Journal of Biomedical Informatics 78 (2018) 78–86

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# 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

ARTICLE INFO

*Keywords:*
Information extraction
Word embedding
Report annotation
Information frames

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.

## 1. Introduction

An enormous amount of electronic information is generated in the major medical centers in the form of unstructured free text clinical reports [1,2]. Clinical reports are a rich resource that can support machine learning advancement since they contain a large amount of expert-defined data that can be used to train these methods. However, the unstructured representation of the clinical information makes it difficult to use these data as the input to computerized systems, such as Clinical Decision Support (CDS) systems and Information Retrieval (IR) applications. Thus, there is substantial interest in developing natural language processing (NLP) systems that can extract structured information from free text clinical narratives.

Automatic extraction of structured information from mammography reports can be separated into two inter-dependent sequential tasks. First task is to identify the named entities of interest and their relations, where the named entities are information items of particular types, such as patient name, date of exam, imaging observations, and diseases. Relations connect one or more named entities to make a factual assertion ("statement") about them, e.g., existence of an imaging observation, presence of a characteristic shape, comparison with previous studies, and association of a targeted imaging observation with other imaging observations. In the case of mammography reports, the named entities are mainly imaging observations/abnormalities, such as characterization of masses and locations. The relations associated with these named entities can be extracted to allow formulating a triplet "subject-action-object" for performing computerized reasoning directly on the extracted information and generate high-level knowledge. Fig. 1 shows examples of few sentences in a mammography report where the relations of interest are shown as underlined texts and the cells separate the set of relations that convey similar semantic knowledge about the respective imaging observations.

The second task is to generate "information frames" from the extracted named entities and relations. An information frame is a collection of statements made about a particular abnormality in the radiology report. In mammography reports, the information frames comprise entities and relations that capture imaging observations about breast lesions, as specified by the BIRADS terminology [3]. For example, an information frame describing a mass lesion on the mammography reports would contain statements about mass size and volume, its margin and shape. It can also represents the asserted facts pertaining to the image, such as a mass is located in the upper outer quadrant of

* Corresponding author.
*E-mail address:* ag3900@columbia.edu (A. Gupta).

https://doi.org/10.1016/j.jbi.2017.12.016
Received 31 August 2017; Received in revised form 29 November 2017; Accepted 30 December 2017
Available online 09 January 2018
1532-0464/ 

| |
|---|
| • **Compared to previous films this focal asymmetric density is less defined.**<br>• **There is a 3.5 cm round mass with a partially obscured margin in the right breast.**<br>• **Compared to previous films this mass is increased in size and more defined.** |
| • Several small axillary lymph nodes **are again noted** on the right.<br>• No significant change **is appreciated** in the asymmetric tissue.<br>• Several small right axillary nodes **are present**.<br>• A small asymmetric density **is identified** posteriorly within the right breast. |
| • There is a benign focal asymmetric density in the right breast in the posterior depth which **represents** a post-surgical scar.<br>• Examination **indicates** an intramammary lymph node in the right.<br>• This **may represent** the early calcifications within a degenerating fibroadenoma.<br>• This mass appears to **correspond to** the known carcinoma as demonstrated by prior biopsy.<br>• It **may represent** a mass or an area of overlapping glandular tissue. |
| • Compared to prior exam this calcification region is **not significantly changed**.<br>• Compared to previous films this mass **is slightly decreased** in size.<br>• The number of calcifications and shape are **not changed**. |

**Fig. 1.** Example sentences in different mammography reports. Relations are underlined and each cell represents semantically similar relations.

the left breast or the mass has spiculated shape.

We demonstrate that it is possible to automatically extract relations and their associated named entities using automated clustering of similar relations in mammography reports, without the need for any labeled data. Our work has two main contributions: (i) identify relations in free-text mammography reports using a statistical model without any labeled data, and group semantically similar relations using distributional semantics; (ii) use the extracted named entities and relations to obtain information frames for each of the imaging observations described in the reports. The strategy relies on a tight integration between standard NLP and unsupervised machine learning techniques, which offers an efficient trade-off between required manual effort and generalizability of the system. The general flow of information and processing stages are shown in Fig. 2. The organization of the rest of the paper is as follows. Section 3 includes a detailed description of our method and presents the most interesting aspects of all processing phases. Results are summarized in Section 4, while Section 5 contains a discussion of the results. Section 6 concludes the paper.

## 2. Related works

Many clinical natural language processing(NLP) systems have been developed to generate structured information from unstructured free text. However, most of the existing information extraction (IE) methods targeting mammography reports are either dictionary-based or rule-based and require the targeted types of relations and the synonymous relations to be pre-defined [1,2,4–6]. For example, Nassif et al. [7] developed a dictionary-based system to detect BI-RADS concepts in mammography reports. Similarly, the MedLEE processor [8,9] detects BI-RADS concepts in clinical reports by identifying the structure of semantic grammar in the text and generalizing the semantic terms using a controlled lexicon. Recently, Sevenster et al. [10] used MedLee to correlate clinical findings and body locations in radiology reports. These approaches are extremely time consuming result in a low recall as the manual rules or patterns are required to be pre-defined by a domain expert.

