Reflective Random Indexing to Develop a Medication-Problem Knowledge Base

Safa Fathiamini, MD, MS, Trevor Cohen, MD, PhD, Allison B. McCoy, PhD, Dean F. Sittig, PhD

The National Center for Cognitive Decision Making in Healthcare and
The School of Biomedical Informatics |The University of Texas Health Science Center at Houston

UTHealth
The University of Texas Health Science Center at Houston
School of Biomedical Informatics

SHARPC NCCD

Introduction

Clinical problems in patient medical records:

- Are important for diagnostic reasoning, indicators of patient's condition, and often referenced in communications and consultations [1].
- Failure to properly present them to the clinician can result in suboptimal decision making [2].
- Maintenance of problem-medication knowledge bases are often an overwhelming task, and automated alternatives are desirable [1].

Aim: To automatically identify medication-problem relationships from medically relevant free text.

Methods

Distributional semantics (DS) [3]**:**

- Draws on the distribution of terms across different context in a corpus of free text.
- Derives meaningful associations between biomedical terms and concepts.

Reflective Random Indexing (RRI) [4]**:**

- A DS technique that identifies relationships between terms that do not directly co-occur
- Particularly useful in literature based discovery.

Approach: We developed a gold standard (explained later) that consisted of medication-problem pairs. For each drug, a list of associated problems were extracted from the corpora, using RRI, and the problems that were not in the gold standard were excluded from the results. We then sorted the problems based on their degree of association with each drug, and set different thresholds to calculate the sensitivity and specificity of the retrieved result set. We assumed that no clinically meaningful relationship existed between drugs and problems in the gold standard that were not explicitly stated as being related to one another.

Materials

Three corpora, with the total number of terms shown, were used: UpToDate medical knowledge base (UD, 280,000), Medline abstracts (ML, 27,000,000), and clinical notes from an electronic medical record system (CN, 3,000,000).

Standard Selection and Preparation

We developed a gold standard using an expert-reviewed set of medication-problem pairs. Two medically trained investigators reviewed the accuracy of 6493 medication-problem pairs linked during e-prescribing by 60 randomly selected clinicians. The reviewers manually annotated the pairs to indicate whether the problem in each pair was or was not related to the medication. We then adapted MetaMap [5] and a customized post-processing to extract the main term for each medication and problem. (Figure 1)

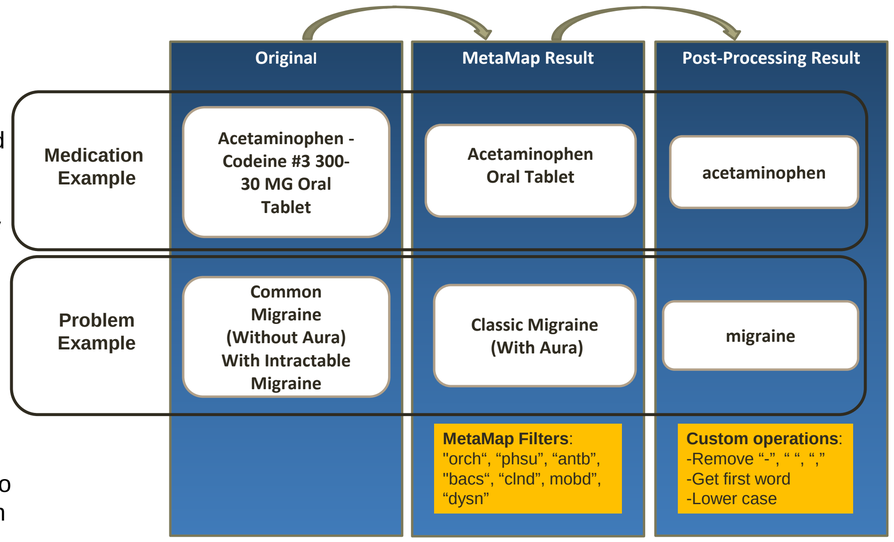
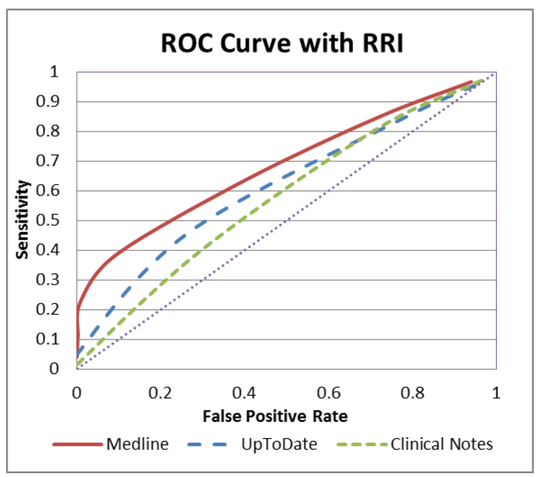


