Document Representation Learning For Patient History Visualization

Halid Ziya Yerebakan, Yoshihisa Shinagawa, Parmeet Bhatia
Siemens Medical Solutions USA / Malvern, PA
halid.yerebakan@siemens-healthineers.com
yoshihisa.shinagawa@siemens-healthineers.com
parmeet.bhatia@siemens-healthineers.com

Yiqiang Zhan
Shanghai Jiao Tong University / Shanghai, China
yiqiang@gmail.com

Abstract

We tackle the problem of generating a diagrammatic summary of a set of documents each of which pertains to loosely related topics. In particular, we aim at visualizing the medical histories of patients. In medicine, choosing relevant reports from a patient's past exams for comparison provide valuable information for precise treatment planning. Manually finding the relevant reports for comparison studies from a large database is time-consuming, which could result overlooking of some critical information. This task can be automated by defining similarity among documents which is a nontrivial task since these documents are often stored in an unstructured text format. To facilitate this, we have used a representation learning algorithm that creates a semantic representation space for documents where the clinically related documents lie close to each other. We have utilized referral information to weakly supervise a LSTM network to learn this semantic space. The abstract representations within this semantic space are not only useful to visualize disease progressions corresponding to the relevant report groups of a patient, but are also beneficial to analyze diseases at the population level. The proposed key tool here is clustering of documents based on the document similarity whose metric is learned from corpora.

1 Introduction

In medicine, examination of a patient's diseases is described in many reports in multiple specialties. Each report specializes in a specific aspect of the diseases, such as the chest, head and bones. For precise treatments, understanding the holistic picture of the patient's clinical history is critical. Unstructured text formats that are widely used further complicates the problem. Usually, relevant report retrieval and comparison are labor-intensive, particularly with patients having crowded clinical histories. As a result, important information may be overlooked due to time limitations.

Automatic matching of reports is not trivial since the reports are generally kept in unstructured text format in electronic health record (EHR) database. Exact keyword matching is not directly useful since the same entities could be written in different forms such as 'cardiac' and 'heart'. Additionally, acronyms are very common in these reports and many irrelevant reports may share same keywords. Semantic understanding of the text is necessary to find relevant report groups that experts consider as clinically similar which we named as *disease lines*.

This paper presents a representation learning algorithm and a visualization mechanism to enable clinicians to have holistic views of patients' history. In order to ensure the clinically meaningful similarity measure for the reports, we have utilized weak label information encoded in previous comparison studies conducted by radiologists.

2 Weakly Supervised Siamese LSTM

Among many alternative approaches for representation learning we decided to utilize siamese LSTM neural network architecture similar to (Mueller and Thyagarajan, 2016) on radiology reports. We de-

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 30–33
Santa Fe, New Mexico, USA, August 20-26, 2018.

rived new insights from extracted continuous space representations of text documents. We applied two different clustering algorithms to analyze extracted representations in patient and population level.

A radiology report often refers to a previous report to understand and compare patient's disease progression. These referrals are used to construct positive ground truth labels for document pairs in order to learn representation space such that clinically similar documents lie close to each other. For a given pair of documents, the label is positive if the reports are directly or indirectly referring to the other report. All the report that do not refer to each other may be considered as negative pairs. However, since the positive labels for all possible pairs is not complete we added additional modality and anatomy constraints using Apache cTakes(Savova et al., 2010) for negative labels.

Similar to (Mueller and Thyagarajan, 2016) we utilized LSTM to reduce variable length documents to space with fixed dimension. The LSTM network can be effectively used to learn very long-term dependencies with a sequence of words which turns out to be useful mechanism for relatively long medical reports. In this network, Siamese structure ensures that both the documents in given pair are represented in same Euclidean space. Another alternative is obtaining document representations using doc2vec. However, unlike *doc2vec* (Le and Mikolov, 2014), our method learns the metric based on ground truth labels because medical reports can be relevant even when the sentences are very different among them. The difficulty of creating a large labeled corpus is tackled in a weakly supervised manner.

