MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-1032

Extracting Dependence Relations from Unstructured Medical Text

Charles Jochim [a], Yassine Lassoued [b], Bogdan Sacaleanu [a], Léa A. Deleris [a]

[a] *IBM Research – Ireland, Dublin, Ireland*
[b] *CMRC – University College Cork, Cork, Ireland*

Abstract

Dependence relations among disease and risk factors are a key ingredient in risk modeling and decision support models. Currently such information is either provided by experts (costly and time consuming) or extracted from data (if available). The published medical literature represents a promising source of such knowledge; however its manual processing is practically infeasible. While a number of solutions have been introduced to add structure to biomedical literature, none adequately recover dependence relations. The objective of our research is to build such an automatic dependence extraction solution, based on a sequence of natural language processing steps, which take as input a set of MEDLINE abstracts and provide as output a list of structured dependence statements. This paper presents a hybrid pipeline approach, a combination of rule-based and machine learning algorithms. We found that this approach outperforms a strictly rule-based approach.

Keywords:

Risk; Risk Factors; Artificial Intelligence; Natural Language Processing; Data Mining.

Introduction

Bayesian belief networks are a convenient tool for macro-level risk models, i.e., risk models articulating the dependence between multiple risk factors, diseases and conditions. They have long been advocated as a useful decision and risk modeling framework in medicine [1]. Our focus on automating the extraction and aggregation of risk information seeks to address the practical challenge of building such networks. We focus specifically on the information contained in MEDLINE, which has been steadily growing at a rate of about 1M publications per year in the past few years and for which manual consumption is no longer practical. The larger context of this research is to build a system, called Medical Recap, which automatically extracts risk information from medical papers and then aggregates this knowledge into a Bayesian network.

Methods

This paper focuses on the sub task of extracting dependence information. From the sentence "Smoking increases the risk of lung cancer", we want to extract the relation (Smoking, Lung Cancer). Our approach to this challenge can be divided into two sub-problems: entity detection and relation construction. Entity detection deals with the identification of the elements in a sentence that are part of a relation. Relation construction is about articulating those elements together (which is not straightforward when multiple relations are expressed in the same sentence). We rely on machine learning for entity detection and propose a rule-based algorithm for defining the relations between the extracted entities. More specifically, we have developed a pipeline that identifies sentences containing dependence relations, then identifies the entities (i.e., dependence relation variables) in the sentence using conditional random fields (CRF) [2]. Dependence relations are constructed from the so-obtained entities using a set of rules, which accurately identify correct dependence relations when given the correct entities.

Results

We evaluated our pipeline both after entity detection and after relation construction. The performance of our proposed hybrid approach consistently outperformed a strictly rule-based baseline that uses shallow syntactic parsing. For entity detection, we achieved an F_1 score of 67.1 vs. the baseline performance of 62.6. Likewise, for relation construction using the entities we extracted, we reached an F_1 score of 53.6 vs. a baseline F_1 of 49.1. The results showed that our approach has higher precision and lower recall across the various steps of the pipeline, which is preferred for our particular use case.

In our error analysis of the entity detection we found a number of false negative errors where correct entities were not detected. Sometimes the CRF classifier also had difficulty detecting the correct entity boundaries. These problems were due in a large part to the heterogeneous nature of the entities and the modest size of our annotated corpus.

Conclusion

Our work investigates how to automatically extract dependence information from the academic medical literature, thereby streamlining an otherwise time-consuming and costly process. We developed a set of NLP steps to address this issue and found that our pipeline of algorithms performed satisfactorily especially compared to a naïve baseline.

Acknowledgments

This project was supported in part through a grant from the Irish Innovation Development Agency (ref. 133954, September 2009) and Science Foundation Ireland (13/IF/I26291).

References

[1] Pauker SG, Wong JB. The influence of influence diagrams in medicine. Decision Analysis 2005: 2: 238–244

[2] Lafferty JD, McCallum A, and Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, 2001: 282–289.

Address for correspondence

Charles Jochim – charlesj@ie.ibm.co