Knowledge Extraction from MEDLINE by Combining Clustering with Natural Language Processing

Jose A. Miñarro-Giménez, PhD, Markus Kreuzthaler, MSc, Stefan Schulz, MD
Institute of Medical Informatics, Statistics, and Documentation,
Medical University of Graz, Austria

Abstract

The identification of relevant predicates between co-occurring concepts in scientific literature databases like MEDLINE is crucial for using these sources for knowledge extraction, in order to obtain meaningful biomedical predications as subject-predicate-object triples. We consider the manually assigned MeSH indexing terms (main headings and subheadings) in MEDLINE records as a rich resource for extracting a broad range of domain knowledge. In this paper, we explore the combination of a clustering method for co-occurring concepts based on their related MeSH subheadings in MEDLINE with the use of SemRep, a natural language processing engine, which extracts predications from free text documents. As a result, we generated sets of clusters of co-occurring concepts and identified the most significant predicates for each cluster. The association of such predicates with the co-occurrences of the resulting clusters produces the list of predications, which were checked for relevance.

Introduction

The state of the art of a scientific discipline and the progress of investigations about a particular topic are described by their set of scientific publications. The identification of new knowledge, which in the past exclusively relied on human effort, has become increasingly difficult due to the accelerating growth in academic literature. Consequently, scholars rely more and more on machine filtering and preprocessing of scientific articles[1].

The MEDLINE[2] database, with currently about 22 million bibliographic records, is the most important source of biomedical literature. Each record has been semantically annotated by experts of the U.S. National Library of Medicine using the MeSH thesaurus[3]. These metadata are not only useful for document retrieval, but also constitute valuable assets for information and knowledge extraction. They include sets of **MeSH terms**, which can be further qualified by one or more **MeSH subheadings** (Table 1), which specify the semantic context of the MeSH term, e.g. Anti-Inflammatory Agents / Therapeutic use, or Nephrotic Syndrome / Drug therapy. As much as this is useful for targeted document retrieval, it falls short of typical knowledge representation use cases, which demand predications like < Anti-Inflammatory Agents — Treats — Nephrotic Syndrome >.

The file MRCOC[4] within the UMLS Metathesaurus[5, 6] contains all annotations of co-occurring MeSH terms and subheading qualifiers for each MEDLINE record (aka citation). In the past, the content of MRCOC was used for diverse objectives, such as knowledge extraction[7], the identification of associative relations[8], semantic relation discovery[9], mining of symbolic and statistical gene-disease relationships[10], and text mining in general[11].

Another UMLS component is the Semantic Network[12] (UMLS SN), an informal upper-level ontology, which provides 133 generic categories, so-called semantic types, linked by 54 directional relationships. All Metathesaurus concepts are assigned to one or more SN semantic types. Table 2 shows how the semantic types Disease or Syndrome, Organism Function and Pharmacologic Substance are linked in UMLS SN, which, in addition, defines directional relationships such as < Disease or Syndrome — Manifestation_of — Organism Function > or < Pharmacologic Substance — Disrupts — Organism Function >. They are pattern for typical predications, with their subject and object positions to be refined by UMLS Metathesaurus concepts.

The identification of relevant semantic relations is crucial for the generation of predications[13, 14]. Natural language processing (NLP) is the method of choice for extracting such predications from textual sources[15]. One example is SemRep[16], a system that recovers predications from biomedical text using syntactic analysis and structured domain knowledge from UMLS. However, ambiguity and complexity of biomedical language hinder the accurate extraction of biomedical facts. Capitalizing on the availability of semantic explicit MeSH annotations, we will investigate how this resource can be used to extract factual statements from MEDLINE, despite the lack of relational predicates in its metadata annotations.

Table 1. Excerpt of the list of MeSH terms and their related MeSH Subheadings in the sample PubMed record "Childhood nephrotic syndrome – Current and future therapies" (id = 22688744).

MeSH term and identifier	MeSH Subheadings and abbreviations
Nephrotic Syndrome (D009404)	Therapy (TH), Metabolism (ME)
Anti-inflammatory Agents (D000893)	Therapeutic use (TU)
Immunosuppressive Agents (D007166)	Therapeutic use (TU)
MAP Kinase Signaling System (D020935)	Drug effect (DE)
Plasmapheresis (D010956)	
Interleukin-13 (D018793)	Antagonists & inhibitors (AI)

Table 2. List of UMLS SN relationships between the semantic types Disease/Syndrome, Organism Function and Pharmacologic Substance.

