Project Final Submission

Information Retrieval and Extraction

Team IRE'ers

Saravanan Senthil

Tejasvi Chebrolu

Prajneya Kumar

Nikhil Bishnoi

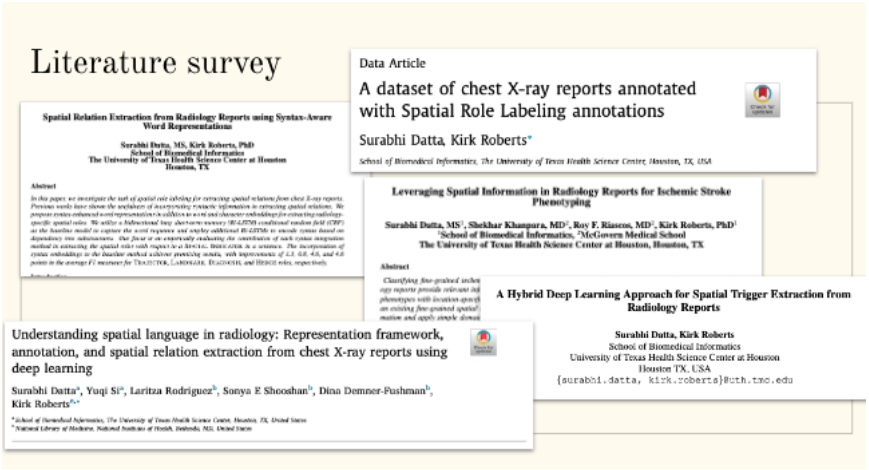
Spatial Role Labelling in Radiology Reports

The goal is to understand and implement techniques for Spatial Role Labelling in Radiology Reports.

It involves identifying spatial relations between spatial entities and spatial connectors mentioned in radiology reports.

Further, we seek to extract spans indicating diagnosis and hedges as part of the spatial relations wherever present.

Related Work



Spatial Relation Extraction from Radiology Reports using Syntax-Aware Word Representations

This paper focuses on spatial role labelling for extracting spatial information from chest X-rays.. They proposed syntax-enhanced word representations in addition to word and character embeddings for extracting radiology- specific spatial roles. This paper uses a bidirectional long short-term memory (Bi-LSTM) as a syntax encoder, conditional random field (CRF) as the baseline model to capture the word sequence and employ additional Bi-LSTMs to encode syntax based on dependency tree substructures.

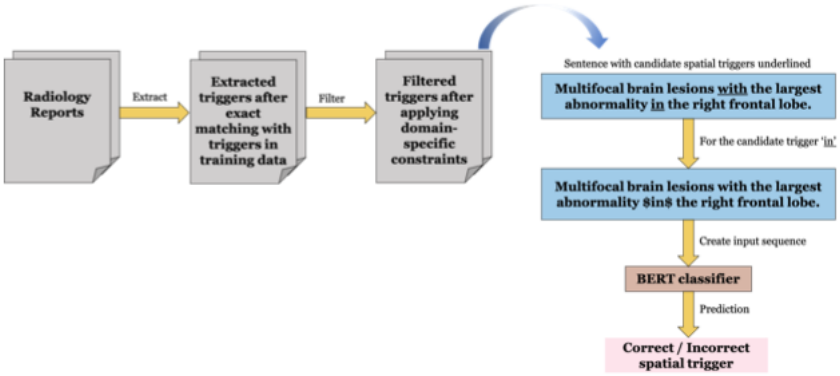
Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning

This paper deals with Radiology reports, which usually describe spatial relations. Significant radiographic findings are primarily described about an anatomical location through spatial prepositions. Such spatial relationships are also linked to various differential diagnoses. The description also may contain uncertain phrases. Thus, the Structured representation of this clinically significant spatial information has the potential to be used in a variety of downstream clinical informatics applications. The paper seeks to extract these spatial representations from these reports. The process consists of making a framework based on the Spatial Role Labeling (SpRL) scheme, which we refer to as Rad-SpRL. In Rad-SpRL, common radiological entities tied to spatial relations are encoded through four spatial roles: TRAJECTOR, LANDMARK, DIAGNOSIS, and HEDGE, all identified in relation to a SPATIAL INDICATOR. We annotated a total of 2,000 chest X-ray reports following Rad-SpRL. A deep learning-based natural language processing (NLP) method involving word and character-level encodings to first extract the SPATIAL INDICATORs, followed by identifying the corresponding spatial roles. A bidirectional long short-term memory (Bi-LSTM) conditional random field (CRF) neural network as the baseline model. Additionally, pre-trained language models

(BERT and XLNet) are used for extracting spatial information. We evaluated both gold and predicted SPATIAL INDICATORs to extract the four types of spatial roles.

