Semantic Relation Discovery by Using Co-occurrence Information

Stefan Schulz[1], Catalina Martínez Costa[1], Markus Kreuzthaler[1], Jose Antonio Miñarro-Giménez[2], Ulrich Andersen[3], Anders Boeck Jensen[4], Bente Maegaard[5]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[2]Medical Expert and Knowledge-Based Systems, Medical University of Vienna, Austria
[3]Sorano, Copenhagen, Denmark
[4]Center for Biological Sequence Analysis, DTU, Copenhagen, Denmark
[5]Institute for Language Technology, University of Copenhagen, Denmark

E-mail: stefan.schulz@medunigraz.at, catalina.martinez@medunigraz.at, markus.kreuzthaler@medunigraz.at, jose.minarrogimenez@meduniwien.ac.at, uan@sorano.dk, anders.boeck.jensen@gmail.com, bmaegaard@hum.ku.dk

Abstract

Motivated by the need of constructing a knowledge base for a patient-centred question-answering system, the potential of exploiting co-occurrence data to infer non-ontological semantic relations out of these statistical associations is explored. The UMLS concept co-occurrence table MRCOC is used as a data source. This data provides, for each co-occurrence record, a profile of MeSH subheading profiles. This is used as an additional source of semantic information from which we generate hypotheses for more specific semantic relations. An initial experiment was performed, limited to the study of disease-substance associations. For validation 20 diseases were selected and annotated by experts regarding treatment and prevention. The results showed good precision values (82 for prevention, 72 for treatment), but unsatisfactory values for recall (67 for prevention, 45 for treatment) for this particular use case.

Keywords: literature databases, knowledge engineering

1. Introduction

Motivated by a medical question answering use case we will investigate the automated construction of a supporting domain fact repository. Such a knowledge resource can be used for a series of possible applications, part of them directly related to biomedical text processing. In this context, the Danish ESICT project (Andersen et al., 2012) aims at developing strategies to provide natural language answers to laypersons' question on chronic diseases. One strategy of its hybrid approach has been the exploration of the content of SNOMED CT (2014), an ontology-based medical terminology. Its representational units (named SNOMED CT concepts, totalling about 300,000) group domain terms (about 700.000), which share a common meaning under specific concept categories (e.g. clinical finding, procedure, etc.) and are related by logical axioms, which express necessary and sufficient definitional conditions. Such axioms are strictly ontological, i.e. they state what is universally true for all individuals that instantiate a given SNOMED CT concept. Therefore, propositions of medical interest such as, for instance, relating typical diagnostic and therapeutic procedures, signs and symptoms to diseases, are not explicitly represented in SNOMED CT, which requires the exploitation of additional knowledge resources.

In this paper we investigate the potential of co-occurrence data as provided by the Unified Medical Language System (UMLS, 2014). They aggregate data on co-occurrences of semantic descriptors in bibliographic records. Their source is the MEDLINE database, in which each entry is indexed by a set of descriptors from the Medical Subject Headings (MeSH, 2014) vocabulary.

The goal of this study is to infer specific semantic relations out of these statistical associations.

Table 1 frames our hypothesis, viz. that co-occurrence patterns characterized by the defined semantic types involved, together with indirect semantic information of the context of a descriptor in the source (MeSH subheadings) may correspond to certain semantic relations. We focus on non-ontological predications, i.e. binary relations that are not used in definitions or universal statements as found in ontologies, thus excluding ontological relations like *part-of*, *has-site*, *has-quality* etc. In contrast, non-ontological predicates are not used in formal definitions as they express context-dependent and less strict associations, which, however, are often more “interesting” (Rector, 2008) from a medical point of view. Instead of stating what is necessarily true – like in formal ontologies – these predicates express what is typical, likely, or relevant. However, such knowledge is subject to change: A drug was indicated to treat a certain disease in the past, but it is mainly used for another purpose today or it has been withdrawn from the market. A clinical sign was frequently seen in the past, but it has become rare now, because the natural course of the underlying disease can no longer be observed, due to effective treatment.

