our aim is to remain domain independent and propose semantic relations that are impervious to specialties. More recently Hatzivassiloglou and Weng also automatically extract verbs, but those relevant to gene and protein interaction [11].

Other related research includes [12] presenting a prototype, MeSHmap where the documents retrieved in response to a PubMed search are analysed for their MeSH concepts and subheadings. Concepts qualified by the same subheading are grouped and the system displays the distribution of MeSH concepts by subheading as a summary of the retrieved set. MeSHmap also supports the comparison of pairs of objects (such as diseases) each represented by its own PubMed search. This is done by comparing the distributions of concepts over the different subheadings that appear across the retrieved sets of documents. In [13] the authors discover new potential relations between MeSH major headings by analyzing and combining known relations. They use binary association rules as a knowledge extraction mechanism. The association rules are based on the co-occurrence of the MeSH headings.

Conclusions

We presented a method for identifying the interesting MeSH based co-occurrences from MEDLINE. Each co-occurring pair is represented by a profile vector composed of verbs depicting the functional aspects of the underlying semantics. These verbs are selected based on their pattern of frequencies in the subset for the pair and in a background dataset. The resultant verb vectors give good performance in their ability to characterize documents in which the co-occurrence is present from the other documents. In future work will focus on the next phase of our research which is to extract the nouns from the individual documents that appear related to each of the key verbs in the profile. We expect the verb - noun combinations to give a more complete representation of the MeSH co-occurrences. We will also test our methods on a larger subset of the co-occurrences.

Acknowledgments

This research was conducted while both authors were visiting scholars at the National Library of Medicine. The authors thank NLM and the ORISE program for support. In addition the first author thanks the University of Iowa for the Faculty Scholar Award (2001-2003) that also supported her research stay at NLM.

References

[1] Liu H, Friedman C: Mining Terminological Knowledge in Large Biomedical Corpora. *PSB* 2003; 415-426.

[2] Shah PK, Perez-Iratxeta C, Bork P, Andrade MA. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* 2003, 4:20

[3] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *PSB* 2000; 529-40.

[4] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. *ISMB* 1999; 60-7.

[5] Shatkay H, Edwards S, Wilbur WJ, Boguski M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *ISMB* 2000; 8:317-28.

[6] Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. *PSB* 2000; 541-52.

[7] Brill E. Some advances in transformation-based part of speech tagging. *Proceedings of the National Conference on Artificial Intelligence* 1994; 722-7.

[8] Sekimizu T, Park HS, Tsujii J. Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. *Genome Inform Ser Workshop Genome Inform* 1998; 9:62-71.

[9] Cimino JJ, Barnett GO. Automatic knowledge acquisition from MEDLINE. *Methods Inf Med* 1993; 32(2):120-30.

[10]Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. *AMIA Symp* 2000; 575-9.

[11]Hatzivassiloglou V, Weng W. Learning anchor verbs for biological interaction patterns from published text articles. Workshop on Natural Language *Process in Biomedical Applications* 2002.

[12]Srinivasan P. MeSHmap: a text mining tool for MEDLINE. *AMIA Symp* 2001; 642-6.

[13]Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Medinfo 2001*; 10(Pt 2):1344-8.

Address for correspondence

Email: padmini-srinivasan@uiowa.edu. Phone: 319-335-5707; Fax: 319-335-5374.