A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports

Radiology reports contain important clinical information about patients, which is often tied through spatial expressions. Spatial expressions are used to describe the positioning of radiographic findings or medical devices with respect to some anatomical structures. As the expressions result from the mental visualization of the radiologist's interpretations, they are varied and complex. The focus of this paper is to automatically identify the spatial expression terms from three different radiology sub-domains. We propose a hybrid deep learning-based NLP method that includes generating a set of candidate spatial triggers by an exact match with the known trigger terms from the training data, applying domain-specific constraints to filter the candidate triggers, and utilizing a BERT-based classifier to predict whether a candidate trigger is a true spatial trigger or not.



Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports

This paper proposes a representation framework for encoding spatial language in radiology based on frame semantics. It builds on the SpatialNet representation in the general domain with the aim of generating more accurate representations of spatial language used by radiologists. We describe Rad-SpatialNet in detail and illustrate the importance of incorporating domain knowledge in understanding the varied linguistic expressions involved in different radiological spatial relations. This work also constructs a corpus of 400 radiology reports of chest X-rays, brain MRIs, and babygrams. Spatial trigger expressions and elements corresponding to a spatial frame are annotated. BERT-based models (BERTBASE and BERTLARGE) were applied to first extract the trigger terms (lexical units for a spatial frame) and then to identify the related frame elements. This frame-based resource can be used to develop and evaluate more advanced natural language processing (NLP) methods for extracting fine-grained spatial information from radiology text in the future.

A dataset of chest X-ray reports annotated with Spatial Role Labelling Annotations

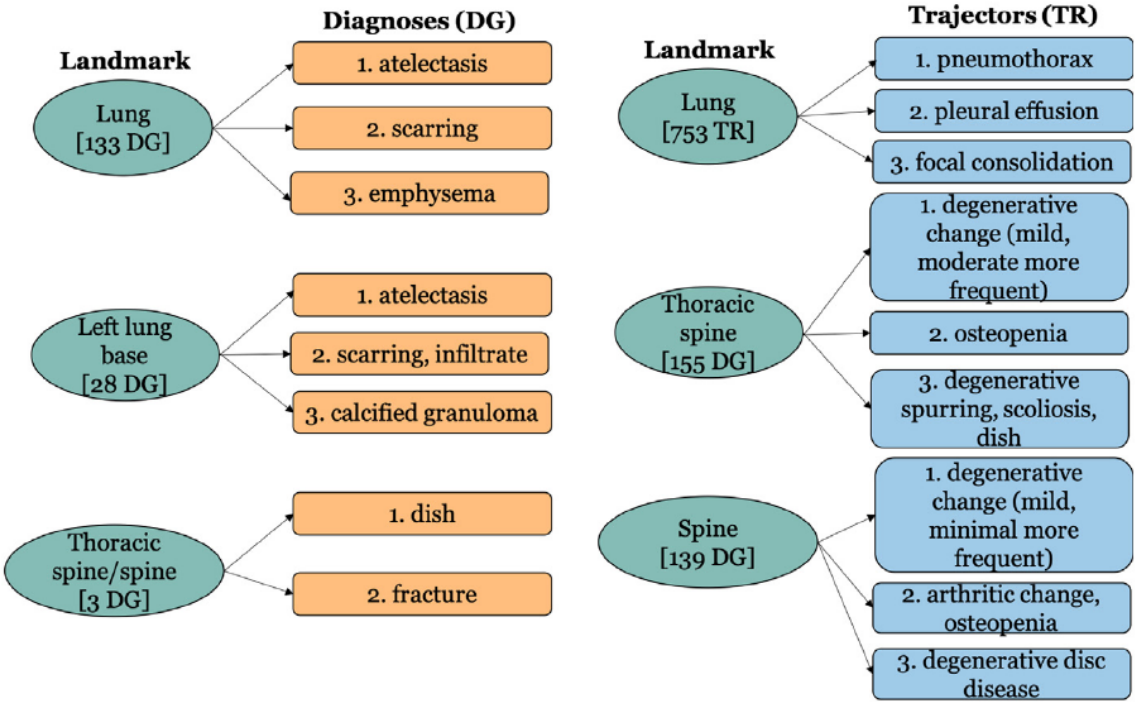
This paper contains information of around 2000 X-ray reports annotated with spatial information. The information includes annotating a radiographic finding, its associated anatomical location, any potential diagnosis described in connection to the spatial relation (between finding and location), and any hedging phrase used to describe the certainty level of a find- ing/diagnosis. All these annotations are identified concerning a spatial expression (or Spatial Indicator) that triggers a spatial relation in a sentence. The spatial roles used to encode the spatial information are Trajector, Landmark, Diagnosis, and Hedge. In total, there are 1962 Spatial Indica- tors (mainly prepositions). The dataset has 2293 Trajectors, 2167 Landmarks, 455 Diagnosis, and 388 Hedges.