Bozkurt et al. [4] developed a rule based NLP system to extract information frames for each lesion described in narrative

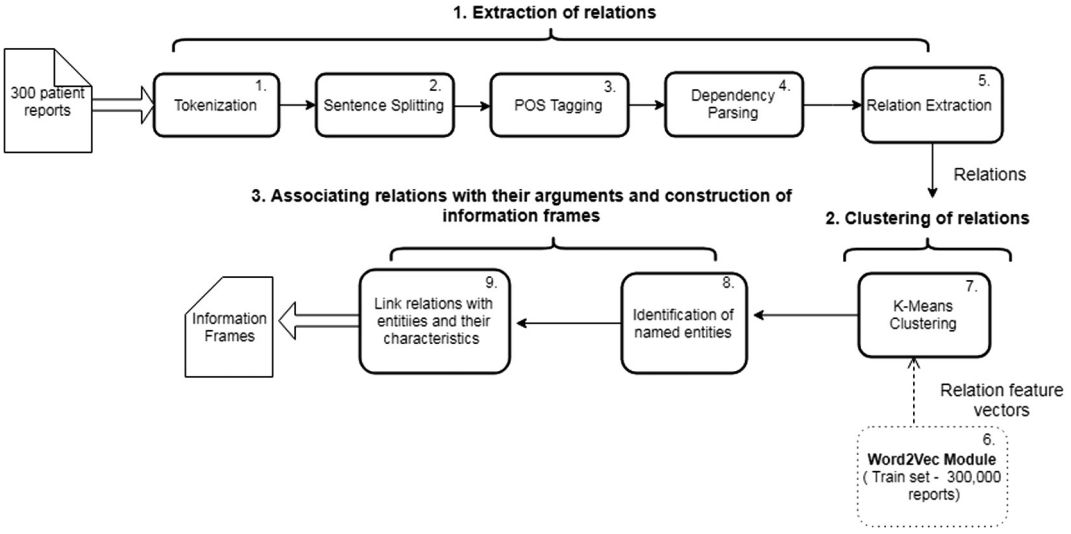


**Fig. 2.** Pipeline to produce structured information frames from free text mammography reports.

mammography reports, including the abnormalities and characteristics of those abnormalities. The main limitation of the Bozkurt's system and similar rule-based systems is that all the kinds of entities and relations need to be pre-specified. It requires substantial manual tuning to extract information, and may fail to extract information for rules that were not pre-specified. For instance, Bozkurt's system in its default configuration did not extract important relations that denote pertinent information about lesions, such as their stability over time (how the lesion has changed as compared to a previous observation or if this lesion is a new observation in the patient mammography), and possible relationships to another important entity (such as a lesion representing either a post-surgical scar or cancer). Moreover, it requires much manual work to build such systems if the targeted number of entities and relations is extensive, and these systems may have limited reusability. Thus, an interesting research problem is – *how to extract information from mammography reports without pre-specified rules or lexicon?*

In addition to rule-based systems, prior works employed supervised learning for automatic extraction of relations and information frames. Taira et al. [5] developed an NLP processor to automatically extract relations from radiology reports. Brunside et al. [11] proposed a statistical method for mapping radiology reports to BI-RADS (Breast Imaging Reporting and Data System) terms. The relations, denoted by single or multiple words, were extracted using hand-tagged training examples. Since there is no standardized lexicon of all the different words or combinations of words that can represents potential relations, it is an impractical task to tag examples of every relation of interest. Besides, the type of relations that can be encountered in free text reports are difficult to know in advance, which makes the task even more challenging.

According to a recent study [12] that reviewed the existing NLP solutions for general medicine, there is an emerging trend of hybrid systems, where rule-based NLP systems have been the most commonly used followed by supervised machine learning approaches. Several hybrid and pure machine learning based methods have been proposed that obtained good performance in clinical information retrieval [13–17]. However, building these supervised systems requires large amounts of hand-annotated data, which is very tedious and time-consuming to produce. For instance, recently, Hassanpour et al. developed an IE system for chest CT reports and used conditional Markov field and conditional random field models for extracting radiological observation from the reports [15]. The IE system required the list of concept classes to be specified by manual annotators. However, obtaining an exhaustive list of concepts in the domain of interest is also an equally challenging task.

## 3. Materials and methods

### 3.1. Algorithm overview

In our unsupervised approach, we first discover all relations from the text reports, then we identify synonyms of the relations to identify semantically unique relations, and finally we apply this to new text to automatically extract relations and their associated named entities. Accordingly, our method comprises three sequentially executed processing modules (see Fig. 2). In the following subsections, we detail the each core processing modules, and describe the datasets on which we train and evaluate the system.

***1. Extraction of relations*** – We define a "relation" to be a word or a group of words in a sentence that convey meaningful information about two related named entities in the text. Following the definition, the relations can capture the presence of an imaging observation (e.g.,"there is", "is present", "is noted"), can express how something affects an observation (e.g., indicates, could obscure, etc.) or a change in characteristics of an observation (e.g., "is unchanged", "is increased", "significantly changed", etc.). We assume that a relation is either a verb or a phrase containing a verb such that each constituent word has at least one syntactic dependency with another word in the same phrase. Syntactic dependency is the grammatical relationship that exists between two words, represented as a directed labeled link in the parse tree structure of a sentence (Fig. 3). These dependencies are defined as triplets: *head-dependency-dependent*. In Fig. 3, an example of such relations is – "Asymmetric breast tissue is again noted in the lateral left breast." The phrase "is again noted" is a relation denoting the presence of an imaging observation ("Asymmetric breast tissue") in a particular location ("lateral left breast"). Similarly, in the sentence – "Compared to previous films this nodular density is more defined.", the phrase "compared to" denotes a comparison of the nodular density between current and the previous films, and the phrase "is more defined" denotes a characteristic change in the nodular density. Additionally, the dependency (shown in Fig. 3) between the words 'previously' and 'described' will be expressed as *advmod* (described, previously), which means that 'previously' is an adverb modifier of the term 'described'.

