Intelligent Word Embeddings of Free-Text Radiology Reports

Imon Banerjee, Ph.D[1], Sriraman Madhavan, B.E.[1], Roger Eric Goldman, M.D., Ph.D.[1], Daniel L. Rubin, M.D.[1]
[1]Department of Radiology, Stanford University School of Medicine, Stanford, USA

Abstract

Radiology reports are a rich resource for advancing deep learning applications in medicine by leveraging the large volume of data continuously being updated, integrated, and shared. However, there are significant challenges as well, largely due to the ambiguity and subtlety of natural language. We propose a hybrid strategy that combines semantic-dictionary mapping and word2vec modeling for creating dense vector embeddings of free-text radiology reports. Our method leverages the benefits of both semantic-dictionary mapping as well as unsupervised learning. Using the vector representation, we automatically classify the radiology reports into three classes denoting confidence in the diagnosis of intracranial hemorrhage by the interpreting radiologist. We performed experiments with varying hyperparameter settings of the word embeddings and a range of different classifiers. Best performance achieved was a weighted precision of 88% and weighted recall of 90%. Our work offers the potential to leverage unstructured electronic health record data by allowing direct analysis of narrative clinical notes.

1 Introduction

The Picture Archiving and Communication Systems (PACS) stores a wealth of unrealized potential data for the application of deep learning algorithms that require a substantial amount of data to reduce the risk of overfitting. Semantic labeling of data becomes a prerequisite to such applications. Each PACS database serving a major medical center contains millions of imaging studies "labeled" in the form of unstructured free text of the radiology report by the radiologists, physicians trained in medical image interpretation. However, the unstructured free text cannot be directly interpreted by a machine due to the ambiguity and subtlety of natural language and variations among different radiologists and healthcare organizations. Lack of labeled data creates data bottleneck for the application of deep learning methods to medical imaging[1].

In recent years, there is movement towards structured reporting in radiology with the use of standardized terminology[2]. Yet, the majority of radiology reports remain unstructured and use free-form language. To effectively "mine" these large free-text data sets for hypotheses testing, a robust strategy for extracting the necessary information is needed. Methods for structuring and labeling the radiology reports in the PACS may serve to unlock this rich source of medical data.

Extracting insights from free-text radiology reports has been explored in numerous ways. Nguyen et al.[3] combined traditional supervised learning methods with Active Learning for classification of imaging examinations into reportable and non-reportable cancer cases. Dublin et al.[4] and Elkin et al.[5] explored sentence-level medical language analyzers and SNOMED CT-based semantic rules respectively, to identify pneumonia cases from free-text radiological reports. Huang et al.[6] introduced a hybrid approach that combines semantic parsing and regular expression matching for automated negation detection in clinical radiology reports.

In recent years, the word2vec model introduced by Mikolov et al.[7,8] has gained interest in providing semantic word embeddings. One of the biggest problems with word2vec is the inability to handle unknown or out-of-vocabulary (OOV) words and morphologically similar words. The challenge is exacerbated in domains, such as radiology, where synonyms and related words can be used depending on the preferred style of radiologist, and a word may only have been used infrequently in a large corpus. If the word2vec model has not encountered a particular word before, it will be forced to use a random vector, which is generally far from its ideal representation. Thus, we explore how the word2vec model can be combined with the radiology domain-specific semantic mappings in order to create a legitimate vector representation of free-text radiology reports. The application we have explored is the classification of reports by confidence in the diagnosis of intracranial hemorrhage by the interpreting radiologist.

Our two core contributions are:

1. We proposed a hybrid technique for a dense vector representation of individual words of the radiology reports by analyzing 10,000 radiology reports associated with computed tomography (CT) Head imaging studies. *(Word embeddings are publicly released in: https://github.com/imonban/RadiologyReportEmbedding)*

2. Using our methods, we automatically categorized radiology reports according to the likelihood of intracranial hemorrhage.

We derived the word embeddings from large unannotated corpora that were retrieved from PACS (10,000 reports), and the classifiers were trained on a small subset of annotated reports (1,188). The proposed embedding produced high accuracy (88% weighted precision and 90% recall) for automatic multi-class (low, intermediate, high) categorization of free-text radiology reports despite the fact that the reports were generated by numerous radiologists of differing clinical training and experience. We also explored the visualization of vectors in low dimensional space while retaining the local structure of the high-dimensional vectors, to investigate the legitimacy of the semantic and syntactic information of words and documents. In the following sections, we detail the methodology (Sec. 2), present the results (Sec. 3) and finally conclude by mentioning future directions (Sec. 4).

2 Methodology

Figure 1 shows the proposed research framework that comprises five components: Dataset retrieval from PACS, Data Cleaning & Preprocessing, Semantic-dictionary mapping, Word and Report Embedding, and Classification. In the following subsections, we describe each component.

