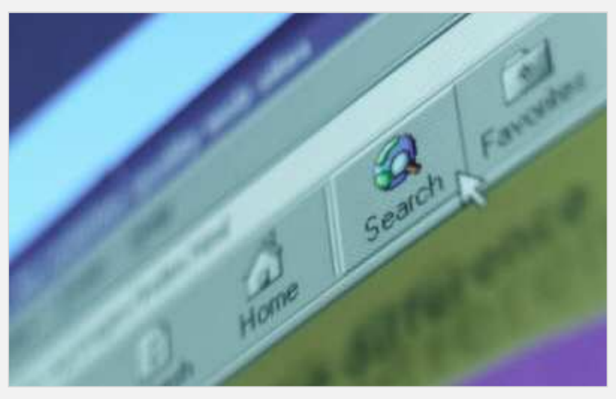
Relation Extraction for Diseases and their Determinants within the DIDO Framework

Patrick Ernst

Introduction

Motivation



Study by Deloitte & Touche and VHA

- Health information is one of the most frequently requested types of information
- 17.5 million adults in the United States, or 43% of the 40.6 million who use the Internet, are searching for health information

Text Corpora

Laymen vs. Professional Sources

- Sources ranges from laymen to professional text corpora
- Laymen sources are the encoclypedias: *Mayo Clinic* and *Wikipedia*
- Professional source are scientific publications: *Medline* and *Pubmed Central*

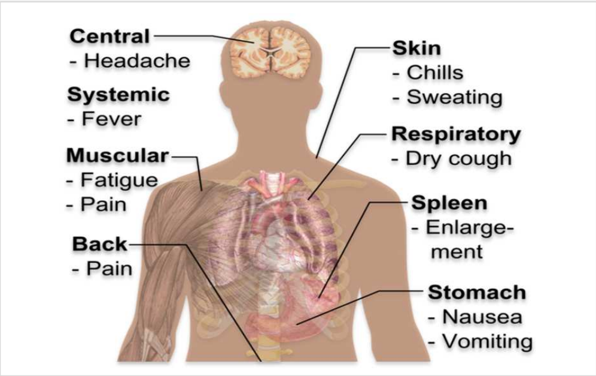
Approach

Entity Recognition

Mapping to UMLS Dictionary

- UMLS integrates approximately 150 medical vocabularies
- MetaMap maps biomedical text to concepts
- Semantic types and semanitc groups are assigned to concepts

Relation Extraction



Pattern-based Approach

- Manually extracted seed relations form the basis
- Approach relies on patterns extracted between recognized entities
- Reasoning is applied for determining the patterns expressing what relation
- New relations are gained by linking patterns to seed relations

Results

Evaluation

Quality Assessment with Web Surveys

- Experts and laymen are evaluating the results
- Evaluaters judge if the relations are correctly extracted from a textual context
- Different text corpora will be characterized

Future Work

Ternary Relations, Complex Negations, Qualified Relations

- "GDM is a condition in which women without previously diagnosed diabetes exhibit high blood glucose levels during pregnancy."
- "Psychosis is not pathognomonic for schizophrenia."
- "However, with aspiration, fevers might also indicate aspiration pneumonia."

References

1. V Nebot, M Ye, J-H Eom, and G Weikum. *DIDO: a Disease-Determinants Ontology from Web Sources.* In Proceedings of WWW 2011.
2. F Suchanek, M Sozio, G Weikum. SOFIE: A Self-Organizing Framework for Information Extraction. In Proceedings of WWW 2009.
3. N Nakashole, M Theobald, G Weikum. Scalable Knowledge Harvesting with High Precision and High Recall. In Proceedings WSDM 2011

max planck institut informatik

UNIVERSITÄT DES SAARLANDES

International Max Planck Research School for Computer Science