In order to extract the targeted "relations", we constitute a pipeline with five processing blocks where the first four blocks are performing standard NLP tasks using the Stanford CoreNLP toolkit [18]: reports tokenization (step 1), sentences splitting (step 2), part of speech tagged (POS) (step 3), and dependency parsing (step 4). The verbs detected by the POS tagger constitute an initial list of relations (named as "seed list"). The relation phrases are formed from the seed list, exploiting the outcome of the dependency parser on the respective sentence (Fig. 3). The links between the words denote the grammatical relationship, and these links can be exploited to extract meaningful multiword relation units. The final block (step 5) is coded based on the following grammatical rules, where multiword units or relation phrases are formed by combining the seed list's terms and their neighboring words:

**Rule 1**: *For each term in the seed list, the immediate next or previous word is concatenated with the term if there exists a dependency relation between the word and the term and the POS tag of the word does not fall into any of the four categories, that is, NN (noun), CD (cardinal number), CC (coordinating conjunction) and DT (determiner). The resultant concatenated term is added to the seed list.*

**Rule 2**: *The two or more phrases obtained in Rule 1 is added to the seed list if they are contiguous sequence of words in the same sentence.*

As an example, consider the sentence shown in Fig. 3: "The previously described mass inferiorly within the right breast is less discretely seen on the current exam." The initial seed list for this sentence consists of the terms "described" and "seen". These terms are combined with their neighboring words according to the constraints described in *Rule 1*. Specifically, "previously" is combined with "described" to form the relation phrase "previously described". Similarly, "discretely" and "seen" combine to form "discretely seen", which further combines with "less" to form "less discretely seen" (according to *Rule 2*) as the final relation phrase that is extracted by the relation extraction pipeline.

***2. Clustering of relations*** – After extracting relations from the text, the next step is to group the similar relations together based on their semantic relatedness to discover the semantically distinct (or canonical) relations of interest in the patient reports, which we refer as "clustering of relations". The intuition behind grouping similar relations together is

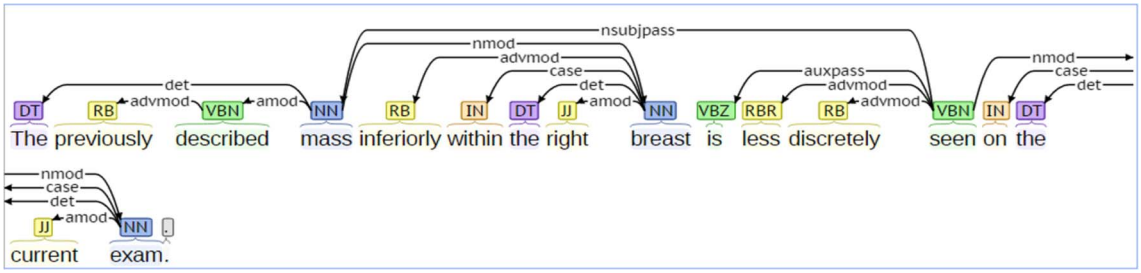


**Fig. 3.** Dependency parse diagram of an example sentence produced using Stanford CoreNLP toolkit, which shows dependency relations between words represented by arcs. An arc from word A (head) to word B (dependent) indicates that word A influences word B.

**Table 1**
Each column represents a separate cluster and the rows represent the relations belonging to the respective cluster.

| Cluster 1: 'Presence of abnormality' | Cluster 2: 'Stability over time' | Cluster 3: 'Removal of abnormality' | Cluster 4: 'Similarity with other observation' | Cluster 5: 'Characteristic of abnormality' | Cluster 6: 'Action taken' |
|---|---|---|---|---|---|
| seen | changed | removed | represents | defined | dated |
| are present | significantly changed | completely removed | indicates | obscured | palpated |
| appearing | decreased | has been removed | corresponds | circumscribed | biopsied |
| demonstrated | increased | been removed | related | more defined | diagnosed |
| identified | are unchanged | have been removed | likely represents | less defined | performed |
| are noted | not changed | | could obscure | partially obscured | stopped |
| are new | persists | | degenerating | partially circumscribed | recommended |
| again noted | improved | | may represent | | guided |

to frame related information of an observation in the report more meaningfully in the final information frames generated by the IE system. For example, in Fig. 5, based on the relation "there is" that denotes presence of abnormality (refer Table 1) the information about the abnormality "architectural distortion" is generated and displayed in the Frame 1. Similarly, Frame 2 groups together the information about associated findings of the abnormality based on the similar relations ("with associated" and "associated with") that link the findings with the abnormality.

In order to cluster the relations, we learn the vector representation of the relation phrases using the unsupervised word2vec algorithm [19] (step 6), and then generate clusters using k-means clustering [20] (step 7). The word2vec is a neural embedding model, which depends on the assumption that the meaning of a word can be inferred from the distribution of words within a predefined window around the word. The word2vec model represents each word in a vocabulary as a vector of floating numbers (or "word embeddings") by learning how to predict a 'key word' given the neighboring words. The word embeddings learned on a large text corpus are typically good at representing semantic similarity between words that are semantically similar, since such words often occur in similar context in the text. Thus, word embeddings often can be used as features in various machine learning applications.