Specifications Table

Subject	Health Informatics
Specific subject area	Spatial information extraction from chest X-ray reports based on Spatial Role Labeling schema for spatial language understanding in radiology reports
Type of data	Table, Figure, Text, Annotated data in XML format
How data were acquired	A subset of 2000 chest X-ray reports were used from a pool of 3996 de-identified reports collected from the Indiana Network for Patient Care (available as one of the Open-i datasets released by the National Library of Medicine.)
Data format	Raw, Processed
Parameters for data collection	2000 chest X-ray reports that are annotated with important spatial information were selected from the set of 2470 non-normal reports in the Open-i chest X-ray report dataset as adjudicated by two annotators.
Description of data collection	These 2000 reports were annotated with four spatial roles using the Brat toolkit. First, the spatial indicators (usually the spatial prepositions) triggering any spatial relation between a radiographic finding and an anatomical location were annotated for each sentence. Then, four spatial roles–the radiographic finding, its corresponding location, hedging phrase, and any potential diganosis were annotated with respect to a specific spatial indicator.
Data source location	Primary data source: Open-i chest X-ray dataset (https://openi.nlm.nih.gov/). *Associated research paper: "Preparing a collection of radiology examinations for distribution and retrieval"*–https://doi.org/10.1093/jamia/ocv080
Data accessibility	Repository name: Mendeley data repository Data identification number: 10.17632/yhb26hfz8n.1 Direct URL to data: https://doi.org/10.17632/yhb26hfz8n.1, https://github.com/krobertslab/datasets/tree/master/rad-sprl
Related research article	S. Datta, Y. Si, L. Rodriguez, S. E. Shooshan, D. Demner-Fushman, K. Roberts, Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning, Journal of Biomedical Informatics 108 (2020) 103473. doi:10.1016/j.jbi.2020.103473.

Data Description

This 20 0 0 chest X-ray reports dataset is a subset of 3996 reports collected from the Indi- ana Network for Patient Care [2]. Specifically, the 20 0 0 report subset is composed of a set of 2470 non-normal reports as judged by two human annotators. The annotation schema has been extended to encode information in a radiology context. We further analyze the variations of Spatial Indicators in the dataset. Besides the five most frequent ones, the other spatial prepositions include – 'at', 'over', 'on', 'throughout', 'under', 'along', 'near', 'to', 'through', 'between', 'adjacent', 'beneath', 'from', 'into', 'be- low', 'above', 'around', 'towards', 'about', 'behind'. This dataset also includes four more verbal spatial expressions – 'overlie', 'overlies', 'overlying', and 'involving'. However, these four expressions occur very infrequently and together account for 30 out of 1962 Spatial Indicators. Also, note that the indicator 'without' denotes a negated spatial relation and is oftentimes present as part of the common negated phrase used in radiology reports – 'without evidence of'.



Experimental Design

In this dataset, an attempt is made to widen the scope of clinically significant information types to be extracted from chest X-ray reports and additionally aim to relate all the information in context to a spatial relation between a finding and a location. This provides more contextual information about radiographic finding. Many of the previous works on radiology information extraction mainly focused on extracting radiological entities (findings, diagnoses, etc.) separately without establishing any relation among these entities.

Baseline Methodologies:

1. Converting the XML data into CoNLL Tags

   Our spatial representation framework (Rad-SpRL) consists of 4 spatial roles (trajector, landmark, hedge, and diagnosis) with respect to a spatial indicator.

   We then convert the data into a CoNLL .csv file.