Subject/ Object	Disease or Syndrome	Organism Function	Pharmacologic Substance
Disease or Syndrome	Associated_with, Co-occurs_with, Result_of, Degree_of, Process_of, Manifestation_of, Precedes, Affects, Occurs_in, Complicates	Result_of, Process_of, Manifestation_of, Affects	
Organism Function	Process_of, Results_of, Affects	Co-occurs_with, Result_of, Degree_of, Process_of, Precedes, Affects	
Pharmacologic Substance	Diagnoses, Treats, Complicates, Affects, Prevents, Causes	Complicates, Disrupts, Affects	Interacts_with

Table 3. Simplified sample record from the UMLS MRCOC file, containing the following fields: PubMed Unique Identifier (PMID); the dates related to the publication of the paper and its related MeSH indexing year; whether both MeSH terms are the main topics in the publication (ZY) or not (ZN); a description of the first MeSH term that consists of the unique identifier for MeSH heading term (MeSH DUI), the corresponding UMLS concept unique identifier (UMLS CUI) and the list of comma-separated MeSH subheadings that qualify the MeSH term; and an analogous description of the second MeSH term.

PMID	Earliest year, pub date, article date, date completed, indexing year	Major Topics	MeSH Descriptor 1			MeSH Descriptor 2		
			MeSH DUI	UMLS CUI	List of Qualifiers	MeSH DUI	UMLS CUI	List of Qualifiers
20278133	19461001|19461001|0|20100318|2010	ZY	D001808	C0005847	AB:Q000002	D006225	C0018563	BS:Q000098

This will be addressed by analyzing, in parallel, (i) patterns of MeSH term / subheading co-occurrence as extracted from MEDLINE metadata, and (ii) the text of the paper abstracts, using SemRep. In both cases, the target representations are triplet-based predications, using UMLS SN relations. Based on the frequency of natural language predicates in the abstracts, we attempt to infer relations that interpret the statistical associations regarding the distribution of 83 subheading types in MeSH annotations.

Rather than formal-ontological relations, we expect, primarily, to extract associative relations between the subject and the object concept, which express what is typical or probabilistic. We are not interested in statements like < Nephrotic Syndrome — Is_a — Kidney Disease > which are already extensively covered by ontologies like SNOMED CT or OBO ("all nephrotic syndromes are kidney diseases"). Instead, we focus on triples like < Corticosteroids — Treats — Nephrotic Syndrome >, which represent contingent or probabilistic ("non-ontological") knowledge, which cannot be translated into first order logics[17] and is not expected to be found in domain ontologies. However, this is exactly the kind of content we consider more "interesting" from a biomedical knowledge representation point of view[18].

We limited the scope of our experiment dataset to the MEDLINE records published in the last 5 years (2009 - 2013). Besides, we focused on the most relevant MeSH terms that (i) co-occur in the MEDLINE records and (ii) are linked to the UMLS SN types Disease/Syndrome and Pharmacologic Substance. Despite of such limitations, the amount of data involved is still huge and therefore requires efficient and scalable methods typical for big data analytics[19], including a distributed framework that supports multithreaded, massively parallelized computing tasks.

Material and Methods

Resources: The main resource is the UMLS MRCOC file, which provides information about pairs of MeSH terms that co-occur in each MEDLINE record related to a specific MeSH indexing year. We used the 2014AA release of the detailed MRCOC version with a size of 131 GB. Each line of this file (Table 3) contains, among other data, a pair of co-occurring MeSH terms, the subheadings assigned to them, and the identifier of the MEDLINE record where these MeSH terms co-occur. The number of subheadings is variable, depending on the content of the article. Each MeSH term points to one UMLS unique concept identifier (CUI), which is linked to one or more UMLS semantic types in the files MRSTY and SRSTRE1.

Tools: SemRep is a system that analyses text and extracts semantic predications as triples. Their subjects and objects are represented as UMLS CUIs and the predicates correspond to one of the 54 UMLS SN relations. The output of SemRep is classified into three categories: TEXT, ENTITY and RELATION. TEXT describes the textual section analyzed by the system. ENTITY is related to a particular UMLS concept identified in the text, and RELATION describes a semantic relation between two entities found in the text.