We have used word embeddings trained on Pubmed central biomedical articles using word2vec(Mikolov et al., 2013) model for the embedding layer of LSTM network. We kept the learned word embeddings fixed during training of siamese LSTM network since changing their weights did not improve results in our experiments.

We have utilized generalized logistic loss as our objective function given in Equation 1(Hu et al., 2014) to train siamese LSTM network. Minimization of such loss essentially reduces the distance between positive pair of document while at the same time increases the distance between negative pairs. In this formula, β and τ are the hyper parameters and y denotes label information. Distance d is selected as Euclidean norm.

$$F(d, y) = \frac{1}{\beta} log(1 + e^{\beta(\frac{3}{2} - y(\tau - d))}) \tag{1}$$

3 Data and Preprocessing

We collected a corpus of radiology reports containing 100,000 de-identified radiology reports including studies on chest x-rays, abdominal CTs, and brain MRIs. The maximum number of reports per patient is 74. There are 25,546 unique patients with 12,677 of them being one time admission only;i.e, these patients have only one report entry in the database. More than 97% of reports are shorter than 250 words. Thus, we decided to limit the maximum number of the words to 250. Furthermore, as a result of data generation step explained in previous section, we have total of 32,000 positive pairs and 91,000 negative pairs among the 165,000 possible pairs of reports. Note that we only considered intra-patient pair of documents to construct our training and test data sets.

For data preprocesing, we applied stemming and lowercase to all the words in documents for normalization. We have removed punctuation marks and numbers as well. For tokenization, we have used NLTK word tokenizer. We used cTAKES to obtain tags such as pathology, anatomies, symptoms and negation from the radiology reports.

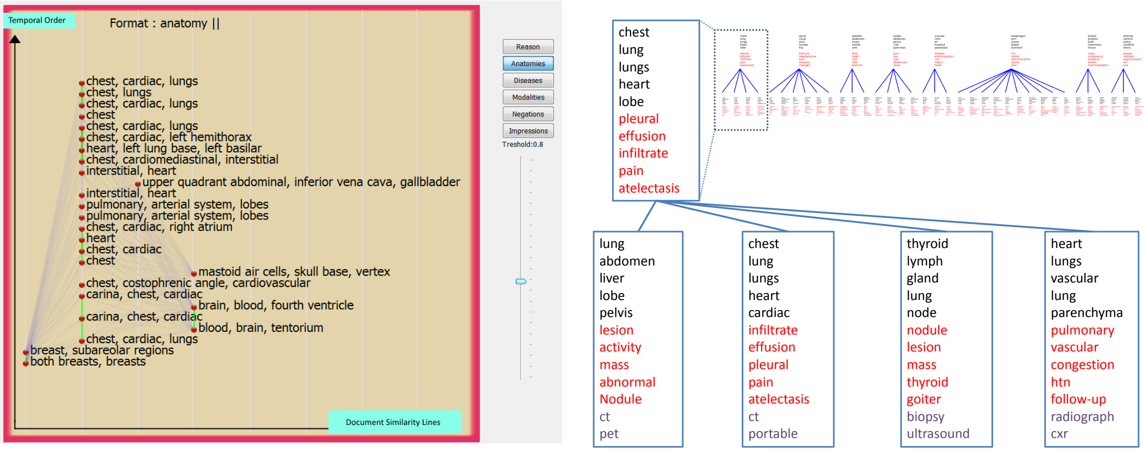
4 Results

In this section, we evaluate the quality of the representations learned by the siamese LSTM network.

In the experiments, the hyper-parameters are chosen as follows: τ=0.25 , separation = 1, β=2. Network width is selected as 32. We have trained the network for 10 epochs with batch size of 200.

4.1 Performance Evaluation

After creating a split of training and test pairs at the patient level, the documents are converted into integer word IDs and are passed through the network. As a result, we have obtained 32-dimensional



(a)

(b)

Figure 1: Patient level visualization system and population level analysis.

floating point representation for each report. There can be multiple alternatives to define a metric in this space to differentiate positive pairs from negative ones. We have chosen Euclidean norm as our metric that indicate similarity of documents. A pair of documents is considered as positive if the distance between their floating point representations is less than 0.5. With this threshold, we obtain 0.976 $F1$ score on the test set.