2. Bi-Directional LSTM

Architecture

Dataset and Data Exploration:

The dataset provided to us was in an XML format:

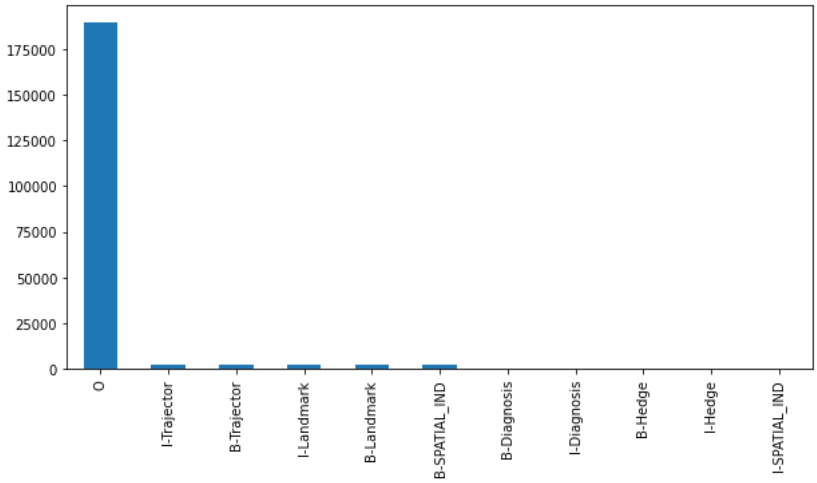
```
<Document version="1.0" id="00001">
  <MetaData />
  <Annotated>
    <Type value="SENTENCE" />
    <Type value="TOKEN" />
  </Annotated>
  <Text><![CDATA[
Chest PA-Lat XR

Imaging Study
Xray Chest PA and Lateral
Exam: 2 views of the chest XXXX/XXXX.

Comparison: None.

Indication: Positive TB test

Findings:
The cardiac silhouette and mediastinum size are within normal limits.
There is no pulmonary edema.  There is no focal consolidation.  There
are no XXXX of a pleural effusion.  There is no evidence of
pneumothorax.

Impression:
Normal chest x-XXXX.
This examination and reported findings have been reviewed and
confirmed by the undersigned.

]]></Text>
  <Annotations>
    <Token cs="1" cl="5" />
    <Token cs="7" cl="2" />
    <Token cs="9" cl="1" />
    <Token cs="10" cl="3" />
    <Token cs="14" cl="2" />
    <Token cs="18" cl="7" />
    <Token cs="26" cl="5" />
    <Token cs="32" cl="4" />
    <Token cs="37" cl="5" />
```

Since we are tackling the problem statement as a sequence labelling task, we convert the given dataset into a CoNLL format as follows:

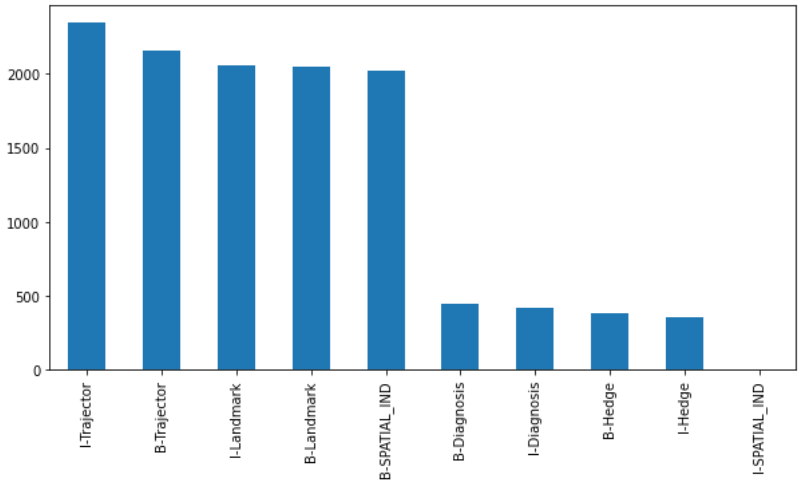
195	this	O
196	XXXX	O
197	.	O
198	Pain	B-Trajector
199	to	B-SPATIAL_IND
200	R	B-Landmark
201	back	I-Landmark
202	,	O
203	R	B-Landmark
204	elbow	I-Landmark
205	and	O
206	R	B-Landmark
207	rib	I-Landmark
208	XXXX	O

The above CoNLL tags have been constructed in the IOB Format.