Ideally, structured data in medical records would be a rich source of such associations. However, they are not commonly available and often lag behind the state of the art of scientific investigation. Medical literature abstracts are much easier to access and their semantic metadata – in this case MeSH annotations – constitute a valuable source of knowledge. If we want to exploit them for knowledge construction, the question arises whether it is reliable to interpret semantic associations between topics in scientific literature in the light of clinical practice so that they can be used as a raw material for the acquisition of medical predications. This topic will be further discussed. The goal of this study is a first exploration of the possibility of constructing symbolic knowledge out of statistical associations. Whereas one could identify a much more complex matrix than Table 1, in this preliminary study we restrict ourselves to the exploration of disease-substance associations from where we attempt to extract knowledge on (i) how a disease can be treated, and (ii) how a disease can be prevented.

	Disease	**Finding**	**Substance**	**Organism**
Finding	sign of symptom of	accompanied by	treated by	affects caused by
Substance	*causes* *treats* *prevents* *metabolite*	causes treats prevents	interacts	affects is produced by
Organism	Causes affected by	Causes observed in organism	sensitive to	interacts with
Body part	possible location of	possible location of	targeted by	targeted by

Table 1. Examples of semantic relations between concepts ordered by semantic types. In this paper only co-occurrences between disease and substance concepts are explored (italics).

2. Materials

The Unified Medical Language System (UMLS) links representational units (classes, concepts, terms) from about 60 families of biomedical vocabularies. (Quasi-)synonymous terms are aggregated as UMLS concepts and identified by a unique identifier (CUI). The U.S. National Library of Medicine (NLM) produces and distributes the UMLS Knowledge Sources (databases) among which three files have been used in this work: (1) MRCOC, (2) MRCONSO and (3) MRREL.

- MRCOC provides pairs of concepts that co-occur in the same entries in some information source. It summarizes the MeSH descriptors that occur together in MEDLINE citations from the MEDLINE/PubMed baseline, a snapshot created at the beginning of each new MeSH indexing year. The co-occurrences are summarized by timeframe (MED, last five years of MEDLINE; MBD, previous five years of MEDLINE (years 6-10). In this study we have only used the most recent data set (MED). Besides the main headings, MRCOC includes a "fingerprint" of MeSH subheadings, such as "DT" (Drug Therapy) or "PC" (Prevention & Control). They characterize the context in which the first concept occurs in the related MEDLINE records. E.g., if an article is about the prevention of Stroke, the concept "Stroke" in the MEDLINE record is refined by the subheading "PC".
- MRCONSO relates UMLS CUIs with their language, source vocabularies, synonyms, translations, and lexical variants. We use this table to extract mappings between UMLS CUIs and SNOMED CT concepts. We are interested in SNOMED CT concepts due to their clinical relevance, but also because they are consistently grouped into semantic types, a few of them being depicted in Table 1.
- MRREL provides relation triples, again indexed by source. We will use this file for extracting hierarchical relationships.

Table 2 shows an example of the content of each of the previous files.

UMLS file	**Example of UMLS file content**
MRCOC	**C0000039**|**C0000506**|**MED**|L|**1**|**CH=1**|
MRCONSO	**C0000039**|**ENG**|S|L3000054|PF|S3260062|Y|A8383517|544223010|**102735002**||**SNOMEDCT**|OF|102735002|**Dipalmitoyl-phosphatidylcholine (substance)**|9|O||
MRREL	**C0002871**|**CHD**|**C0002891**||MSH|MSH||

Table 2. The UMLS files used, with examples. The fields relevant for this work are picked out in bold face, such as concepts, co-occurrence frequency and subheading from MRCOC, the linkage to SNOMED CT in MRCONSO and a hierarchical relationship between two concepts in MRREL.

3. Methods

Our methods can be divided into:

- Mappings of co-occurrence pairs to SNOMED CT
- Calculation of the relative co-occurrence value
- Analysis of MeSH sub-headings
- Prototypical Implementation
- Evaluation strategy

Mapping of co-occurrence pairs to SNOMED CT

In order to add the corresponding SNOMED CT concept ID to each of the new UMLS concepts added, the UMLS MRCONSO file was used, which contains the SNOMED CT correspondences of the UMLS concepts when there is any. As there is a 1:n relation between UMLS and SNOMED CT concepts, mappings from numerous SNOMED CT concepts to the same UMLS concept occur. In MRCOC records we also find UMLS concepts that have no SNOMED CT correspondence. In such cases, the co-occurrence pair is discarded.