Methodology: The creation of predications from MEDLINE metadata consists of five main phases: (i) aggregation of the co-occurring concepts and their MeSH subheading information; (ii) filtering of the aggregated co-occurrences based on their log-likelihood rates[20] (LLRs); (iii) clustering of the co-occurring concepts based on the accumulated MeSH subheadings; (iv) extraction of semantic predications from MEDLINE records associated with the co-occurrences of every resulting cluster using SemRep; and (v) identification of the statistically significant predicates of each cluster.

- The **first phase** is the aggregation of co-occurring concept pairs in order to select the most relevant ones. The input data is the list of co-occurrences from MRCOC described in Table 3, and the output will be the list of aggregated PubMed identifiers, the MeSH and UMLS IDs of the first term with a list of aggregated MeSH subheadings, the MeSH and UMLS IDs of the second term with a list of aggregated MeSH subheadings (Figure 1). Due to the large data volume in the detailed version of MRCOC, we have applied the MapReduce[21] programming paradigm and the Amazon cloud services (Amazon EMR[22] and Amazon S3[23]) together with Apache Hadoop[24] for generating the aggregated version of MRCOC. MapReduce provides two types of procedures: MAP and REDUCE. The functionality of MAP is to filter and sort elements, where REDUCE takes the output of MAP and performs certain operations on the processed values. The data is represented as key/values pairs, thus facilitating data access and distribution across several computers. Hadoop is a software framework that supports the distributed processing of large data sets across clusters of computers. Consequently, processing the MRCOC file using both, MapReduce and Hadoop, can be easily scaled up and parallelized.

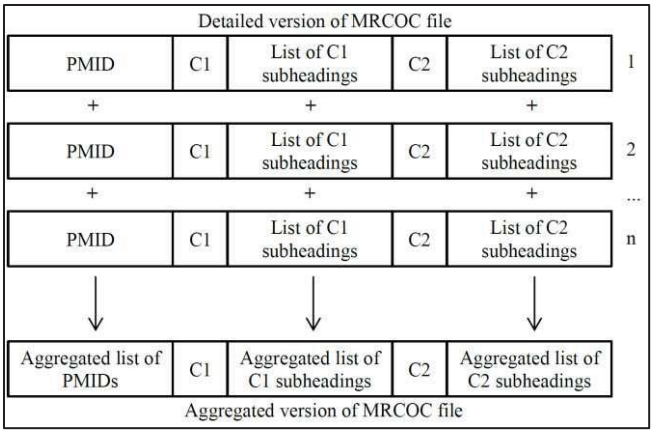

Figure 1. Depiction of the aggregation of co-occurring concept pairs from the MRCOC file.

- The **second phase** of our methodology is the calculation of log likelihood ratios (LLRs) and the co-occurrence filtering based on such rates. LLR reflects the statistical relevance of a pair of concepts regarding their percentage in the dataset and the percentage of other concept pairs. The required parameters to calculate the LLR of concept C1 that co-occurs with concept C2 are:

  1. Number of co-occurrences of C1 and C2 (#C1∩C2);
  2. Number of co-occurrences of C1 without C2 (#C1∩¬C2);
  3. Number of occurrences C2 without C1 (#¬C1∩C2);
  4. Number of co-occurrences where neither C1 nor C2 occur (#¬C1∩¬C2).

  These parameters where obtained during the aggregation of list of MeSH subheadings using Amazon cloud services. The formula in Figure 2 explains how the log-likelihood ratio is computed. The function H represents the Shannon entropy[25]. H(matrix) indicates the entropy of #C1∩C2, #C1∩¬C2, #¬C1∩C2 and #¬C1∩¬C2. H(mRows) is the sum of entropies of the pairs <#C1∩C2, #¬C1∩C2> and <#C1∩¬C2, #¬C1∩¬C2>. Finally, H(mCols) is the sum of entropies of the pairs <#C1∩C2, #C1∩¬C2> and <#¬C1∩C2, #¬C1∩¬C2>. In order to filter the less significant co-occurrences we applied a threshold of 10.83, which corresponds to the chi-squared test with one degree of freedom and a p-value < 0.001.

$$LLR = 2 \times \left(H(\text{matrix}) - H(\text{mRows}) - H(\text{mCols}) \right)$$

Figure 2. Formula of the Log-likelihood ratio (LLR) using Shannon entropy.