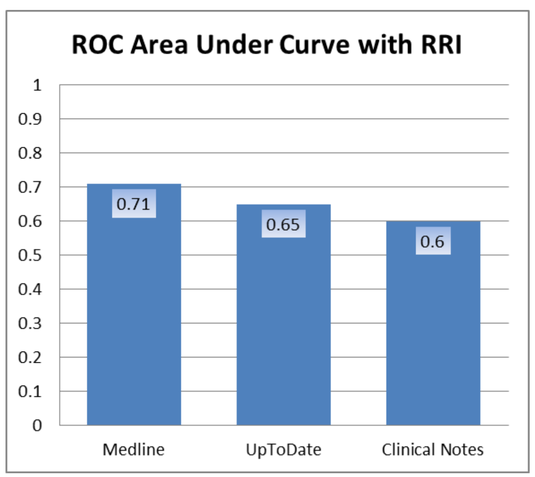
Figure 1. Examples of terms and processes used in the gold standard preparation.

Results

To evaluate RRI, we generated receiver operating curves (ROC) (Figure 2.a), and calculated the area under curve (AUC) to be 0.71 for ML, 0.65 for UD, and 0.6 for CN (Figure 2.b).



a. Receiver operating curve for the three corpora



b. Area under curve comparison for the three corpora

Figure 2. Receiver operating curve (ROC), and area under curve (AUC) for RRI with different corpora.

Summary of Conclusions

- RRI is a DS technique that provides a way to extract meaningful information and relationship from any corpus of data.
- In this research, RRI was applied to a range of different corpora, for the purpose of medication-problem pair extraction.
- Contrary to expectations, Medline outperformed UpToDate, potentially due to Medline's much larger size.
- We demonstrated that RRI can be used to identify medication problems relationships from medically relevant free text.

Limitations and Future Directions

- Determine the effect of size and type of corpus on the performance of the model for this particular task.
- Determine how RRI would compare against more established DS methods like Latent Semantic Analysis (LSA).
- The results would have been better if we had a complete medication-problem test set. All relationships between potential drug-problem pairs in the set were not explicitly annotated.

References

1. McCoy A, Wright A, Laxmisan A, Ottosen MJ, McCoy J, Butten D, et al. Development and Evaluation of a Crowdsourcing Methodology for Knowledge Base Construction: Identifying Relationships between Clinical Problems and Medications. J Am Med Inform Assoc. 2012 Sep 1;19(5):713–8.
2. Gandhi TK., Zuccotti G, Lee TH. Incomplete care—on the trail of flaws in the system. New England Journal of Medicine. 2011(6) 365: 486-488.
3. Cohen T, Widdows D. Empirical Distributional Semantics: Methods and Biomedical Applications. J Biomed Inform. 2009 April ; 42(2): 390–405.
4. Cohen T, Schvaneveld R, Widdows D. Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. Journal of Biomedical Informatics. April 2010;43(2):240-256.
5. MetaMap Portal. Available at: http://mmtx.nlm.nih.gov/. Accessed August 1, 2012.

Acknowledgements

This project was supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects (SHARP) from the Office of the National Coordinator for Health Information Technology.

Please contact: safa.fathiamini@uth.tmc.edu