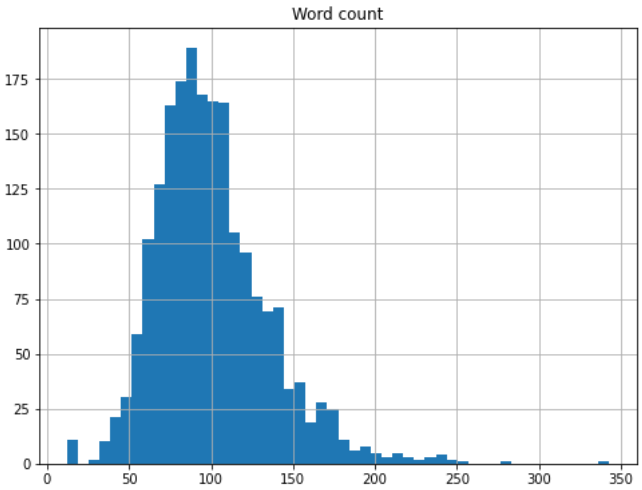
We also looked at the distribution of different tokens in the data, as can be seen in the statistical analysis below:



We notice that there are a lot of 'O' tokens in the dataset. This will skew accuracy, as we will also see in the upcoming sections. Apart from the 'O' tokens, the distributions of tags are:



If we look at the average number of words per document that we need to parse,



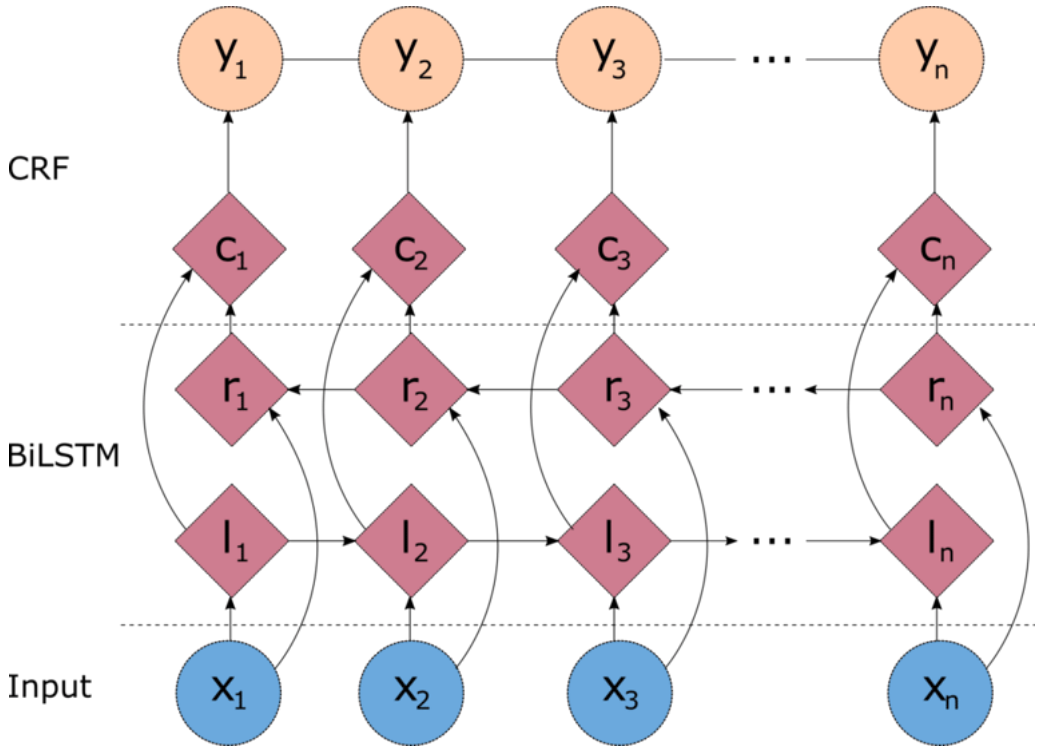
we can see that, on average, there are around 100 words per document. Keeping these statistics about the data in mind, we proceed to train our models.

