Learning a Health Knowledge Graph from Electronic Medical Records

Ole Berg

November 19, 2024

Table of Contents

1 Motivation and earlier work

- Demand for decision support systems in clinical settings
- Existing knowledge bases created manually or *"using simple pairwise statistics"* [1, p. 1]
- E. g. 15 person-years needed for *Internist-1/QMR* knowledge base
- Manually developed systems very brittle and difficult to extend
- Automatic compilation speeds up development of KBs
- *WatsonPath* by IBM and *Isabel* use NLP to find relations betweens diseases and symptoms in textbooks and journals

2 Goal and methods

- Utilize electronic medical record (EMR) to construct a knowledge graph
- Validation against Google health knowledge graph (GHKG)
- Three steps for knowledge graph generation:
  1. Data collection and preparation
  2. Learning of statistical models
  3. Transformation of models into knowledge graphs

3 Electronic Medical Record (EMR)

- EMR sometimes used interchangeably with electronic health record (EHR)
- Some authors distinguish between these terms: EHR is comprehensive collection of EMRs
- EMR is information on a patient from one **TODO**
- EMRs useful as data source because they represent diseases and their symptoms in real-world environment
- Difficult data source for four reasons:
  1. Notes from physicians and notes less formal
  2. Comorbidities, confounding factors and nuances present
  3. Associations between diseases and symptoms are statistical
  4. Pre-filtered by physician

4.1 Data collection and preparation

- Focus on positive mentions of diseases and their symptoms
- Structured data:
  - ICD-9 codes
- Unstructured data:
  - Triage Assessment
  - Nursing Notes
  - MD comments
- Diseases and symptoms chosen from GHKG
- Mapping of extracted concepts to a concept ID

4.2 Learning of statistical models

- Three statistical models:
  - Naive Bayes
  - Logistic regression
  - Noisy OR gates
- Parameter learning with maximum likelihood estimation
- L1 regularization used for logistic regression
- Laplacian smoothing used for naive Bayes

4.3 Transformation of models into knowledge graphs

- Estimating the importance of edges (connections between diseases and symptoms)
- One importance measure for each statistical model
- Maximum of five symptoms per disease

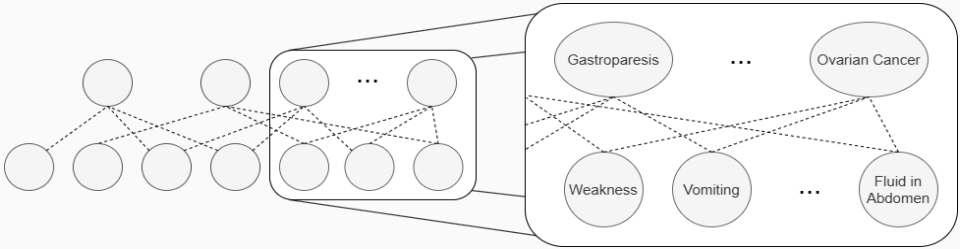

Figure 1: Resulting knowledge graph (own illustration according to [1, p. 4])

5 Evaluation

5.1 Methods

- Automatic evaluation against GHKG
- Assumption: GHKG is precise, but not complete
- -> relative comparison between models, not an absolute evaluation
- Two best performing models and GHKG evaluated by physicians
- Physicians tag suggested disease-symptom edges on a 4-point scale ranging from "Always happens" to "Never"
- Binarization with "Never" as negative and other three as positive
- Precision-recall curve as evaluation measure

5.2 Results

- Naive Bayes and noisy OR perform considerably better than logistic regression
- Performance better when models compared to evaluations by physicians
- With recall of 0.5: Precison of noisy OR at 0.87, of naive Bayes at 0.8
- Conclusion: Noisy OR better than naive Bayes; statsitically significant ($p = 0.01$)

6 Discussion

- Three kind of differences between edges suggested by the model and by GHKG:
  - GHKG focuses on information for web users
  - GHKG uses less precise language
  - Less severe edges in GHKG
- Naive Bayes and logistic regression suggest symptoms casued by confounding factors
- Noisy OR often suggests general symptoms
- Difficulty inferring causation from correlation
- Confounding factors difficult to recognize and eliminate

7 Future improvements

- Edges between symptoms
- Softer boundary between symptom and disease
- Introduce a manual filter step
- Use other, non-parametric models
- Higher coverage, more input data

8 Conclusion

Thank you for attending my presentation!

Do you think that Schema.org still has **unused potential**? Or are the extensions all we can expect?

9 Sources I

- [1] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, und D. Sontag, „Learning a Health Knowledge Graph from Electronic Medical Records“, Sci Rep, Bd. 7, Nr. 1, S. 5994, Juli 2017, doi: 10.1038/s41598-017-05778-z.

9 Sources II

- U. Serles und D. Fensel, “Analysis of Schema.org at Five Levels of KR”, in *An Introduction to Knowledge Graphs*, Cham: Springer Nature Switzerland, 2024, S. 259–270. doi: 10.1007/978-3-031-45256-7_15.
- P. Hitzler, “A review of the semantic web field”, *Commun. ACM*, Bd. 64, Nr. 2, S. 76–83, Jan. 2021, doi: 10.1145/3397512.