Computation of relative co-occurrence values

The overall frequency of MEDLINE annotations varies across several orders of magnitude. A relatively low co-occurrence value may be more expressive than a higher one, in case the latter combines concepts of very high frequency such as, e.g. "Diabetes mellitus" and "Antibiotics". We hypothesize that this could be a source of error. We therefore used two thresholds, viz. the log likelihood ratio (Dunning, 1993) and an absolute threshold. A co-occurrence was seen as significantly expressive if the absolute co-occurrence was greater than five (McDonald, 2009), and the log likelihood ratio was greater than 6.63, which corresponds to $p < 0.01$.

Analysis of MeSH subheadings

After a thorough analysis of the 85 subheadings provided by MeSH we concluded that the subheadings "DT" and "PC" are highly indicative for the meaning of "treats" and "prevents", i.e. the two predicates we expect to induce. We decided to require the respective subheading for more than half of the co-occurrences. E.g., with a co-occurrence of 11, and the subheading distribution of "DT=5; PC=6", only "prevents" would be asserted.

Source concept	*Name (SNOMED CT)*	Bipolar disorder
	UMLS ID (CUI)	C0005586
	SNOMED ID	13746004
	SNOMED CT semantic type	Disorder
Target concept	*Name (SNOMED CT)*	Tricyclic antidepressant
	UMLS ID (CUI)	C0003290
	SNOMED ID	373253007
	SNOMED CT semantic type	Substance
MeSH subheadings		DT=9,CI=7,DI=5, PX=4,CO=2,EP=2, GE=2,BL=1,ET=1, PA=1,PC=1,PP=1, TH=1
Original Table	*Absolute co-occurrence*	17
	Relative co-occurrence (log-likelihood)	54.57

Table 3. Example record. The pairing of the two UMLS concepts has a co-occurrence in MED (last 5 year MEDLINE) of 17. If adding pairings from all UMLS subconcepts this value raises to 714. Relative co-occurrences are the decadic logarithms of the ratio of an absolute co-occurrence and a theoretic baseline (random co-occurrence). The MeSH subheading counts (for the original co-occurrence) refer to subheadings assigned to the first concept (or better the MeSH term mapped to it), e.g. nine DT (drug therapy), one PC (prevention and control)

Prototypical implementation

We developed a command line based interface in Java, using Apache Lucene 2.0 for indexing and the Apache Mahout Math library for log likelihood calculation. The *tempusfugit* package from Google Code was used for co-occurrence calculation. The interface provides a combined search of the fields shown in Table 3. The output result is shown in the console in descending order by relative co-occurrence in a short human readable format for quick search result feedback. Additionally, it is recorded as comma-separated values in an output file for further analysis. Further options include how many results should be returned as well as a cut-off-threshold for the co-occurrence value. VBA scripts were produced to facilitate the interpretation of the result by ordering, filtering, and colour coding the content in Excel spreadsheets.

Evaluation strategy

A reference standard was acquired by a random extraction of twenty disease terms from a textbook of internal medicine (Herold, 2014). For each disease, two of the authors, who are MDs, created a reference standard by collecting references to substances that are recommended for therapy and prevention. Various online and offline sources were used, with a focus on Wikipedia, as well as Danish treatment guidelines. Both MDs included only recommendations that appeared to be based on scientific evidence. For each disease, the co-occurrence table was filtered according to the following criteria, which had been heuristically acquired in a series of pre-tests with training data. The following parameters were measured, each for DT and PC:

Recall 1: Strict recall of reference standard concepts, regardless of hierarchical level.

Recall 2: Generous recall: here descendants of the concept in the reference standard were equally accepted.