To train the word2vec model, we sequentially parse the sentences in the corpus in two passes. In the first pass, the frequency counts of the sentence tokens are collected and stored in a dictionary data structure. In the second pass, a neural network model is trained to learn the vector representations of the tokens. We explored the following configurations of the word2vec model: (i) type of architecture: continuous bag-of-words and skipgram; (ii) the dimension of word vectors: {50, 100, 200, 300, 400, and 500}; (iii) the size of the context window: {3, 5, 10 and 20}. On the training dataset (see below), the optimized performance achieved with the skip-gram architecture, vector size of 300 and a window size of 5.[1]

In the word2vec embeddings space, semantically related word vectors should appear in a close vicinity. Thus, we can perform unsupervised space clustering using k-means algorithm to find similar words. However, word2vec only creates embedding of single words, not the multi-word terms. In order to perform clustering of multi-word relation phrases, we need to create a representation of the relation vectors (which we refer to as "relation phrase embeddings"). The most common method for combining the vectors of a multi word term is averaging. Using the vector averaging method, we take the average of all the constituent individual word vectors of the respective relation phrase to obtain the final embedding as: $V(w_1,w_2,...,w_n) = \frac{V(w_1)+V(w_2)+\cdots+V(w_n)}{n}$, where, $V(w_1,w_2,...,w_n)$ denotes the "relation phrase embeddings" and $w_1,w_2,...,w_n$ denotes the words that formed the phrase. After obtaining the relation phrase embeddings, we use k-means++ algorithm [20] to cluster the relations phrase embedding space. We use this version of k-means because the accuracy of the standard k-means are very sensitive to the choice of initial cluster centers (seeds), whereas the k-means++ chooses the initial cluster centers using a randomized seeding technique before proceeding with optimization. In order to determine the optimum value of k, we use the elbow method [21]. After obtaining the clusters of relations, we label the clusters to indicate the underlying concept of the relations belonging to each cluster. Table 1 lists the relations classified into the six clusters using k-means++. With the help of domain experts, we manually reviewed the relations belonging to each cluster to infer the cluster topic based on the most important terms/relations that represent the cluster.

***3. Associating relations and extracting information frames*** – Finally, we generated the information frames by linking the extracted relations (from the prior step) with the associated named entities, and entitled the frames on the basis of the class on which the relation belongs (see Table 1). In order to identify the named entities and link them to the relation to which they belong, we used an information extraction (IE) module that scans and identifies all the named entities (step 8, Fig. 2), followed by a filter that preserves only the entities that are arguments to the relations that have been identified in the text (step 9, Fig. 2). For this case-study, named entities are sequences of words that belong to a particular class.

Mammography reporting generally follows a very specific vocabulary and it is important to capture the semantics of these domain-

[1] The trained word embeddings are available at https://github.com/anupama-gupta/word-embeddings-mammography.

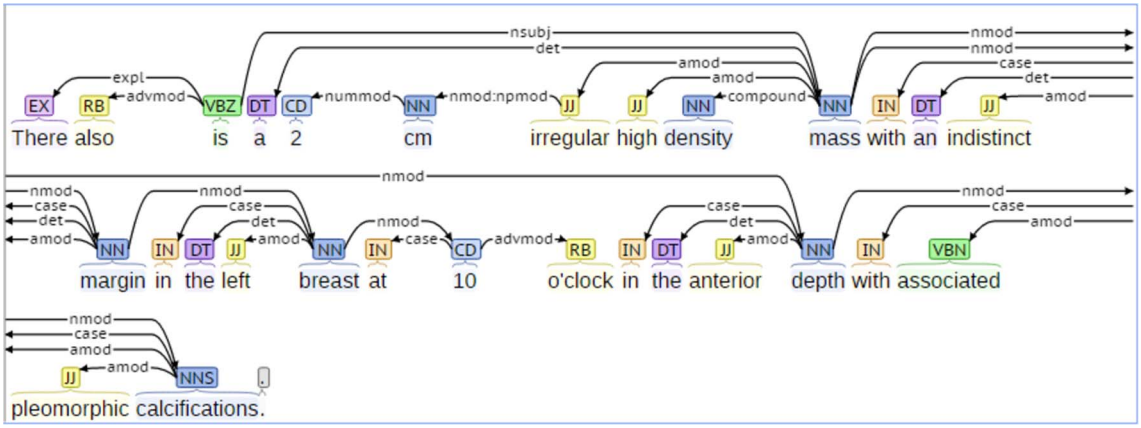


**Fig. 4.** Parse Tree representation of an example sentence obtained using the Stanford CoreNLP.

| **Input: Mammography report** | | |
|---|---|---|
| There is an area of architectural distortion in the left breast 9 o'clock in the anterior depth with associated punctate calcifications which represents a post surgical scar. Associated with this area of architectural distortion is skin thickening. | | |
| **Output: Information Frames** | | |
| **Imaging Observation** | **Frame1** | **Relation** = there is<br>**Observation** = architectural distortion<br>**Laterality** = left<br>**Clock face** = 9<br>**Depth** = anterior |
| | **Frame 2** | **Relation** = with associated<br>**Observation** = calcification<br>**Type** = punctuate<br><br>**Relation** = associated with<br>**Observation** = skin thickening |
| | **Frame 3** | **Relation** = represents<br>**Observation** = post surgical scar |