Figure 1: Components of the proposed framework

2.1 Dataset

The dataset consists of the radiology reports associated with all computed tomography (CT) studies of the head located in the PACS database serving of our adult and pediatric hospitals and all affiliated outpatient centers for the year of 2015. Through an internal custom search engine, candidate studies were identified on the PACS server based on imaging exam code. The included study codes captured all CT Head, CT Angiogram Head, and CT Head Perfusion studies. A total of 10,000 radiology reports were identified for this study. In order to provide a gold standard reference for the vector-space embedding algorithm, a subset of 1,188 of the radiologic reports were labeled independently by two radiologists. For each report, the radiologists read the previous interpretation and then graded the confidence of the interpreting physician with respect to the diagnosis of intracranial hemorrhage. For each study, a numeric label was provided on a scale ranging from 1 to 5 with labels as follows: 1) No intracranial hemorrhage; 2) Diagnosis of intracranial hemorrhage unlikely, though cannot be completely excluded; 3) Diagnosis of intracranial hemorrhage possible; 4) Diagnosis of intracranial hemorrhage probable, but not definitive; 5) Definite intracranial hemorrhage. These labels were chosen to reflect heuristics employed by radiologists and treating physicians to interpret the spectrum of information produced by the imaging study.

2.2 Data Cleaning & Preprocessing

All 10,000 radiology reports were transformed through a series of pre-processing steps to truncate the free-text radiology reports and to focus only on the significant concepts, which would enhance the semantic quality of the resulting word embeddings. We developed a python-based text processor - *Report Condenser*, that executes the pre-processing steps sequentially. First, it extracted the *'Findings'* and *'Impressions'* sections from each report that summarizes the CT image interpretation outcome, since our final objective was to classify the reports based on radiological findings.

In the next pre-processing stage, the Report Condenser cleansed the texts by normalizing the texts to lowercase letters and removing words of following types: general stop words, words with very low frequency (<50), unwanted terms and phrases (e.g. medicolegal phrases - *"I have personally reviewed the images for this examination and agreed with the report transcribed above."*, headers - *'FINDINGS', 'IMPRESSION', 'Additional comment'*). These words usually appear either in all the reports or in a very few reports, thus of little or no value in document classification. We used the NLTK library[9] for determining a stop-word list and discarded them during indexing. Examples of the stop-words are: *a, an, are,...,be, by,...,has, he,...,etc*. The Report Condenser also discarded datestamps, timestamps, the radiologist details (e.g. names, contacts) and other recurring phrases in reports. Removal of these terms significantly reduced the number of words that the system had to handle.

Following the removal steps, Report Condenser searched the updated corpus to identify frequently appearing pairs of words based on pre-defined threshold value of occurrence (> 500) and concatenated them into a single word to preserve useful semantic units for further processing. Some examples of the concatenated words are: *'midline shift'* $\rightarrow$ *'midline_shift', 'mass effect'* $\rightarrow$ *'mass_effect', 'focal abnormality'* $\rightarrow$ *'focal_abnormality'*.

In the next step, Report Condenser identified and encoded negation dependencies that appear in the radiology reports via simple string pattern matching. For example, in the phrase *'No acute hemorrhage, infarction, or mass'*, negation is applied to 'acute hemorrhage', 'infarction' as well as 'mass'. Therefore, the Report Condenser encodes the negation dependency as: *'No_acute_hemorrhage', 'No_infarction', 'No_mass'*. Such phrases were identified automatically by analyzing the whole corpus and transformed accordingly.

2.3 Semantic-dictionary mapping

The main idea of the Semantic-dictionary mapping is to use a lexical scanner that recognizes corpus terms which share a common root or stem with pre-defined terminology, and map them to controlled terms. In contrast with traditional NLP approaches, this step does not need any sentence parsing, noun-phrase identification, or co-reference resolution. We used dictionary style string matching where we directly search and replace terms, by referring to the dictionary. We implemented a lexical scanner in python which can handle 1 kilobyte of text per millisecond. On average, the size of each radiology report after cleaning was 1 kilobyte and our scanner took less than 10 seconds to complete the whole mapping process for 10,000 radiology reports. We applied the following two-stage process.

1. *Common terms mapping*: First, we used the more general publicly available CLEVER terminology[10] to replace common analogies/synonyms for creating more semantically structured texts. We focused on the terms that describe family, progress, risk, negation, and punctuations, and normalized them using the formal terms derived from the terminology.

For instance, {*'mother', 'brother', 'wife' ..* } $\rightarrow$ 'FAMILY', {*'no', 'absent', 'adequate to rule her out' ..* } $\rightarrow$ 'NEGEX', {*'suspicion', 'probable', 'possible'* } $\rightarrow$ 'RISK', {*'increase', 'invasive', 'diffuse', ..* } $\rightarrow$ 'QUAL'.