Bi-LSTM:

```
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         (None, 343)               0
_________________________________________________________________
embedding_3 (Embedding)      (None, 343, 50)           165500
_________________________________________________________________
bidirectional_3 (Bidirection (None, 343, 1000)         2204000
_________________________________________________________________
crf_3 (CRF)                  (None, 343, 12)           12180
=================================================================
Total params: 2,381,680
Trainable params: 2,381,680
Non-trainable params: 0
_________________________________________________________________
```

We build a model according to the parameters given by the authors of the original paper. Our baseline model consists of an Input Layer, an Embedding Layer, A Bidirectional Layer along with a final CRF Layer.

Our architecture can be described by the following diagram:



The CRF Layer:

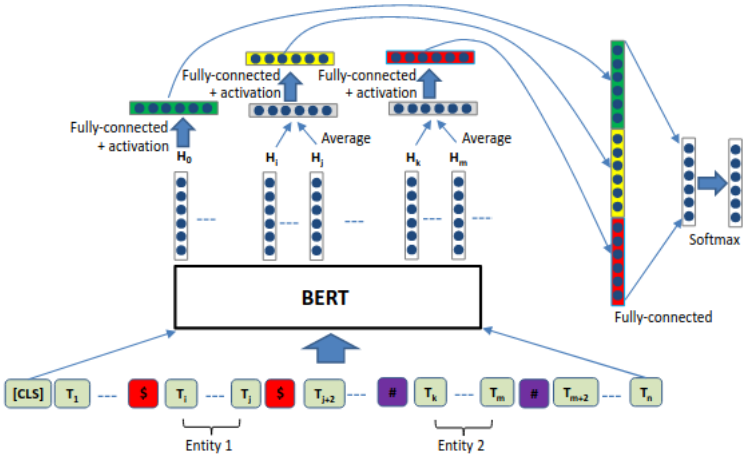
Because it considers context, the conditional random field (CRF) statistical model is highly suited for dealing with sequence labeling issues. To put it another way, a CRF model models a prediction as a graphical model in order to account for the influence of nearby data. Since the topology of CRF is an undirected graph, it is similar to Hidden Markov Model in that it assumes that the tag for the current word is only reliant on the tag of one prior word. CRF is a common type of CRF model.

R-BERT:

Relation classification is an important NLP task to extract relations between entities. The state-of-the-art methods for relation classification are primarily based on Convolutional or Recurrent Neural Networks. The pre-trained BERT model has recently achieved very successful results in many NLP classification/sequence labelling tasks. Relation classification differs from those tasks in that it relies on the information from both the sentence and the two target entities. The pre-trained BERT model (Devlin et al., 2018) is a multi-layer bidirectional Transformer encoder
(Vaswani et al., 2017). The design of the input representation of BERT is to be able to represent both a single text sentence and a pair of text

sentences in one token sequence. The input representation of each token is constructed by the summation of the corresponding token, segment and position embeddings. '[CLS]' is appended to the beginning of each sequence as the first token of the sequence. The final hidden state from the Transformer output corresponding to the first token is used as the sentence representation for classification tasks. In case two sentences are in a task, '[SEP]' separates the two sentences.

The architecture can be seen as follows:



Hyperparameters Used:

- Batch Size: 16
- Max Sentence Length: 128
- Adam Learning Rate: $2e^{-}5$
- Epochs: 5
- Dropout Rate: 5

For the pre-trained BERT model, the basic uncased model was used.

Methodology:

The basic methodology can be summarized as seen:

1. Get three vectors from BERT.
    a. [CLS] token vector
    b. Averaged entity_1 vector
    c. Averaged entity_2 vector
2. Pass each vector to the fully-connected layers.
    a. *Dropout → tanh → fc-layer*
3. Concatenate the three vectors.
4. Pass the concatenated vector to the fully-connected layer.
    a. *Dropout → fc-layer*