**Fig. 5.** Generated information frames (grouped by the class of relations that they represent) for an imaging observation present in a mammography report.

specific entities for generating meaningful information frames. Thus, we adopted BI-RADS [5] which mainly represents mammography imaging observations (conditions, problems, etc.), location (laterality, clock face location, etc.) and report entities. We used the BI-RADS terminology to create a mapping between the terms that may appear in mammography reports, and their respective classes. The information extraction module exploits this ontology to annotate the named entities. For examples, the terms – (mass, density, nodule, calcification) are annotated as *abnormality*, the terms – (skin retraction, skin thickening, fibroadenoma) as *special case*, the terms – (anterior depth, posterior depth, middle depth) as *depth*, and the terms – (clustered, coarse, diffuse, linear) as *type*.

The final module filters the named entities that are candidate arguments of the relations (step 9, Fig. 2) and associates them with the correct relation. The module first obtains the dependency tree (Fig. 4) for each sentence using the Stanford Dependency Parser and then applies an algorithm to the named entities in the parse tree (pseudo shown in Algorithm 1) to associate the named entities to their corresponding relations. We assume that the relations are binary which means relating at most two arguments.

We illustrate the algorithm using the example of the first sentence in Fig. 4. After step 7, 'mass' is added to the subject list corresponding to the relation 'also is'. At the end of step 11, we obtain all the attributes ⟨2 cm, irregular, high density, indistinct margin⟩ as well the location descriptors ⟨left breast, 10 o'clock, anterior depth⟩ of the 'mass'. Finally, the information frames are generated by listing all the imaging observations present in the reports along with their relationships with other entities. In Fig. 5, we present an information frame generated by the Algorithm 1 that contains the relations and their associated named entities.

**Algorithm 1.** Algorithm to identify arguments of the relations

1: **procedure** ARGUMENTSIDENTIFIER
2: $r \leftarrow$ list of extracted relations
3: Run Stanford Dependency Parser for each sentence and store outcome in tagged *EntityList*
4: Annotate the named entities using BIRADS-Ontology and store outcome in *DicAnnotation*
5: **while** $r[i] \neq null$ **do** ▷For each relation presents
6: **for** `i is in` *EntityList* **do**
7: $r[i].$ *namedentities* $\leftarrow$ *EntityList*$[r[i]]$.tagged[subj or dobj] ▷construct a entity list that are linked to $r[i]$ with a nominal subject (subj) or direct object (dobj)
8: **end for**
9: **for** $j$ is in $r[i].$ *namedentities* & stop word is not encountered **do**
10: $r[i].$ *modifiers* $\leftarrow$ DicAnnotation $[r[i].$ *namedentites*$[j]]$.annotation["location" or "characteristic"] ▷obtain a list of modifiers linked with $r[i].$ *namedentities*$[j]$ by a syntactic dependency.
11: **end for**
12: *InformationFrame*$[r[i]] \leftarrow r[i].$ *namedentities* and $r[i].$ *modifiers*
13: **end while**
14: **end procedure**

### 3.2. Dataset

Under institutional review board approval, we assembled a corpus of 300,000 mammography reports from our healthcare institution's clinical data repository, which contains a variety of breast abnormalities. We used this corpus of 300,000 un-curated reports to train the word2vec model. We pre-processed the corpus of 300,000 mammography reports for training the word2vec model – (1) punctuations are removed and '.' is replaced by ' dot' from the text so that the same terms occurring in the forms 'cyst.' and 'cyst' are not treated as different terms; (2) all the words are converted to lower case form so that the terms such as 'Compared' and 'compared' treated equally.

We also obtained an independent set of 300 manually curated mammography reports that were used in a prior study to evaluate the state-of-the-art system [4]. We used these 300 reports to extract the relations for identifying meaningful clusters and used their annotations as a gold standard to test the information extraction performance of our proposed system.

### 3.3. Evaluation scheme

To evaluate relation clustering, we created a gold standard of optimal relation partitions by manually labeling each relation instance belonging to the 300 curated mammography reports with its correct class label. In total, a set of 162 relation instances were extracted. Our goal is to quantify the degree of agreement between the true and predicted clusters (outcome of Section 3.1 module 2) which measures how closely the unsupervised clustering of the relations coincides with the human expectation of categorizing relations according to the semantics of abnormalities. In order to measure the degree of similarity between the automatically extracted and the manually labeled relations, we used the adjusted Rand index [22]. The Rand index [23] is a commonly used measure of agreement between two partitions which is defined by: $RandIndex = \frac{(TP+TN)}{(TP+FP+FN+TN)}$, where $TP$ and $FP$ are the number of correct and incorrect decisions of assigning a pair of relations to the same class, and $TN$ and $FN$ are the number of correct and incorrect decisions of assigning a pair of relations to different classes. The standard Rand index lies between 0 and 1 and may result in quite large values even when clustering two random partitions. To address this issue we use adjusted Rand index (ARI) [22] which is a better metric as it puts the expected value at 0 for random labeling and 1 when the clusters are identical. Moreover, it also allows measuring agreement even when the partitions compared have different numbers of clusters.

