BioTxtM2014 – Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing

Semantic Relation Discovery by Using Co-occurrence Information

Stefan Schulz[1], Catalina Martínez Costa[1], Markus Kreuzthaler[1], Jose A. Miñarro-Giménez[2], Ulrich Andersen[3], Anders B. 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

MeSH annotations of bibliographic records are explored as a source for generating factoid statements, based on their statistical co-occurrence and their subheading profile, which helps disambiguate between competing interpretations like 'substance causes disease' vs. 'substance prevents disease' vs. 'substance treats disease'.

INTRODUCTION

ONTOLOGY vs. KNOWLEDGE BASE

Ontologies represent what is universally true, using predicates like "has-part", "has-quality" , "is located in". Probabilistic, contingent, and default statements are often more "interesting" for knowledge intensive applications (Retrieval, Question Answering, Decision support) [1,2].

CHALLENGE: To automatically build a "knowledge" layer on top of the clinical ontology SNOMED CT, constituted by <SUBJ; PRED; OBJ> triples, with predicates like:

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

MATERIALS / METHODS

- **MEDLINE database:** 22 M bibliographic records semantically annotated with MeSH thesaurus main headings (medical terms) and subheadings (qualifiers)
- **UMLS:** cross-mapping to other medical terminologies, e.g. SNOMED CT; MEDLINE annotations aggregated in co-occurrence table.

Source concept	*Name*	Bipolar disorder
	Type	Disorder
Target concept	*Name*	Tricyclic antidepressant
	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
Absolute co-occurrence		17
Log-likelihood		54.57

- Identify typical subheading profiles for combinations of semantic types. The above example suggests a high rate of DT ("drug therapy") co-occurrences as indicative for the semantic relation "treated by".
- **Implementation**: *Java* command line / *Lucene*
Evaluation: Two of the authors (MDs) created reference standard of substances for treatment and prevention of 20 randomly selected diseases. Parameters: strict recall (considers only the reference standard concept); generous recall (considers also hierarchically related concepts); precision (considers sources of generally accepted clinical evidence).

RESULTS

Thresholds: absolute co-occurrence > 5, log-likelihood ratio > 6.63 (corresponding to $p < 0.01$). Requested rate of subheadings "DT" > 0.5 for "treats" and "PC" > 0.5 for "prevents".

Disease	**# Target concepts**	**Recall (strict)**	**Recall (generous)**	**Precision (Correctness)**
Giant Cell Arteritis	13 / 0	1.00 / –	1.00 / –	0.77 / –
Cerebrovascular accident	40 / 36	0.50 / 0.57	0.83 / 0.86	0.62 / 0.83
Appendicitis	3 / 0	0.67 / –	1.00 / –	1.00 / –
Anthrax disease	1 / 2	0.10 / 0.30	0.10 / 1.00	1.00 / 1.00
Pre-eclampsia	6 / 6	0.50 / 0.33	0.50 / 0.33	0.50 / 0.16
Yellow fever	1 / 1	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Gallbladder Carcinoma	3 / 0	0.33 / –	1.00 / –	1.00 / –
Membr.glomerulonephritis	10 / 0	0.67 / –	0.67 / –	0.90 / –
Hemolytic Anemia	2 / 0	0.33 / –	0.33 / –	1.00 / –
Hepatitis B	13 / 5	0.63 / 1.00	0.63 / 1.00	0.62 / 1.00
Impetigo	1 / 0	0.12 / –	0.12 / –	1.00 / –
Infectious mononucleosi	0 / 0	– / –	– / –	– / –
Pertussis	1 / 1	0.25 / 0.50	0.25 / 0.50	1.00 / 1.00
Malaria	14 / 16	0.36 / 0.67	0.36 / 0.67	0.79 / 0.75
Osteitis Deformans	2 / 0	0.22 / –	0.22 / –	1.00 / –
Neurosyphilis	2 / 0	0.20 / –	0.20 / –	1.00 / –
Gastric ulcer	19 / 7	0.22 / 0.00	0.22 / 0.00	0.53 / 0.00
Syncope	/ 0	– / –	– / –	– / –
Tachycardia, Paroxysmal	2 / 0	0.50 / –	0.50 / –	1.00 / –
Erysipelas	1 / 0	0.25 / –	0.25 / –	1.00 / –

DISCUSSION

So far, only small segments of knowledge space analysed (diseases vs. substances). Known issues**:**

- **Coverage**: (MeSH annotations are coarse grained, especially regarding substance concepts) -> low recall;
- **Validity**: hypotheses without scientific evidence, animal studies, ongoing research -> low precision;
- **Lacking interest**: trivial associations, not subject to current research -> low recall;
- **Underspecification of predicates**: E.g., substance effects like adverse reactions, metabolites missed.

OUTLOOK

CONCLUSION: MeSH subheading information – underused resource, but promising for inferring factoid statements.

CURRENT STATE:

- analysis of further subheading profiles
- implementation as scripts, e.g.:

```
treats [subject term] [object term] [# results]
```

OUTLOOK:

- Inclusion of additional information from MEDLINE, e.g. publication type, year (for trend information);
- Use text mining for acquiring additional entities from abstracts (as new entry terms / hyponyms for asserted concepts in co-occurrence records);
- Accumulate co-occurrence counts upwards along MeSH hierarchies;
- Improve precision by matching output triples against external sources (e.g. web mining);
- Use clustering methods for finding new, highly predictive co-occurrence profiles.

References:

[1] Rector A. (2008) Barriers, approaches and research priorities for integrating biomedical ontologies. SemanticHealth Deliverable D6.1 www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf.

[2] Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearbook of Medical Informatics 2013;8(1):132-46.

Acknowledgements: This work 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/.