- The **third phase** is the co-occurrence clustering. The resulting list of co-occurrences from the previous phase is firstly, filtered by the semantic types of both concepts and, then, clustering them by the list of aggregated MeSH subheadings. Thus, we obtain different sets of co-occurrences for each pair of semantic types, such as Disease or Syndrome with Organism Function, or Pharmacologic Substance with Pharmacologic Substance. We hypothesize that each pair of semantic types can be interpreted as one or more UMLS SN relations. Their number can be used as input parameter to define the number of clusters to obtain during the clustering process. We use WEKA[26], which provides a set of clustering and machine learning algorithms, which can be applied to data mining tasks, from which we chose the k-means clustering algorithm. The variables to be used to cluster the co-occurrences are the list of aggregated MeSH subheadings. The number of occurrences of each subheading in the list is normalized by the total number of co-occurrences found for its related concept pair, and, therefore, the value for each subheading will be in the range [0, 1]. As there are 83 types of MeSH subheadings and because each concept in the co-occurrence has a different list of subheadings, the total number of features is 166.
  In our experiments, we have focused on the combination of the semantic types Disease or Syndrome with Pharmacologic Substance, for which UMLS SN defines six predicates: Diagnoses, Affects, Treats,

Complicates, Prevents and Causes (Table 2). The number of suggested predicates provided by UMLS SN can be used as the input parameter of the number of clusters to be produced by the k-means algorithm. Moreover, we can also investigate the use of clustering algorithms that can estimate the best number of clusters.

- In the **fourth phase** of the methodology, predications are extracted from MEDLINE abstracts using SemRep. However, the accuracy of such predications is limited by the ambiguity and complexity of natural language. Nevertheless, we can use the outcome of this analysis as indicators for the most relevant predicates for each co-occurring concept pair. Subject to analysis are the abstracts of those MEDLINE records that are related to the co-occurrences of the resulting clusters from the third phase. This phase yielded a list of predicates directly related to each pair of co-occurring concepts.

- The goal of the final **fifth phase** is the association of the most relevant predicate to each resulting cluster. To this end, we analyze the predications generated by SemRep. The analysis is focused on the study of the frequencies of the predicates in each cluster. According to our hypothesis, the concept co-occurrences grouped together in a cluster are related with the same type of predicate. Therefore, the predicate frequencies found by SemRep for those co-occurrences should be also statistically significant.

  We compared the results between clustering into five, six, and seven different clusters with the k-means clustering algorithm. The assignation of the relevant predicates to each resulting cluster must be consistent with the predominant types of MeSH subheadings. Besides, the inference of unrelated predicates for the same co-occurrences will be analyzed, i.e. if in the extracted predications two concepts are associated with both Treats and Cause relations.

Results

The total number of aggregated co-occurrences from MRCOC is around 99,000. However, the input dataset for clustering only includes the resulting list of co-occurrences with LLR > 10.83, which reduces the number to 15,886 co-occurrences. Examples of co-occurrences with higher LLRs in our dataset are < HIV infection, AIDS Drugs >, < Hypertension, Antihypertensive Agents >, or < Grippe, Influenza vaccines >. Such co-occurrences with a high LLR are very frequent in MEDLINE records for the analyzed period.

The k-means algorithm produced centroids for each cluster and, hence, we can classify the co-occurrences into the cluster with the smallest Euclidean distance to its centroid. Nevertheless, co-occurrences can be ranked depending on their distance to the centroid of each cluster, as a consequence, we could compare the ranks of co-occurrences of each cluster and discover which ones are representative of more than one cluster.

The parameter of the number of clusters for the k-means algorithm was, firstly, obtained from the six relations suggested by UMLS SN between our selected semantic types; and, secondly, we used the expectation — maximization (EM) algorithm[27], which gave us an estimated number of five or seven clusters depending on the provided minimum standard deviation. Thus, we compute the clustering using five, six, and seven clusters. The clustering is based on the list of aggregated subheadings of each co-occurrence. Consequently, some subheadings are more predominant than others are. In Table 4, Table 5, and Table 6, we show the list of MeSH subheadings that have higher frequency in the co-occurrences for each cluster.

