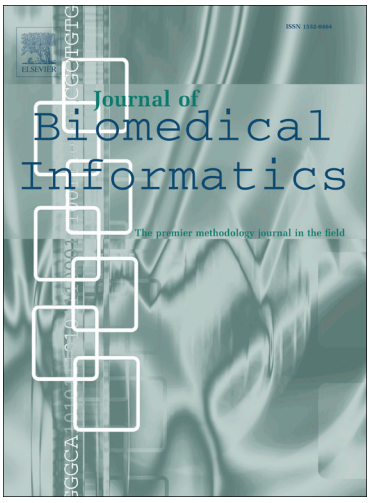

Automatic Information Extraction from Unstructured Mammography Reports Using Distributed Semantics

Anupama Gupta, Imon Banerjee, Daniel L. Rubin

PII: S1532-0464(17)30290-3
DOI: https://doi.org/10.1016/j.jbi.2017.12.016
Reference: YJBIN 2911

To appear in: *Journal of Biomedical Informatics*

Received Date: 31 August 2017
Accepted Date: 30 December 2017

Please cite this article as: Gupta, A., Banerjee, I., Rubin, D.L., Automatic Information Extraction from Unstructured Mammography Reports Using Distributed Semantics, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.12.016

Automatic Information Extraction from Unstructured Mammography Reports Using Distributed Semantics

Anupama Gupta[a], Imon Banerjee[b], Daniel L. Rubin[c,b]

[a] *Department of Computer Science, Columbia University*
New York City, NY, USA
[b] *Department of Biomedical Data Science, Stanford University School of Medicine*
Stanford, CA, USA
[c] *Department of Radiology, Stanford University School of Medicine*
Stanford, CA, USA

Abstract

To date, the methods developed for automated extraction of information from radiology reports are mainly rule-based or dictionary-based, and, therefore, require substantial manual effort to build these systems. Recent efforts to develop automated systems for entity detection have been undertaken, but little work has been done to automatically extract relations and their associated named entities in narrative radiology reports that have comparable accuracy to rule-based methods. Our goal is to extract relations in a unsupervised way from radiology reports without specifying prior domain knowledge. We propose a hybrid approach for information extraction that combines dependency-based parse tree with distributed semantics for generating structured information frames about particular findings/abnormalities from the free-text mammography reports. The proposed IE system obtains a F_1-score of 0.94 in terms of completeness of the content in the information frames, which outperforms a state-of-the-art rule-based system in this domain by a significant margin. The proposed system can be leveraged in a variety of applications, such as decision support and information retrieval, and may also easily scale to other radiology domains, since there is no need to tune the system with hand-crafted information extraction rules.

Keywords – Information extraction, word embedding, report annotation, information frames.