In order to ultimately evaluate the success of information extraction based on our methods, we used our gold standard dataset of 300 mammography reports to determine the actual imaging observations, their characteristics, stability value and relationships with other abnormalities. We applied our automatic system on the test set of 300 mammography reports and compared the results of processing the 300 reports with our system and the rule-based system described by Bozkurt et al. [4] that had been used to evaluate the same reports. During the evaluation, we consider both *complete detection of a lesion* – all the characteristics of the lesion present in the text are detected by the system, and *partial detection of a lesion* – at least one characteristic the detected lesion is either incorrect or missing.

We use the Precision-Recall metrics for the evaluation of the completeness of the information frames. For the complete detection cases, $Precision = \frac{n_c}{n_d}$, where $n_c$ is the total number of lesions detected completely, and $n_d$ is total number of lesions detected, $Recall = \frac{n_c}{nl}$ where $n_l$ is the total number of lesions present in the report, and $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$. For the partial detection cases, $Precision = \frac{n_c + \frac{1}{2}n_p}{n_c + \frac{1}{2}n_p + n_{fp}}$, where $n_c$ is the total number of lesions detected completely, $n_p$ is the total number of lesions detected partially and $n_{fp}$ is total number of falsely detected lesions, $Recall = \frac{n_c + \frac{1}{2}n_p}{n_c + \frac{1}{2}n_p + n_{fn}}$ where $n_{fn}$ is the total number of lesions present in the report that were not detected.

## 4. Results

### 4.1. Validation of the clusters

The set of relations in the final clusters formed are shown in Table 1. We inferred that the optimal number of clusters is 6 from the elbow plot

**Table 2**
Comparison of results from our system based on complete and partial match cases of lesions with an existing information extraction system.

| Methods | Absolute counts | | | Complete match cases results | | | Partial match cases results | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total lesions present | Total lesions detected | Lesion characterized completely | Precision | Recall | F1 Score | Precision | Recall | F1 Score |
| Our method | 797 | 783 | 746 | 0.95 | 0.94 | 0.94 | 0.99 | 0.97 | 0.98 |
| Bozkurt et al. [4] | | 815 | 723 | 0.89 | 0.91 | 0.90 | 0.96 | 0.96 | 0.96 |

**Table 3**
Confusion matrices of lesion detection, where 'n' denotes the total lesions present.

| Our method | | | Bozkurt et al. [4] | | |
|---|---|---|---|---|---|
| n = 797 | Predicted | | n = 797 | Predicted | |
| Observed | TP = 774<br>FP = 9 | FN = 23<br>TN = 0 | Observed | TP = 780<br>FP = 35 | FN = 17<br>TN = 0 |

and reaffirmed the value of k to be 6 on k-means + +. As seen from the table, relations belonging to the first cluster are: ⟨ "seen", "are present", "are noted", "identified", "are new" ⟩, and, based on the semantic nature of these relations, we concluded that the corresponding cluster denotes "presence of an abnormality". Similarly, the rest of the clusters (from the left to right column in Table 1) denote: "stability over time", "removal of abnormality", "similarity with another observation", "characteristic of abnormality", and "action taken". The ARI (adjusted rand index) for the predicted clusters against the gold standard of clusters was 0.76 which means that there is a high agreement between the automatically extracted and the manually labeled relations.

### 4.2. Evaluation of the information frames

On the 300 test radiology reports, the proposed method was able to automatically extract relations, grouped them according to their semantic similarity, and generated final information frames for the imaging observations present in a report. Our system extracted relation phrases like "there is", "is decreased" and "associated with" to generate the final information frames. The information frames for each imaging observation are grouped by the different types of relations associated with the observation. Table 2 presents side-by-side comparison of the respective output of our proposed system with the state-of-the-art method. We use two different assessment metrics – (1) *Total Lesion Detected* – number of lesions identified automatically from the reports with or without their characterizations, (2) *Lesion characterized completely* – number of lesions identified with all the relevant characteristics.

In Tables 3 and 4, we present the confusion matrix for lesion detection and the complete extraction of breast lesions characteristics, respectively, where 'n' in each cell is the total number of the lesion properties denoted by column name. As seen from the evaluation, the rule-based system against which we compared, is able to detect larger number of lesions than our proposed system (774 by our method and 780 by Bozkurt et al. [4]) while our method outperforms the rule-based system in-terms of more comprehensive characteristics extraction of the lesion, including relationships between findings, descriptions of stability over time, and historical mentions(see Table 4). Our method produced an $F_1$-score of 0.94 in extracting the lesions with all the relevant characteristics pertaining to them. The precision and recall are both higher than that of the existing rule-based system. The higher precision is mainly due to the fact that our system was able to reduce the number of false positives in lesion detection by successfully categorizing previously noted and removed lesions with the help of the extracted relations. The proposed system detected larger number of the relevant characteristics of a given imaging observation mentioned in the gold standard reports and results in a significant number of complete extractions.