2. *Domain-specific dictionary mapping*: For this case-study, we used the domain-specific RadLex ontology[11] for mapping the variations of radiological terms that are related to hemorrhage, to a controlled terminology. We created an ontology crawler using SPARQL that grabs the sub-classes and synonyms of the domain-specific terms from Radlex, and creates a focused dictionary for *"Intracranial hemorrhage"* radiology reports. Using the dictionary all the equivalent terms of hemorrhage are formalized in the corpus as: {*'apoplexy', 'contusion', 'hematoma', ...* } $\rightarrow$ *'hemorrhage'*.

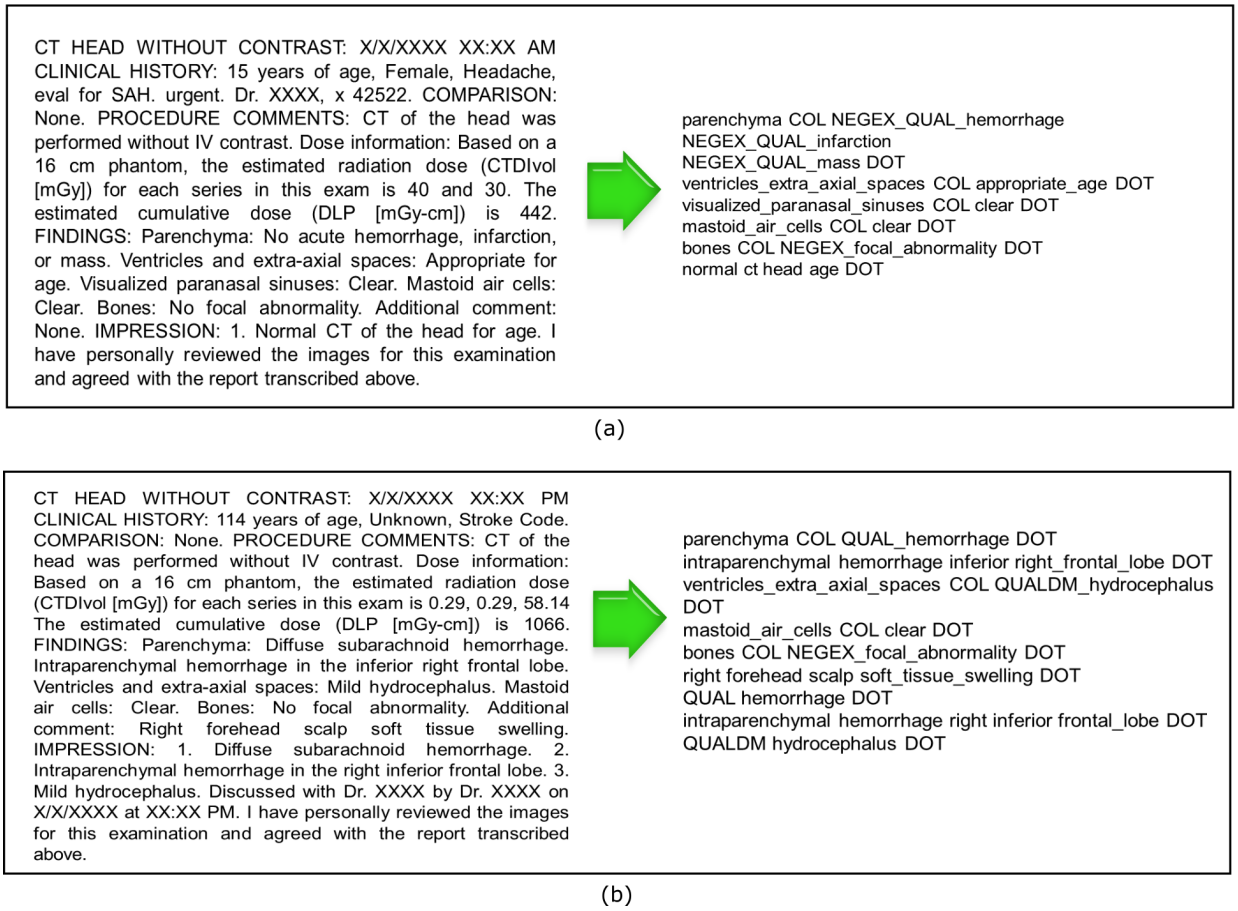

Figure 2: Examples of preprocessing and semantic-dictionary mapping - on the left FINDINGS and IMPRESSION sections of the original reports and on the right processed reports of (a) low and (b) high likelihood of intracranial hemorrhage. (Names and dates have been redacted to preserve anonymity)

In Figure 2, we present the outcome of preprossessing and semantic dictionary mapping by showing free-text reports and the corresponding processed texts side-by-side. In our corpus, average word count of original free-text reports is 285 and the average word count of processed reports is 98, which is approximately 3x reduction in size.