Because of the clustering, we obtained the lists of co-occurrences that belong to each generated cluster. The identifiers of the MEDLINE records, which are related to each co-occurrence, are also included in the MRCOC file. During the aggregation of the list of MeSH subheadings, the identifiers of such records were also collected and, therefore, we can use them to gather abstracts that are going to be processed with SemRep.

The collected corpus contains around 1,500 abstracts related the top co-occurrences of each cluster where, at most, 50 abstracts of the same co-occurring concept pair were collected. Each abstract corpus was processed by SemRep. The results were analyzed to extract the percentage of each predicate from the resulting triples where the subject and object are identical with co-occurring concepts of the cluster. However, this exact match occurred only in roughly 10% of the abstracts. Nevertheless, we assumed this subset reasonably representative for the whole. Thus, we normalize the frequency of each predicate in each cluster by dividing the resulting frequency by the total number of predicates that match the co-occurrences in a cluster. In particular, SemRep could identify through the different corpora the following predicates: Treats, Prevents, Affects, Causes, Associated with, Predisposes, Augments, and Disrupts. The predicates Predisposes, Augments, and Disrupts do not belong to UMLS SN, but they were proposed by SemRep. Figures 3 - 5 visualize the percentage of each predicate per cluster.

Table 4. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of five clusters.

	Subheadings 1st concept	**Subheadings 2nd concept**
Cluster 0	Drug Therapy (DT)	Therapeutic Use (TU), Administration and Dosage (AD)
Cluster 1	Prevention and Control (PC), Immunology (IM)	Immunology (IM), Administration and Dosage (AD)
Cluster 2	Drug Therapy (DT), Prevention and Control (PC)	Therapeutic Use (TU), Administration and Dosage (AD)
Cluster 3	Metabolism (ME), Blood (BL)	Metabolism (ME), Blood (BL)
Cluster 4	Chemically Induced (CI)	Adverse Effects (AE), Therapeutic Use (TU), Administration and Dosage (AD)

Table 5. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of six clusters.

	Subheadings 1st concept	**Subheadings 2nd concept**
Cluster 0	Drug Therapy (DT)	Therapeutic Use (TU), Adverse Effects (AE), Administration and Dosage (AD)
Cluster 1	Immunology (IM)	Immunology (IM), Administration and Dosage (AD)
Cluster 2	Drug Therapy (DT), Complications (CO) Prevention and Control (PC)	Therapeutic Use (TU), Administration and Dosage (AD)
Cluster 3	Metabolism (ME), Blood (BL)	Metabolism (ME), Blood (BL)
Cluster 4	Chemically Induced (CI)	Adverse Effects (AE), Therapeutic Use (TU) Administration and Dosage (AD)
Cluster 5	Drug Therapy (DT), Metabolism (ME) Pathology (PA)	Pharmacology (PD), Therapeutic Use (TU)

Table 6. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of seven clusters.

	Subheadings 1st concept	**Subheadings 2nd concept**
Cluster 0	Drug Therapy (DT)	Therapeutic Use (TU)
Cluster 1	Prevention and Control (PC), Immunology (IM)	Immunology (IM), Therapeutic Use (TU), Administration and Dosage (AD)
Cluster 2	Blood (BL), Diagnosis (DI)	Blood (BL)
Cluster 3	Metabolism (ME)	Metabolism (ME)
Cluster 4	Chemically Induced (CI)	Adverse Effects (AE), Therapeutic Use (TU) Administration and Dosage (AD)
Cluster 5	Drug Therapy (DT), Metabolism (ME) Pathology (PA)	Pharmacology (PD), Therapeutic Use (TU)
Cluster 6	Drug Therapy (DT)	Therapeutic Use (TU), Adverse Effects (AE), Administration and Dosage (AD)