## 5. Discussion

In this paper, we describe a novel approach to automated information extraction from narrative mammography reports by identifying relations in the text and grouping them together with respect to their semantic similarity, and finally using the information to automatically extract the entities and relations in the reports and associate

**Table 4**
Results of complete information extraction of breast lesions detected where 'n' denotes total number of the respective lesion properties.

| | | | |
|---|---|---|---|
| **Characteristics (laterality, size, shape, margin, clock position, depth, density)** | | | |
| ***Our method*** | n = 2682 | Predicted | |
| | Observed | TP = 2643<br>FP = 9 | FN = 39<br>TN = N/A |
| **Bozkurt et al.** | n = 2577 | Predicted | |
| | Observed | TP = 2475<br>FP = 3 | FN = 102<br>TN = N/A |
| **Stability** | | | |
| ***Our method*** | n = 516 | Predicted | |
| | Observed | TP = 476<br>FP = 4 | FN = 2<br>TN = N/A |
| ***Bozkurt et al.*** | | N/A | |
| **Relationship with other abnormalities** | | | |
| ***Our method*** | n = 256 | Predicted | |
| | Observed | TP = 232<br>FP = 6 | FN = 24<br>TN = N/A |
| ***Bozkurt et al.*** | | N/A | |

them to produce output information extraction frames. To identify the relations, we generated their word embeddings using a distributional similarity model, word2vec, and then cluster the relations to define information extraction patterns. The clusters are manually labeled to enable us to identify the type of information conveyed by each information frame, such as the characteristics of an abnormality, the presence of an abnormality, its stability value, and its similarity to another observation. Note that the unsupervised clustering approach automatically unifies synonyms in relations which allows to deal with the unseen relations in the execution phase.

Our system successfully extracted more comprehensive information from the reports than the rule-based system described previously by Bozkurt et al. [4], which includes relationships between findings, descriptions of their stability over time, and their historical mentions. For example, in the sentence, "A coarse calcification noted in this region was removed", the system described in Bozkurt extracted the calcification as being present, as their conclusion is based simply on the presence of the abnormality term 'calcification' in the sentence. However, our system extracts the calcification and also links it with the relation "was removed" in the sentence. Hence, we accurately extract the fact that the calcification is no longer present and has been removed.Our system thus is advantageous and may recognize more relations than using hand-crafted rule-based systems. In addition, our distributional semantic approach to relation discovery requires much less hand-curation effort than a rule-based systems to develop.

We also noted that the generic state-of-the-art regular expressions based NLP tools – Negex [24], failed to detect negation of findings on sample sentences from reports – "A coarse calcification noted in this region is no longer seen due to biopsy.", "The nodule in the superior left breast questioned previously is not seen now.". This performance is due to the reason that negation detection techniques require in-domain annotated data examples of negation on the target corpus to yield optimal performance [25]. In the future, we could adopt the distributional relation phrase embeddings to further improve negation detection by automatically learning patterns from clinical text that indicate negation of observations.

Recently, there has been an increasing interest in applying distributional semantic methods for various information extraction tasks in the biomedical domain, such as named entity recognition [26,27],

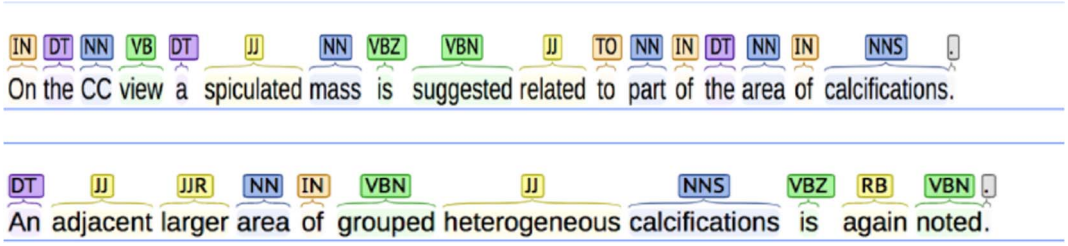


**Fig. 6.** Examples of incorrect pos tagging. In the first sentence 'view' is incorrectly tagged as a verb (correct tag: noun). In the second sentence 'grouped' is incorrectly tagged as a verb (correct tag: adjective).

medical semantics modelling [28], relation between pharmaceuticals and diseases [29], and event extraction [30]. The popularity and advancement of distributional semantics highlights the great potential for accomplishing more advanced biomedical information extraction tasks. A key contribution of our work is the use of distributional semantic features (learned using a predictive neural network model) to cluster the relations extracted from unannotated reports. Moreover, our work is novel in showing the possibility of using these distributional semantic methods to create information extraction systems that can be competitive in performance with rule-based systems.

There are several limitations to our approach. First, the 162 relations extracted in our work are limited in number, which is due to the availability of only a small, expert reviewed dataset of 300 mammography reports used for our gold standard. However, the semantics of these relations seem to be reasonably captured based on a Rand Index score of 0.76. Ultimately in order to show our results are convincing and generalizable, we would need a much larger set of reports and relations.

Secondly, our approach does not extract complex relations or links between different relations. For example, in the sentence "Compared to prior exam this calcification region is not significantly changed", our system extracts the relations "compared to" and "is not significantly changed" separately, and links both of them to the abnormality "calcification". However, both of the relations together actually convey common information about the "calcification". Also, the recall of our method can be improved by introducing more robust rules in identifying imaging observations which share the common name but different characteristics. For example in the sentence "Scattered benign appearing calcifications and vascular calcifications are present in the right breast", our method fails to identify the second set of "vascular calcifications" which are also present in the right breast. Future work can be done to deal with the representation of such complex relation to avoid re-iteration of known information about the observations.

Finally, we have used Stanford POS tagger to extract the relations, which is not always correct in parsing the radiology reports. This is due to that fact that the radiology report narrative style does not always follow the syntactic rules. In the few cases, we found that the POS tagger made mistakes. It was mainly because of the ambiguity of free text narrative style and lexical variations. In Fig. 6, we present an example of incorrect POS tagging that affects the performance of our information extraction system. Future work should attempt to experiment with more precise POS tagging to improve the detection of relations. In the current study, we used simple averaging of word vectors for creating multi-term phrase representations. We are also planing to explore the performance of various vector composition operations (e.g. convolution, max pool, min pool) for relation phase creation [31].