2.4 Word and Report Embedding

After pre-processing and dictionary mapping, the corpus of 10,000 processed reports (see examples in Figure 2) was used to create vector embeddings for words in a completely unsupervised manner using the word2vec model that can be trained on a large text corpus to produce dense word vectors. Two unsupervised algorithms were introduced to obtain word to vector representation: Continuous Bag of Words (CBOW) and Skip-gram[7]. Those algorithms learn word representations that maximize the probabilities of a word given other contextual words (CBOW) and of a word occurring in the context of a target word (Skip-gram).

Our semantic dictionary mapping step considerably reduced the size of our vocabulary by mapping the words in corpus to their root terms, thereby making the words in the vocabulary more frequent. CBOW is several times faster to train than the Skip-gram, with slightly better accuracy for frequent words. The CBOW architecture also captures the semantic regularities of words. Thus, CBOW approach appeared to be more suitable to be integrated into our framework, and, as expected, results of preliminary experiments with Skip-gram and CBOW showed CBOW to be the better performing model.

We first constructed a vocabulary from our pre-processed tokenized corpus that contains 10,000 free-text radiology reports, and then learned *vector representations of words* in the vocabulary. We build our predictive model using the Gensim 2.1.0 library[12]. The CBOW word2vec model predicts a word given a context where context is defined by the window size. The loss function of CBOW is: $E = -{v_{w_o}}'.h + \log \sum_{j=1}^{V} \exp({v_{w_j}}'.h)$, where w_o is the output word, ${v_{w_o}}'$ is its output vector, h is the average of vectors of the context words, and V is the entire vocabulary. Once the model constructs the vectors, we can use the cosine distance of vectors to denote similarity, thereby deriving analogies. The resulting word vectors can be used as features in many natural language processing and machine learning applications.

As the training algorithm, we used both Hierarchical Softmax as well as Negative Sampling. Based on preliminary results, we found Negative Sampling to be better training algorithm. Mikolov et al.[8] also described Negative Sampling as the method that results in faster training and better vector representations for frequent words, compared to more complex hierarchical softmax. The cost function of Negative Sampling is: $E = -\log \sigma({v_{w_o}}'.h) - \sum_{w_j \in \omega_{neg}} \log \sigma(-{v_{w_j}}'.h)$, where ω_{neg} is the set of negative samples, w_o is the output word, ${v_{w_o}}'$ is its output vector and h is the average of vectors of the context words.

Finally, the *document vectors* were created by simply averaging the word vectors created through the trained model. According to Kenter et al.[13], averaging the embeddings of words in a sentence has proven to be a successful and efficient way of obtaining sentence embeddings. Each document vector was computed as: $v_{doc} = \frac{1}{\|V_{doc}\|} \sum_{w \in V_{doc}} v_w$, where V_{doc} is the set of words in the report and v_w refers to the word vector of word w.

2.5 Visualization of the embeddings

Our idea is to visualize the vector representation of words and documents to validate the semantic quality of the embeddings in two different levels. In the first level, the visualization of the trained individual word embeddings can verify the positioning of synonyms (and related words), antonyms and other word-to-word relations, and can show at the very low scale that if our vector embedding is able to preserve legitimate semantics of the natural words and clinical terms. Second, the visualization of the document vectors can fulfill the purpose of analyzing the proximity of documents that have different levels of likelihood of intracranial hemorrhage. If the documents corresponding to the same class (risk) appear close to each other and form clusters, we can infer that our embedding can be useful to boost the performance of any standard classifier.

Our trained embeddings are expected to be high dimensional and may lie near a low-dimensional, non-linear manifold. Therefore, standard linear dimensionality reduction techniques (e.g. Principal Component Analysis) are not well-suited for preserving the distance between similar data points in low-dimensional representation of the vector space. We adopted t-Distributed Stochastic Neighbor Embedding (t-SNE) technique[14] to visualize the trained embeddings using sklearn python library. t-SNE is a technique for dimensionality reduction that is particularly well suited to serve our application since it is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. It employs Gaussian kernel in the high-dimensional space and defines a soft border between the local and global structure of the data. For pairs of data points that are close to each other relative to the standard deviation of the Gaussian, t-SNE determines the local neighborhood size for each data point separately based on the local density of the data. We describe the results of t-SNE visualization of word and document vectors in the following section (Sec. 3.2).