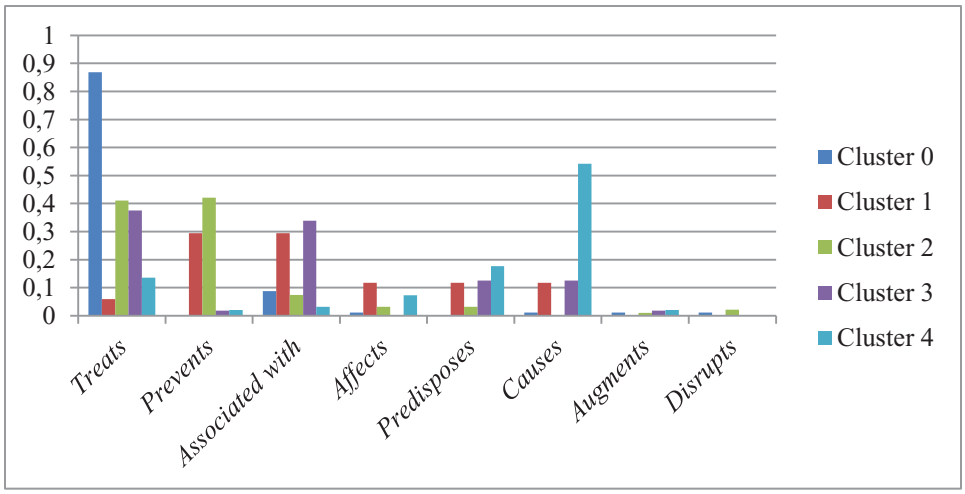

Figure 3. Percentage of relations that were extracted by SemRep within the five clusters generated using k-means.

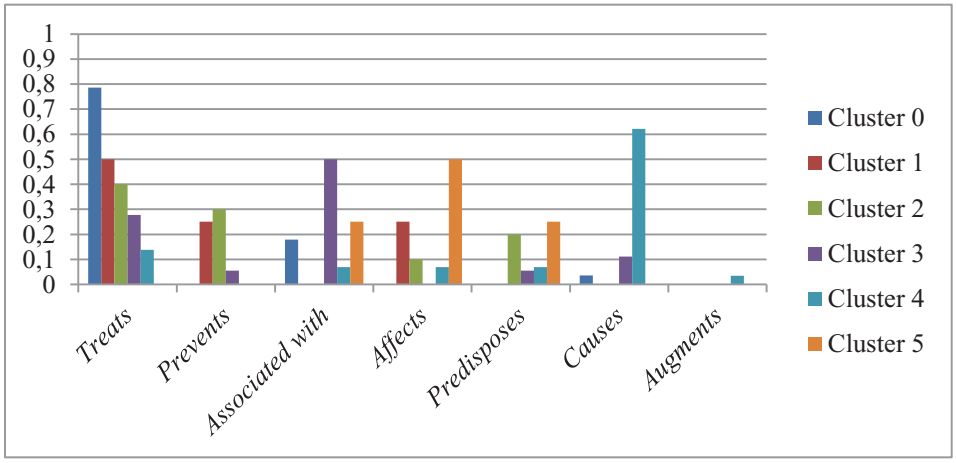

Figure 4. Percentage of relations that were extracted by SemRep within the six clusters generated using k-means.

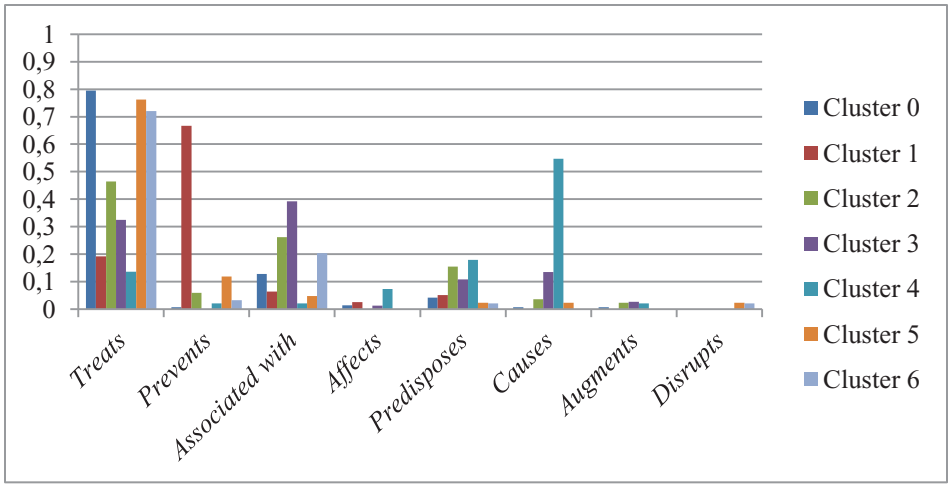

Figure 5. Percentage of relations that were extracted by SemRep within the seven clusters generated using k-means.