Precision: Each retrieved concept is checked for correctness, i.e. whether it treats or prevents the disease under scrutiny. The criterion here is: given the state of the art, there is at least some clinical evidence that the treatment or prevention strategy is recommendable for humans. A thorough assessment would require a clinical review board. As this was not possible, we performed a cursory check of primary and secondary literature.

4. Results

Table 4 gives the result for the recall and precision analyses for all 20 sample diseases. Only for two diseases (*Infectious mononucleosis*, *Syncope*) the thresholds were not reached. More than ten results were only found for five diseases regarding therapy and two results regarding prevention.

In detail, the number of results per disease ranged for treatment from 0 to 40 (median = 2; mean = 6.7), and for prevention from 0 to 36 (median = 0; mean = 3.7). The strict recall values ranged from 0 to 1 (median = 0.35; mean = 0.42) for treatment and from 0 to 1 for prevention (median = 0. 50; mean = 0.49). The generous recall values

Disease	*# Target concepts*	*Recall (strict)*	*Recall (generous)*	*Precision (Correctness)*
Giant Cell Arteritis *C0039483*	13 / 0	1.00 / –	1.00 / –	0.77 / –
Cerebrovascular accident *C0038454*	40 / 36	0.50 / 0.57	0.83 / 0.86	0.62 / 0.83
Appendicitis *C0003615*	3 / 0	0.67 / –	1.00 / –	1.00 / –
Anthrax disease *C0003175*	1 / 2	0.10 / 0.30	0.10 / 1.00	1.00 / 1.00
Pre-eclampsia *C0032914*	6 / 6	0.50 / 0.33	0.50 / 0.33	0.50 / 0.16
Yellow fever *C0043395*	1 / 1	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Gallbladder Carcinoma *C0235782*	3 / 0	0.33 / –	1.00 / –	1.00 / –
Membranous glomerulonephritis *C0017665*	10 / 0	0.67 / –	0.67 / –	0.90 / –
Hemolytic Anemia *C0002878*	2 / 0	0.33 / –	0.33 / –	1.00 / –
Hepatitis B *C0019163*	13 / 5	0.63 / 1.00	0.63 / 1.00	0.62 / 1.00
Impetigo *C0021099*	1 / 0	0.12 / –	0.12 / –	1.00 / –
Infectious mononucleosis *C0021345*	0 / 0	– / –	– / –	– / –
Pertussis *C0043167*	1 / 1	0.25 / 0.50	0.25 / 0.50	1.00 / 1.00
Malaria *C0024530*	14 / 16	0.36 / 0.67	0.36 / 0.67	0.79 / 0.75
Osteitis Deformans *C0029401*	2 / 0	0.22 / –	0.22 / –	1.00 / –
Neurosyphilis *C0027927*	2 / 0	0.20 / –	0.20 / –	1.00 / –
Gastric ulcer *C0038358*	19 / 7	0.22 / 0.00	0.22 / 0.00	0.53 / 0.00
Syncope *C0039070*	0 / 0	– / –	– / –	– / –
Tachycardia, Paroxysmal *C0039236*	2 / 0	0.50 / –	0.50 / –	1.00 / –
Erysipelas *C0014733*	1 / 0	0.25 / –	0.25 / –	1.00 / –

Table 4. Recall and Precision for DT (“Drug Therapy”) / “Prevention & Control” (PC), with UMLS identifiers

(in which more general terms were allowed for matching) equally ranged from 0 to 1 (median = 0.35; mean = 0.45) for treatment and from 0 to 1 for prevention (median = 0. 77; mean = 0.67). Finally, the precision values ranged from 0 to 1 (median = 1; mean = 0.82) for treatment and from 0 to 1 for prevention (median = 0. 92; mean = 0.72). The large variation of the result is only understandable in a case to case analysis, which also reveals sources of error and demonstrates routes to improvement:

- Concept mismatch. Low recall values are often explained by the fact that the reference standard contained rather comprehensive lists of substance concepts (especially antibiotics), which were not contained in the co-occurrence table, to a large extent. A reference standard restricted to concepts that occur in the co-occurrence table would have yielded better results, as well as variations of the matching criteria, in the sense that a general term suggested by the system (e.g. *Antibacterial agent*) would match all existing antibiotics in the reference standard.
- Underspecifications of what was a therapy and what a preventive measure were observed in symptomatic treatments (e.g. *Intensive care treatment* in cases of *Yellow fewer*) or preventive measures that consist in the drug treatment of an underlying cause, e.g. treatment of *Arrhythmia* and *Arterial hypertension* as prevention of *Stroke*.
- Another issue is how to deal with widely practiced treatments of debatable evidence, such as, e.g. of

Glucosamine in *Osteoarthritis*.