2.6 Classification

In this study, the resulting document vectors were used as features to develop a computerized hemorrhage likelihood assessment system that aims to assign a *'risk'* label to the free-text radiology reports while being trained on the subset of reports with the ground truth labels created by the experts (see Sec. 2.1). We observed that our dataset had imbalanced distribution of training data, i.e. class 2, 3, and 4 had fewer instances than class 1 and 5. Thus, we grouped classes 2-4, and re-defined the class labels to ensure variation of the likelihood of intracranial hemorrhage as: (1) 'no risk' - no intracranial hemorrhage; 2) 'medium risk' - probability of having intracranial hemorrhage; (3) 'high risk'- definite diagnosis of intracranial hemorrhage. The re-definition of the class labels were validated by forming a mutual agreement between the two expert radiologists. In Table. 1, we show the number of examples per class for

Table 1: Number of examples per category in our dataset

	Class labels		
No. of cases	**'norisk'**	**'medium risk'**	**'high risk'**
	946	43	199

the final three categories. To quantify the performance of the classifier, the 1,188 annotated reports were randomly divided into 80% training set (950 reports) and 20% test set (238 reports). To demonstrate the true power of our vector embedding, we performed experiments using three classifiers - Random Forests, Support Vector Machines, K-Nearest Neighbors (KNN) in their default configurations.

2.7 Evaluation

We experimented with different types of kernels in SVM classifier (Radial kernel & Polynomial kernel), and different values of 'k' in kNN (k= 5,10) classifiers. To investigate the benefits of the proposed hybrid framework, we also tested each classifier's performance by creating vector embeddings of the radiology reports without the domain-specific semantic mapping (Sec. 2.3) where we skipped replacing the radiology terms and their synonyms using RadLex. However, we still substituted the common terms using the CLEVER base terminology for preserving the semantic structure of the radiology reports. In the Result section (see Sec. 3.3), we describe the performance of each classifier on the hold-out test set (238 reports) in a tabular format. Standard precision, recall and F1 score were used as metrics to quantify the classification performance.

3 Results

3.1 Word analogies

On feeding the entire corpus to the system, the final size of the resulting vocabulary was 4,442 words. We created word embeddings or semantic vector representations of words appearing in the corpus, from which several kinds of analogies could be derived by computing the similarity. The similarity score between the word vectors was computed as cosine similarity which is inner product on the normalized space that measures the cosine of the angle between two words: $Similarity = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$. Table 2 shows some synonyms/closely associated words and the cosine similarity scores of their respective word embeddings. Table 3 shows some antonyms and the cosine similarity scores of their respective word embeddings. The data demonstrate that the system has formed embeddings such that pairs of synonyms have high similarity scores while antonyms have negative similarity scores.

Table 2: Similarity scores of word embeddings of synonyms/closely associated words

Word 1	**Word 2**	**Similarity**
new	recent	0.941
overinflated	balloon_appears	0.999
infarction	evidence_hemorrhagic_conversion	0.910
infarction	acute_infarction	0.928
hemorrhage	rightward_midline_shift	0.958
hemorrhage	subdural_hemorrhage	0.964
hemorrhage	intraventricular_hemorrhage	0.959
hemorrhage	subarachnoid_hemorrhage	0.968

3.2 Vector Visualization

Figure 3 shows the 2D visualization of word vector embedding constructed using the t-SNE approach (Sec. 2.5) where each data point represents a word. A total of 4,442 words are visualized in the figure. As seen from the figure, similar words reside fairly close together and form a cluster in the map without even inclusion of any prior knowledge. This map illustrates that our word embedding can preserve semantics of the terms.

Table 3: Similarity scores of word embeddings of antonyms, NEGEX represents negation and QUAL represents severe terms (see Sec. 2.3)

Word 1	Word 2	Similarity
large	NEGEX_enlarged	-0.245
hemorrhage	NEGEX_QUAL_hemorrhage	-0.074
hemorrhage	NEGEX_QUAL_intracranial_hemorrhage	-0.245
infarction	NEGEX_QUAL_infarction	-0.070
large_territory_infarction	NEGEX_QUAL_large_territory_infarction	-0.157
midline_shift	NEGEX_QUAL_midline_shift	-0.206
abnormalities	NEGEX_QUAL_abnormalities	-0.283
mass_effect	NEGEX_QUAL_mass_effect	-0.170