Table 7. List of the most relevant predicates for each cluster generated by 5 k-means, 6 k-means and 7 k-means.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
5 k-means	Treats	Prevents Associated with	Treats Prevents	Treats Associated with	Causes		
6 k-means	Treats	Treats Prevents Affects	Treats Prevents	Associated with Treats	Causes	Affects Associated with Predisposes	
7 k-means	Treats	Prevents	Treats Associated with	Associated with Treats	Causes	Treats	Treats

The analysis of the results of SemRep produces the association of the predicates with the corresponding clusters that are indicated in Table 7. These associations were obtained by selecting those predicates that, at least, were present in 25% of the total predications extracted by SemRep. Using Table 7 we could generate predications such as: < Nesiritide — Treats — Heart failure >, < BCG vaccine — Prevents — Tuberculosis >, < Estrogen — Associated with — Endometrioses >, < Erythropoietin — Affects — Ischemia > and < Nevirapine — Causes — Stevens-Johnson Syndrome >. Moreover, when two or more predicates were assigned to a cluster we could generate predications such as < Anticoagulants — Treats — Arterial Obstructive Disease > and < Anticoagulants — Prevents — Arterial Obstructive Disease >. A first inspection of 50 predications by an expert did not spot any clearly wrong predication. However, due to the high threshold used (p-value < 0.001), we assume that the recall is rather low. A systematic assessment is still outstanding and will be done in the next round after the inclusion of more semantic types and improvements of the matching process.

Discussion and Conclusions

This study is mostly descriptive and refrains from a detailed quantitative analysis, as it constitutes the first important milestone of a larger research project. The inspection of Fig. 3 - 5 seems to be only partly discriminative, but exhibit some interesting aspects that will guide our further work:

- Only a few relations exhibit a clear profile, viz. Treats, Prevents, and Causes. However, there is a cluster that includes in similar proportion Treats and Prevents, which is easily explained by the fact that there are substances used for both treatment and prevention.
- There is a clearly different profile of the relation Causes, because what causes a disease, is rather unlikely to be used in its prevention or treatment.
- Augments and Disrupts are too infrequent to allow any statement. Associated with, Predisposes, and Affects are very little discriminative. This is not surprising, because these relations are very vague. Affects may include causation, prevention, and treatment, Predisposes is difficult to interpret for disorder – substance associations, and Associated with could be as the most generic predicate which subsumes all of the other ones. This shows that the predicates generated by SemRep, but which are not present in the UMLS SN are not helpful for a clear semantic interpretation of the clusters.
- We hypothesize that other relations are more important to describe disorder – substance associations, especially the relation Diagnoses, which describes a substance that can be used to diagnose a disorder.
- The lack of discrimination may also be due to the genericity of many MeSH concepts such as Dermatologic Agent or Vaccines, for which there might be specializations for all predicates under scrutiny. This may lead to the decision to ignore too general concepts in the future refinement of the method.

The main advantage of selecting the most relevant predicates for each cluster is that a low p-value of the predicate could reduce the errors in the resulting set of predications obtained with NLP. For example, from the text: “Even conventional immunosuppressive agents, such as glucocorticoids and cyclosporine, directly affect podocyte structure and function, challenging the immune theory; of the pathogenesis of childhood nephrotic syndrome in which disease is caused by T cells.” SemRep extracts the predication < Immunosuppressive Agent — Causes —

Nephrotic Syndrome >. It is not surprising that a relation extractor fails with syntactically complex sentences like this one. This explains why a certain background noise is unavoidable when relying on NLP tools.

Another limitation is that the scope of SemRep is the whole UMLS Metathesaurus and not only its MeSH subset. This explains the low rate of matching between the concept pairs. This could be improved by inferring predications between MeSH concepts from predications between UMLS Metathesaurus concepts in general by traversal of hierarchical links. This approach can increase the number of matching predications and, thus, provide a bigger dataset.