We further established baseline using bag-of-words model with support vector machines and logistic regression. These baselines gives 0.823 and 0.820 $F1$ scores, respectively. This clearly suggests that LSTM network can learn semantic relationship among words that bring positive pair of documents close to each other, whereas classical algorithms like bag-of-words fails to do so effectively.

4.2 Clustering Patient Lines

The learned representations captures the intrinsic comparison information given in referral of reports. In order to obtain meaningful report groups we further utilize connected component based clustering algorithm. The algorithm first calculates the distance matrix and then applies a threshold to the distance matrix, followed by calculation of connected components on the binarized distance matrix. Each connected component represents one cluster. This system allows to have different threshold levels which could change the granularity of obtained clusters. Other clustering algorithms could be used as well, however this approach provides a non parametric clustering with an interactive interface.

Using relevant reports groups, we have developed a 2-dimensional visualization methodology to get an overview of the patient's history. The reports in a cluster are given a unique y-coordinate and are aligned on a temporal line parallel to y axis according to their temporal order. Different clusters are represented by the lines at different x-coordinates. In this way, every cluster is represented by a distinct temporal line. Thus, the progression of diseases can be easily followed along the y-axis for a particular group (cluster) of reports. In order to qualitatively understand the performance of system, we have created a visualization where in we connected all the pairs with positive ground truth label with green lines and all negative pairs with blue lines as shown in Figure 1a. Ideally, there should be no green line across clusters and vice-versa for blue lines which is indeed the case as can be seen in Figure 1a.

Furthermore, our system facilitates visualization of patient's history in multiple perspectives such as anatomies, modalities and negations. The tagging system allows to extract the informative tags in the reports such as anatomies, modalities, and negations and these tags can be used to highlight particular perspective on visualization dashboard based on selection made by the medical practitioner. In this way, our system provides an interactive and holistic view of overall patient's history.

As it could be seen in the Figure 1a, the patient's reports are clustered into four groups, which match the ground truth labels shown in green lines. This patient first visited the hospital for breast screening, which is represented by the left most disease line. At one point, however, there was bleeding in the brain and the patient had CT exams, which is represented by right-most disease line. The patient was intubated and closely monitored, which is represented by the 2nd disease line from the left.

4.3 Population Level Analysis

Relationships in a large corpus of cross-patient radiology reports can help to understand disease patterns and their possible treatments. Manually analyzing such large radiology corpora is impractical for most clinically relevant applications. Thus, clustering algorithms could be used to visualize the patterns. However, basic clustering algorithms do not consider the hierarchical relations that exist in medical reports. Exploration of such structure within clinical data will further facilitate to understand the correlations across different diseases and sub-types. We choose the two-layer clustering algorithm named I2GMM (Yerebakan et al., 2014) that not only perform clustering but also extract sub-clusters for all the clusters.

We obtained representations of all the reports using shared LSTM network from the learned siamese network and applied I2GMM clustering algorithm on these representations. In order to understand the details of each cluster we extracted tags from each cluster. Most frequent tags are shown in Figure 1b. Zoomed cluster consist of different chest studies. The four sub-clusters from left to right in this cluster could be differentiated as malignant neoplasms, lung pathologies, neck related problems and vessel problems, respectively. In this figure, we used black color for anatomy tags, red for symptom or pathology tags, and purple for modalities. This result indicates that the learned document representations provide population level groups and sub-groups to relate patients at more abstract level. Such visualization could potentially be used to obtain information about different pathways of various diseases in more detail.

5 Conclusion

In this paper, we have presented a visualization system displaying a summary of the medical history of individual patients. The summary is obtained via clustering algorithm on top of the learned representations of documents encoding prior comparison information. Later, we have used these representations to analyze the whole corpus at population level by extracting clusters and sub-clusters.

For future studies combining image data with the corresponding free text information to focus on specific anatomical regions in images could facilitate the overall navigation of patient history.

References

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

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

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Halid Z Yerebakan, Bartek Rajwa, and Murat Dundar. 2014. The infinite mixture of infinite gaussian mixtures. In *Advances in neural information processing systems*, pages 28–36.