Annotation for Information Extraction from Mammography Reports

Selen BOZKURT[a], Kemal Hakan GULKESEN[a,1] and Daniel RUBIN[b]
[a] *Biostatistics and Medical Informatics, Akdeniz University, Antalya, Turkey*
[b] *Department of Radiology, Stanford University, Stanford, CA,USA*

Abstract. Inter and intra-observer variability in mammographic interpretation is a challenging problem, and decision support systems (DSS) may be helpful to reduce variation in practice. Since radiology reports are created as unstructured text reports, Natural language processing (NLP) techniques are needed to extract structured information from reports in order to provide the inputs to DSS. Before creating NLP systems, producing high quality annotated data set is essential. The goal of this project is to develop an annotation schema to guide the information extraction tasks needed from free-text mammography reports.

Keywords. mammography, decision support systems, natural language processing

Introduction

Variability in mammographic interpretation is a substantial problem [1,2], and methods to improve mammographic interpretation are needed [1,3]. The American College of Radiology (ACR) developed Breast Imaging-Reporting and Data System (BI-RADS), a reporting system developed for breast imaging [4]. Likewise, the Radiological Society of North America (RSNA) developed The RadLex® vocabulary, which is intended to reduce variation and improve clarity in radiology reports and image annotations [5]. While adoption of BI-RADS or RadLex® can reduce the variation in the language of mammography reporting, it does not solve the problem of variation in decision making. To reduce variation in decision making, decision support systems (DSS) are advocated. Since radiology reports are created as unstructured text reports, Natural language processing (NLP) techniques are needed to extract structured information from reports in order to integrate DSS into the radiology reporting workflow.

Before creating an NLP system, it is necessary to create a training corpus of radiology reports, in which the named entities and relations that the NLP system will extract are annotated. Accordingly, an annotation schema is needed to clarify the information requirements of the text processing task and the domain of interest [6].

In this work, we created an annotation schema to focus and clarify the requirements of information extraction in mammography reports. With the help of the annotation schema, named entities of interest in mammography reports needed as inputs to DSS were annotated manually and an initial corpus of 35 radiology reports. Our work provides an initial gold standard for evaluation during the development steps

[1] Corresponding Author: Kemal Hakan Gulkesen, e-mail: hgulkesen@gmail.com

of NLP systems. We also describe a pipeline for automatic annotation of those critical entities and conducted an initial evaluation.

1. Method

We used GATE— an open source architecture for language engineering [7]—for manual annotation of texts, and it also provides an NLP development platform. To develop preliminary corpus, we collected 35 free-text mammography reports (five reports for each of the seven BIRADS categories).

We created an annotation schema, comprising the set of named entities critical in mammography reporting (Figure 1). Using this schema, we annotated section headers and the sentences in each section of the radiology reports, and then we focused on "findings" section and defined two entity types as "Anatomic Entity" and "Imaging Observation". We also defined modifiers of the named entities reported in findings section (Table 1).

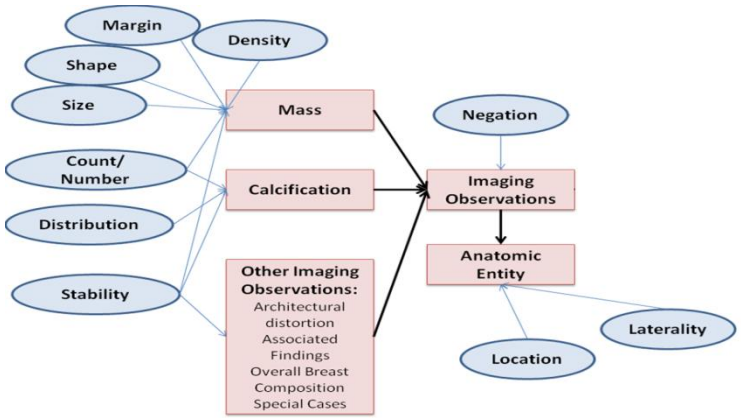

Figure 1. Annotation schema: Rectangles: entities; ovals: modifiers; solid lines: relationships.

Table 1. Some Entity Types and Modifiers in Findings section of a Mammography report.

Entity Type	Description	Example
Negation signal (modifier)	Relates a condition to its negation or uncertainty about it	no focal dominant mass
Margin signal (modifier)	Relates an Imaging Observation to information about the margin.	spiculated mass
Shape signal (modifier)	Relates an Imaging Observation to information about the shape.	irregularly-shaped mass
Density signal (modifier)	Relates an Imaging Observation or to information about the density.	breast tissue is largely fatty
Stability signal (modifier)	Relates an Imaging Observation to information about the stability.	stable focal asymmetric density

Based on BIRADS, we constructed an ontology and annotated entities as concepts described in 2010 i2b2 Concept Annotation Guidelines with their semantic types [8]. Figure 1 also shows the entities, modifiers and their basic relations.

After manual annotation, we built a pipeline in GATE to perform named entity recognition in mammography report texts. In our pipeline, sequential processing is performed to accomplish the following tasks: (1) tokenization of words and punctuation; (2) annotation of the sections of the mammography report (so that we can recognize the Findings section which is most pertinent to our work); (3) annotation of each sentence in a section; (4) annotation of terms using BI-RADS Onto-Gazetteer; (5) identification of concepts.

2. Results

The sections and sentences within all 35 reports were all correctly classified. Based on the list of concepts in Table 1, the "Anatomic Entities", "Imaging Observations" without their modifiers were correctly identified in all reports.

As described in the i2b2 guideline, only complete noun phrases (NPs) and adjective phrases (APs) were marked as annotation. For example in "no focal dominant mass", "focal dominant mass" marked as Imaging Observation entity without negation signal, likewise, "right breast" was marked as Anatomic Entity with its laterality signal.

3. Conclusion

Our preliminary study aims to provide an annotation schema based on BI-RADS terminology for attempting to extract key information from mammography reports needed to input into a decision support model, specifically anatomic entities and imaging observations obtained from specific sections of text. We will be developing a NLP pipeline to extract unique imaging observations, considering co-references and relationships among entities, and will incorporate our NLP into a real-time DSS integrated with reporting. We will also be conducting an evaluation in a larger corpus of reports.

References

[1] J. G. Elmore et al., Variability in radiologists' interpretations of mammograms, *N Engl J Med* **331** (1994), 1493-9.

[2] R. Smith-Bindman et al., Physician predictors of mammographic accuracy, *J Natl Cancer Inst* **97** (2005), 358-67.

[3] K. Kerlikowske et al., Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System, *J Natl Cancer Inst* **90** (1998), 1801-9.

[4] E.S. Burnside et al., The ACR BI-RADS experience: learning from history. *J Am Coll Radiol* **6** (2009), 851-60.

[5] C.P. Langlotz, RadLex: A new method for indexing online educational materials (vol 26, pg 1595, 2006), *Radiographics* **27** (2007).

[6] A. Roberts et al., Building a semantically annotated corpus of clinical texts. *J Biomed Inform* **42** (2009), 950-66.

[7] H. Cunningham, GATE, a general architecture for text engineering. *Computers and the Humanities* **36**, (2002), 223-254.

[8] O. Uzuner et al., 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* **18** (2011), 552-6.