The comparison between the inferred predicates (Table 7) and the resulting clusters from the three clustering experiments (Table 4, Table 5, and Table 6) allows us to identify the combination of MeSH subheadings that are closely related to each particular relation. From our experiments we obtained: (1) Treats is related to co-occurrences that the disease term is annotated with Drug Therapy and the substance with Therapeutic Use; (2) Prevents is associated with co-occurrences that the disease term is annotated with Prevention and Control, and Immunology and the substance with Immunology, and Administration and Dosage; (3) Causes is related to the co-occurrences that the disease term has Chemically Induced subheading and the substance has Adverse Effects, Therapeutic Use, and Administrative and Dosage; and (4) Associated with and Treats are related to the co-occurrences that both the disease and substance have the MeSH subheadings Metabolism and Blood.

Finally, an evaluation of the generated predications would be necessary to rate their plausibility. It is obvious that predicates that represent more than 80% of the extracted predicates in a cluster by SemRep are more plausible that those which are closer to 25%. Besides, the evaluation of the distance of co-occurrences to the cluster centroid might detect weaknesses that could guide the generation of predicates.

Acknowledgements

This paper was performed as a part of the BMFacts project (BMFacts: Knowledge acquisition for a biomedical fact repository), funded by the Austrian Science Fund (FWF): [M 1729-N15].

References

1. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 2004; 20 (3): 389-398.
2. National Library of Medicine (US). MEDLINE fact sheets; 2015 [Updated 2015 Feb 12]. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html
3. National Library of Medicine (US). Medical Subject Headings (MeSH); 2015 [Updated 2013 Dec 9]. Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html
4. National Library of Medicine (US). MEDLINE Co-Occurrences (MRCOC); 2013 [Updated 2014 Dec 3]. Available from: http://mbr.nlm.nih.gov/MRCOC.shtml
5. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 7: D267 — D270.
6. National Library of Medicine (US). UMLS Statistics - 2014AB Release; 2015 [Updated 2014 Nov 7]. Available from: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html
7. Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp. 2000: 575-579.
8. Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput. 2005: 91-102.
9. Schulz S, Costa CM, Kreuzthaler M, Miñarro-Giménez J A, Andersen U, Jensen AB, Maegaard B. Semantic relation discovery by using co-occurrence information. LREC 2014.
10. Cantor MN, Sarkar IN, Bodenreider O, Lussier YA. Genestrace: phenomic knowledge discovery via structured terminology. Pac Symp Biocomput. 2005: 103-114.
11. Srinivasan P, Hristovski D. Distilling conceptual connections from MeSH co-occurrences. Stud Health Technol Inform. 2004; 107(Pt 2):808-812.
12. UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 5, Semantic Network. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9679/
13. Barbara R, Hearst MA. Classifying semantic relations in bioscience texts. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04); 2004; Barcelona, Spain. Association for Computational Linguistics, Stroudsburg, PA, USA, Article 430.

14. Giles CB, Wren JD. Large-scale directional relationship extraction and resolution. BMC Bioinformatics. 2008 Aug 12; 9 Suppl 9:S11.
15. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005; 6 (1): 57-71
16. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec; 36(6):462-77.
17. Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearb Med Inform. 2013; 8(1):132-46.
18. Rector A. Barriers, approaches and research priorities for integrating biomedical ontologies. SemanticHealth Deliverable 6.1. Available from: www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf.
19. Mohammed EA, Far BH, Naugler C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. 2014; 7:22.
20. Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational linguistics. 1993: 61-74.
21. Dean J, and Ghemawat S. "MapReduce: simplified data processing on large clusters." Communications of the ACM. 2008: 107-113.
22. Amazon Web Services, Inc. Amazon Elastic MapReduce (Amazon EMR); 2015. Available from: http://aws.amazon.com/elasticmapreduce/?nc2=h_ls
23. Amazon Web Services, Inc. Amazon Simple Storage Service (Amazon S3); 2015. Available from: http://aws.amazon.com/s3/?nc2=h_ls
24. White T. Hadoop: the definitive guide. O'Reilly Media, Inc.; 2009
25. Shannon CE.A mathematical theory of communication. Bell System Technical Journal. 1948; 27: 379-423.
26. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 2009; 11 (1): 11-18.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statiscal Soc. 1977; 39 (1): 1-38