- Interference with substance side effects. The zero precision at *Gastric ulcer* prevention was partly due to the fact that chemicals were found that *cause* gastric ulcerations. This phenomenon could only be observed here, because the sample did not contain other diseases that can be caused by substances. This shows the weakness of the subheading information we used, which was restricted to the analysis of the source concept only. The methods could be mitigated improved by a more detailed analysis of the subheading profile and the formulation of more rules, also including the relation "causes" as a possible outcome. In addition, the converse co-occurrence records could be included into the analysis.
- Lacking interest by the scientific community. The missing results for *Syncope* and *Impetigo* result from overall low (co-)occurrences. Treatment and prevention of these conditions have not changed for a long time, so that little information is contained in the co-occurrence dataset. In such cases, co-occurrence data from earlier time intervals could be useful. However, in case that co-occurrences of interest are only found in older datasets, a competing interpretation must be considered, *viz.* that a certain therapy became obsolete. For well-established, non-changing therapeutic measures, clinical co-occurrence data would probably yield less ambiguous results.
- No positive outcome of research. It is obvious that a high co-occurrence marked with an appropriate subheading profile will also occur in those cases where intensive research of drug effects could not be translated into clinical practice for several reasons, e.g. lack of superiority in clinical trials. An example found in our data was *Antiviral therapy* for *Yellow fever* or the use of *Zinc* in *Malaria* prevention. To identify such cases time series of co-occurrence data might be helpful, as well as filtering by publication types (e.g. *review* or *randomized controlled trial*).
- Ongoing research. This may produce high co-occurrence values even if a study is still restricted to animal models, such as current investigations of peptic ulcer prevention in rats using plant extracts. Here, information easily available in MEDLINE (*human* vs. *non-human*), but not connected to the co-occurrence dataset, could be used.

5. Related Work

Several authors have used the UMLS co-occurrence data, but there is a general impression that this resource has been rather underused. Burgun and Bodenreider (2001) analysed MRCOC, using three levels of semantic granularity, categorising by concept clusters that semantically cover a restricted area of interest. They found co-occurrence information helpful for this clustering task insofar as the redundancy between co-occurrence linkages and symbolic linkages were low. This corresponds to our division between ontological relations and non-ontological predicates. UMLS SN relationships could be relatively well inferred from co-occurrence information, like in our study, with a focus on the relation between disorders and chemicals. Question answering and enhanced information retrieval was a major driver of a study performed by Mendonça & Cimino (2000). They built their own co-occurrence table based on PubMed clinical queries and exploited its usefulness for gathering additional medical knowledge for knowledge base building. An automated approach for harvesting disease-chemical relationships was proposed by Zeng & Cimino (1998), based on UMLS MRCOC. They further evaluated the quality of the extracted knowledge by comparing the acquired relations with the expert system DXplain and manually extracted medical knowledge from literature. In disease-drug chemical relationships they achieved 93% sensitivity and 68% in disease-lab chemical relationships. Cantor et al. (2005) inferred gene-to-disease relationships using statistical and semantic relationships exploiting MRREL and MRCOC. Like in our approach they considered a threshold of five co-occurrence instances necessary to infer a concept-concept relation. They interlinked the relevant concepts with the Gene Ontology (GO) and evaluated the retrieved gene-to-disease relationships with the Online Mendelian Inheritance in Man's morbidmap (OMIM).