## 6. Conclusion

We have presented a semi-supervised system for information extraction based on distributional semantics and clustering of similar relations, which is followed by the generation of structured information frames from free text mammography reports. The system successfully extracted the relations and recognized their semantic sense. We were also able to extract more accurate information from the reports than a recently proposed rule-based information extraction system, such as the negation and temporality of concepts. Our system performed well in extracting complete information about lesions in mammography reports, with a F-Score of 0.94. In the future, we will work on better feature representation methods for relations, and deal with complex relations to avoid redundancy of information conveyed in the output.

## Conflict of interest

Authors declare no conflict of interest.

## Acknowledgement

This work was supported in part by a grant from the National Cancer Institute, National Institutes of Health, U01CA190214.

## References

[1] F.M. Hall, Language of the radiology report: primer for residents and wayward radiologists, Am. J. Roentgenol. 175 (5) (2000) 1239–1242.
[2] H.J. Tange, H.C. Schouten, A.D. Kester, A. Hasman, The granularity of medical narratives and its effect on the speed and completeness of information retrieval, J. Am. Med. Inform. Assoc. 5 (6) (1998) 571–582.
[3] L. Liberman, J.H. Menell, Breast imaging reporting and data system (bi-rads), Radiol. Clin. 40 (3) (2002) 409–430.
[4] S. Bozkurt, J.A. Lipson, U. Senol, D.L. Rubin, Automatic abstraction of imaging observations with their characteristics from mammography reports, J. Am. Med. Inform. Assoc. 22 (e1) (2014) e81–e92.
[5] R.K. Taira, S.G. Soderland, R.M. Jakobovits, Automatic structuring of radiology free-text reports, Radiographics 21 (1) (2001) 237–245.
[6] I. PAN, Information extraction from mammogram reports, in: KONVENS 2004 Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS) 14–17 September 2004, 2004, p. 113.
[7] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, D. Page, Information extraction for clinical data mining: a mammography case study, in: Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference, IEEE, 2009, pp. 37–42.
[8] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. 11 (5) (2004) 392–402.
[9] D.A. Lindberg, B.L. Humphreys, A.T. McCray, et al., The unified medical language system, IMIA Yearbook (1993) 41–51.
[10] M. Sevenster, R. Van Ommering, Y. Qian, Automatically correlating clinical findings and body locations in radiology reports using medlee, J. Digital Imaging 25 (2) (2012) 240–249.
[11] B. Burnside, H. Strasberg, D. Rubin, Automated indexing of mammography reports using linear least squares fit, in: Proc. of the 14th International Congress and Exhibition on Computer Assisted Radiology and Surgery, 2000, pp. 449–454.
[12] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S.F. Jones, R. Forshee, M. Walderhaug, T. Botsis, Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review, J. Biomed. Inform. 73 (2017) 14–29.
[13] H.-C. Wang, Y.-H. Chen, H.-Y. Kao, S.-J. Tsai, Inference of transcriptional regulatory network by bootstrapping patterns, Bioinformatics 27 (10) (2011) 1422–1428.
[14] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[15] S. Hassanpour, C.P. Langlotz, Information extraction from multi-institutional radiology reports, Artif. Intell. Med. 66 (2016) 29–39.
[16] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufman, 2016.

[17] Y. Xu, K. Hong, J. Tsujii, E.I.-C. Chang, Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries, J. Am. Med. Inform. Assoc. 19 (5) (2012) 824–832.
[18] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit., in: ACL (System Demonstrations), 2014, pp. 55–60.
[19] T. Mikolov, word2vec: Tool for Computing Continuous Distributed Representations of Words, 2016.
[20] D. Arthur, S. Vassilvitskii, k-means + +: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
[21] A. Ng, Clustering with the k-means algorithm, Mach. Learn. (2012).
[22] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[23] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.
[24] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310.
[25] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D. Carrell, C. Clark, Negations not solved: generalizability versus optimizability in clinical natural language processing, PloS One 9 (11) (2014) e112774.
[26] S. Liu, B. Tang, Q. Chen, X. Wang, Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries, Information 6 (4) (2015) 848–865.
[27] B. Tang, H. Cao, X. Wang, Q. Chen, H. Xu, Evaluating Word Representation Features in Biomedical named Entity Recognition Tasks, BioMed Research International 2014.
[28] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, P. Bruza, Medical semantic similarity with a neural language model, in: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, ACM, 2014, pp. 1819–1822.
[29] Z. Jiang, L. Jin, L. Li, M. Qin, C. Qu, J. Zheng, D. Huang, A crd-wel system for chemical-disease relations extraction, in: The Fifth BioCreative Challenge Evaluation Workshop, 2015, pp. 317–326.
[30] C. Li, R. Song, M. Liakata, A. Vlachos, S. Seneff, X. Zhang, Using word embedding for bio-event extraction, in: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), Association for Computational Linguistics, Stroudsburg, PA, 2015, pp. 121–126.
[31] O. Irsoy, C. Cardie, Deep recursive neural networks for compositionality in language, in: Advances in Neural Information Processing Systems, 2014, pp. 2096–2104.