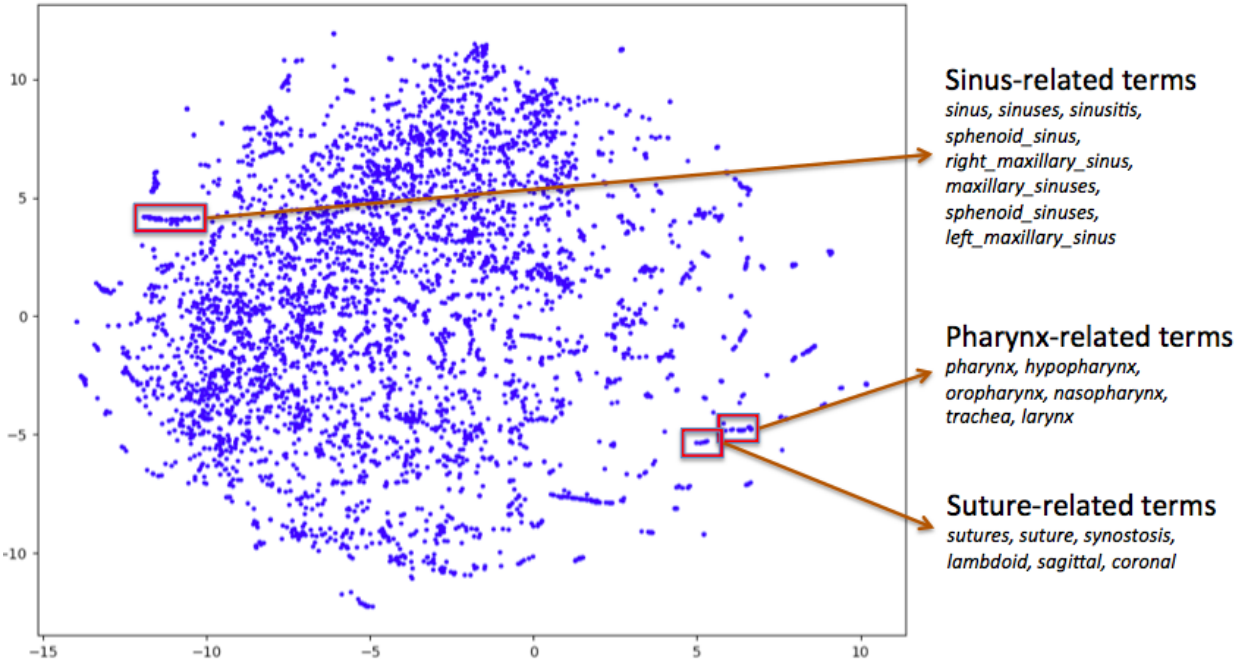

Figure 3: All word embeddings (4,442 words) - visualized in two dimensions using t-SNE

In Figure 4, we also highlight a group of clinical terms particularly relevant for this case-study and their negations using the same t-SNE visualization technique. The figure illustrates ability of the embedding to automatically organize concepts and implicitly learn the relationships between them. To show the word-to-word relations, we visualize only a few significant terms and their negations, but same technique can be used to infer other analogies among the terms present in our vocabulary (e.g. synonyms, antonyms, finding-finding, finding-diagnosis).

We also visualize the subsequent vectors of complete reports projected in two dimensions using the t-SNE technique (Figure 5). This visualization has been created only for the 1,188 annotated reports since the main idea is to see if our proposed embedding can be useful to compute clusters with varying risk factors. From the Figure 5, we can see that the reports denoting high risk of intracranial hemorrhage cluster together, and the reports with intermediate risk are mostly residing close to high risk reports. Though this is a two dimensional projection of the original high dimensional document vector, the result clearly shows that the embeddings carry signals that could be very informative to automatically annotate the reports using state-of-the-art classifiers.