There are several systems that have successfully implemented information extraction methods to process biomedical literature databases. Examples are GOPubMed (Doms & Schroeder, 2005), MedlineR (Lin et all, 2004), FACTA (Tsuruoka et al. 2008), Alibaba (Plake et al., 2006), PolySearch (Cheng et al, 2008) and SemMedDB (Kilicoglu et al., 2012). All these systems concentrate on the knowledge extraction from title and abstracts using natural language processing methods. We have not found any previous work on the use of MeSH subheadings as an additional source of semantic information. The use of this information has been central in our work, which is, admittedly, preliminary and still restricted to the narrow scope of disease-substance associations. Another distinguishing feature of our approach is its reference to SNOMED CT as the emerging worldwide terminological standard. Although the conceptual space covered by MRCOC is much more coarse-grained, the availability of hierarchical links via the MRREL table allows inferring SNOMED-SNOMED co-occurrence values for concepts that have no direct representation in the MRCOC table.

The error analysis we performed on disease-substance co-occurrences demonstrated not only the need to refine the matching criteria but also to consider relations other than "treats" and "prevents", especially causation, we have ignored in this study, with the effect that the system suggested alcohol for prevention of peptic ulcers. Besides the need for improved criteria for the construction of the reference standard it has shed light on the (non-surprising) fact that research hypotheses and outcomes only partly translates in clinical practice.

6. Conclusion and future work

This study, motivated by the need to construct a knowledge base for patient-centred question-answering systems has been restricted to the investigation of substance-drug association, which is only one aspect. Next steps will be other semantic relationships as shown in Table 1, especially the relations between findings and diseases, with a focus on early diagnosis, risks, and prognostic factors. We have to keep in mind that the use of this kind of output can only be one of several knowledge sources in a question answering or decision support pipeline. Support by several sources and careful weighting are mandatory to prevent wrong answers or recommendations.

For the continuation of our work we still focus on MeSH annotations, and place special emphasis on the analysis of co-occurrences, but additional information from MEDLINE records such as timestamps, organisms, and publication types should additionally be exploited. This will require processing the whole body of MEDLINE. We will systematically analyse and categorise errors and try to identify indicative patterns for them. Currently we are incorporating co-occurrence based predications into a Web application, which compares several question-answering methodologies as developed by the ESICT project.

7. Acknowledgements

This paper was performed as a part of the ESICT project (Experience-oriented Sharing of health knowledge via Information and Communication Technology), funded by the Danish Strategic Research Council. http://cst.ku.dk/english/projekter/esict/.

8. References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., Wedekind, J. (2012). Creation and use of Language Resources in a Question-Answering eHealth System. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Burgun, A., Bodenreider, O. (2001). Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. *Studies In Health Technology And Informatics* 84, 171-175.

Cantor, M.N., Sarkar, I.N., Bodenreider, O., Lussier, Y.A. (2005). Genestrace - phenomic knowledge discovery via structured terminology. *Pacific Symposium On Biocomputing* 114, 103-114.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research* 36: W399-405

Doms, A., Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Research 33: W783-786.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Herold G. (2014) *Herold's Internal Medicine*, eBook at http://www.herold-internal-medicine.com/

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C. (2012). SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications. *Bioinformatics* 28(23): 3158-60.

Lin, S.M., McConnell, P., Johnson, K.F., Shoemaker, J. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics* 20(18): 3659-61.

McDonald, John H. (2009) Handbook of biological statistics. Sparky House Publishing Baltimore, MD (2).

Mendonça, E.A., Cimino, J.J. (2000). Automated knowledge extraction from MEDLINE citations. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 575-579.

Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., Leser, U. (2006) AliBaba: PubMed as a graph. *Bioinformatics* 22(19): 2444-2445.

Rector A. (2008) Barriers, approaches and research priorities for integrating biomedical ontologies. Available from: www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf.

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: http: //www.ihtsdo.org/snomed-ct.

Tsuruoka, Y., Tsujii, J., Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24(21): 2559-2560.

Unified Medical Language System. (2014) http://www.nlm.nih.gov/research/umls/

Zeng, Q., Cimino, J.J. (1998). Automated knowledge extraction from the UMLS. *Proceedings of the AMIA Symposium* 568-572.