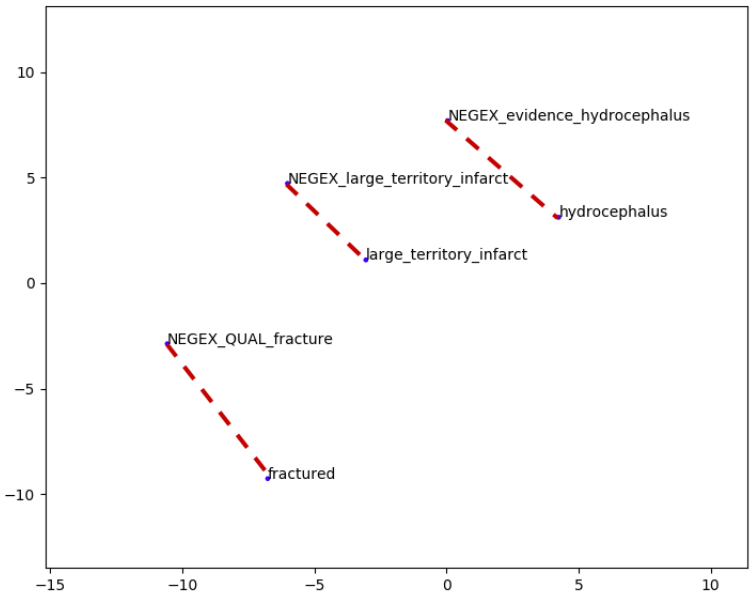

Figure 4: Word embeddings: relation between terms and their negation

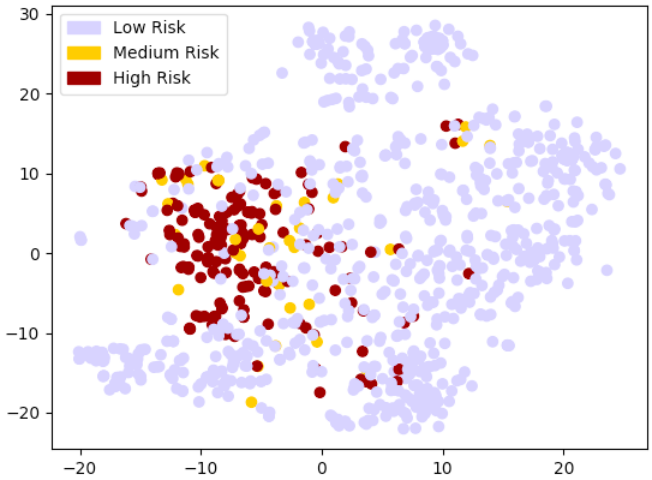

Figure 5: 1,188 CT Head radiology report vectors visualized in two dimensions

3.3 Classification performance

We used the document vectors to classify each report into one of three classes denoting varying likelihood of intracranial hemorrhage (see Sec. 2.6). As mentioned earlier in the paper, our radiology report embedding is flexible enough to be combined with both parametric and non-parametric classifiers. We experimented with three state-of-the-art classifiers - Random Forests, Support Vector Machines and K-Nearest Neighbors (KNN).

To give more insight into the quality of the learned vectors, we used the grid search approach to tune the two main hyperparameters of our embedding for the targeted annotation, i.e. *Window Size* and *Vector Dimension*. The hyperparameter search was done individually for each classifier using cross-validation on the training data set. The effects of the hyperparameters on the resulting classifier performance are shown in Figure 6 where the optimal points of the classifier's performance are highlighted. Based on the optimal points, we selected the hyperparameters and evaluated the classifiers' performance on the test set. For instance, Random Forest was evaluated with the word embeddings that

Table 4: Performance of different classifiers with and without semantic mapping, and with unigrams features.

	With Domain-specific dictionary			Without Domain-specific dictionary			Baseline with unigrams feature		
Classifier	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
Random Forests	88.64%	90.42%	89.08%	87.59%	89.17%	87.78%	87.5%	66.03%	75.26%
KNN (n = 10)	88.60%	89.91%	88.88%	86.73%	88.90%	87.47%	64.79%	80.49%	71.8%
KNN (n = 5)	88.54%	89.62%	88.76%	87.52%	88.65%	87.74%	82.62%	82.36%	75.9%
SVM (Radial kernel)	64.19%	80.09%	71.25%	63.98%	79.96%	71.07%	60.52%	77.80%	68.08%
SVM (Polynomial kernel)	63.25%	79.49%	70.43%	62.40%	78.97%	69.70%	60.52%	77.80%	68.08%

were created with window size 36 and vector dimension 730. The standard classifiers are intentionally applied in their default configurations (as in the scikit learn framework) to demonstrate the ability to achieve high performance using the embedding created by our pipeline and improvement of performance over unigrams and out-of-the-box word2vec.

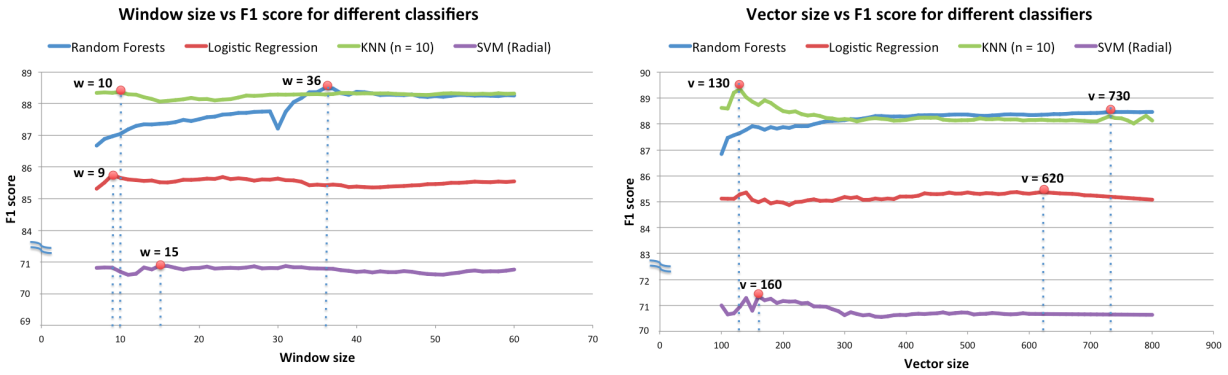

Figure 6: Hyperparameter optimization of the embeddings using grid search: window size on the left and vector dimension on the right

The classifiers' performance on the test set is reported in Table 4 with optimal hyperparameters. We also present performance of the classifiers only using unigrams as features which can be considered as the baseline performance to be compared with word embedding. While the reported performance accuracy (F1-score) of baseline with unigrams is on average 71%, the word embedding resulted F1 score over 80% for most cases which demonstrates that our vector representation was able to capture the significant facets of the radiology reports. The Random Forest classifier yielded a weighted precision of 88.64% and weighted recall of 90.42% with 730 dimensional word vectors, and closely outperforms all the other classifiers used in this study. However, KNN $(n = 10)$ produces a weighted precision of 88.60% and weighted recall of 89.91% that is close to the Random Forest's performance, employing a reduced optimal word vector dimension (130).

In Table 4, we present the classifiers' performance with and without dictionary mapping as well as with unigrams as feature. In general, the word embedding improves the performance of the baseline classifiers and every classifier's performance is consistently better with the proposed hybrid technique. However, performance difference is incremental for the particular case study which is hypothesized to be due to the choice of dataset in which all the reports are associated to a very narrow domain and from the same institution, i.e. CT Head reports, and thus the variation in the vocabulary is relatively small. We expect that superiority in the performance of the proposed hybrid method may be more significant when multi-topic and multi-institutional free-text reports will be considered where the semantic and syntactic variations are more prominent.

4 Conclusion

In this study, we have shown how to efficiently learn dense vector representations of individual words as well as entire radiology reports by using a hybrid technique that combines word2vec and semantic dictionary mapping. Our experimental results show that our proposed embeddings were able to learn the actual semantics of the ra-

diological terms from free-text reports. Thanks to the embeddings, we successfully annotated the radiology reports according to the likelihood of intracranial hemorrhage with 89.08% F1 score. We have publicly released (https://github.com/imonban/RadiologyReportEmbedding) our trained embeddings that have been used to test the classifiers performance (Table 4), which can be directly reused to support similar radiological applications, e.g. inferring relations between clinical terms, annotation of radiology reports, etc. The techniques introduced in this paper can be used also for creating vector representation from clinical notes of different domains (e.g. oncology) given a domain-specific ontology that can be used to reduce underlying term variations in the corpus.

In the prospective future studies, we will compare alternative neural word embedding methods (e.g. GloVe) since we believe that the performance of any such method will be boosted by the semantic mapping, as the models are initialize with random vector for out-of-vocabulary words which is far for reality. In the future version of the pipeline, we will incorporate log-likelihood ratio and mutual information for identify frequently appearing pairs, and will consider different linear functions (max pool, average pool, min pool etc.) to create document embedding from word vectors.

Acknowledgement

This work was supported in part by grants from the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947

References

[1] Xiaosong Wang, Le Lu, Hoo-Chang Shin, Lauren Kim, Isabella Nogues, Jianhua Yao, and Ronald M. Summers. Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database. *CoRR*, abs/1603.07965, 2016.

[2] Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. Toward best practices in radiology reporting 1. *Radiology*, 252(3):852–856, 2009.

[3] Dung HM Nguyen and Jon D Patrick. Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, 21(5):893–901, 2014.

[4] Sascha Dublin, Eric Baldwin, Rod L Walker, Lee M Christensen, Peter J Haug, Michael L Jackson, Jennifer C Nelson, Jeffrey Ferraro, David Carrell, and Wendy W Chapman. Natural Language Processin to identify pneumonia from radiology reports. *Pharmacoepidemiology and drug safety*, 22(8):834–841, 2013.

[5] Peter L Elkin, David Froehling, Dietlind Wahner-Roedler, Brett E Trusko, Gail Welsh, Haobo Ma, Armen X Asatryan, Jerome I Tokars, S Trent Rosenbloom, and Steven H Brown. NLP-based identification of pneumonia cases from free-text radiological reports. In *AMIA*, 2008.

[6] Yang Huang and Henry J Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[9] Steven Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[10] Kenneth Jung, Paea LePendu, and Nigam Shah. Automated detection of systematic off-label drug use in free text of electronic medical records. *AMIA Summits on Translational Science Proceedings*, 2013:94, 2013.

[11] Jose LV Mejino Jr, Daniel L Rubin, and James F Brinkley. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In *AMIA*, 2008.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[13] Tom Kenter, Alexey Borisov, and Maarten de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*, 2016.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.