Running

To train the model:

```
python3 main.py --do_train --do_eval
```

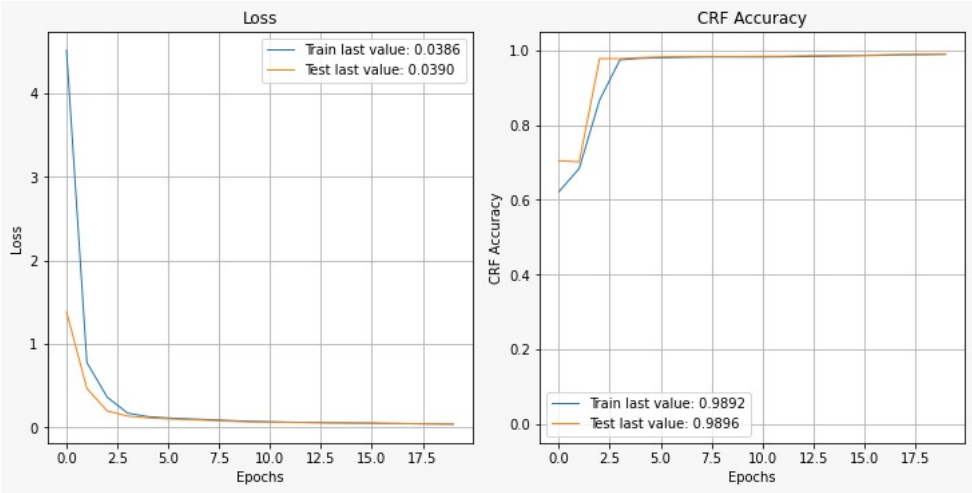
After training, the model can predict given a file containing the input sentences and the path to the output:

```
python3 predict.py --input_file (INPUT_FILE_PATH) --output_file (OUTPUT_FILE_PATH) --model_dir (SAVED_CKPT_PATH)
```

Evaluation and Results

Baseline Model (Bi-LSTM CRF)

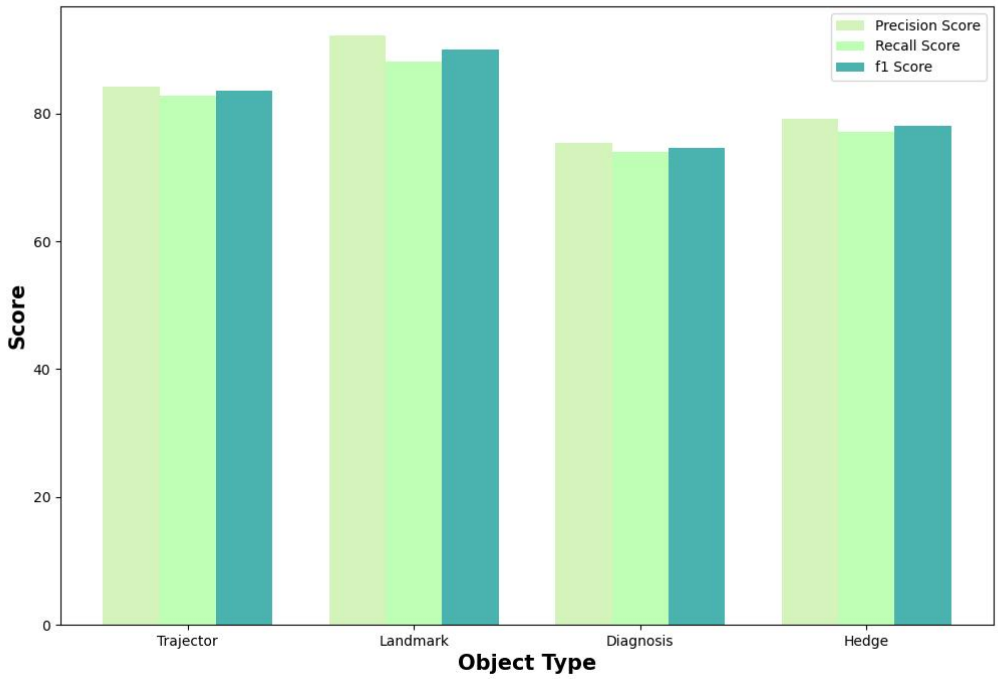
The loss and accuracy charts from the model training help us evaluate our model. Both of them looked acceptable, and the model didn't seem to be overly fitted. Hyperparameter optimization would undoubtedly help the model training. However, this kind of fine-tuning is outside the scope of this project.



The model seems to be performing really well. However, this is a little deceptive. Because there are so many O-tags in both the training and test datasets, this dataset is severely unbalanced. The samples, including the various tag classes, are further unbalanced. Building confusion matrices for each tag and assessing model performance using those would be a more thorough evaluation.

After fixing the evaluation metrics, we get the following recall, precision, and F1 scores:

TRAJECTOR			LANDMARK			DIAGNOSIS			HEDGE		
P	R	F1	P	R	F1	P	R	F1	P	R	F1
84.3	82.9	83.6	92.2	88.2	90.1	75.5	74.0	74.7	79.2	77.2	78.1

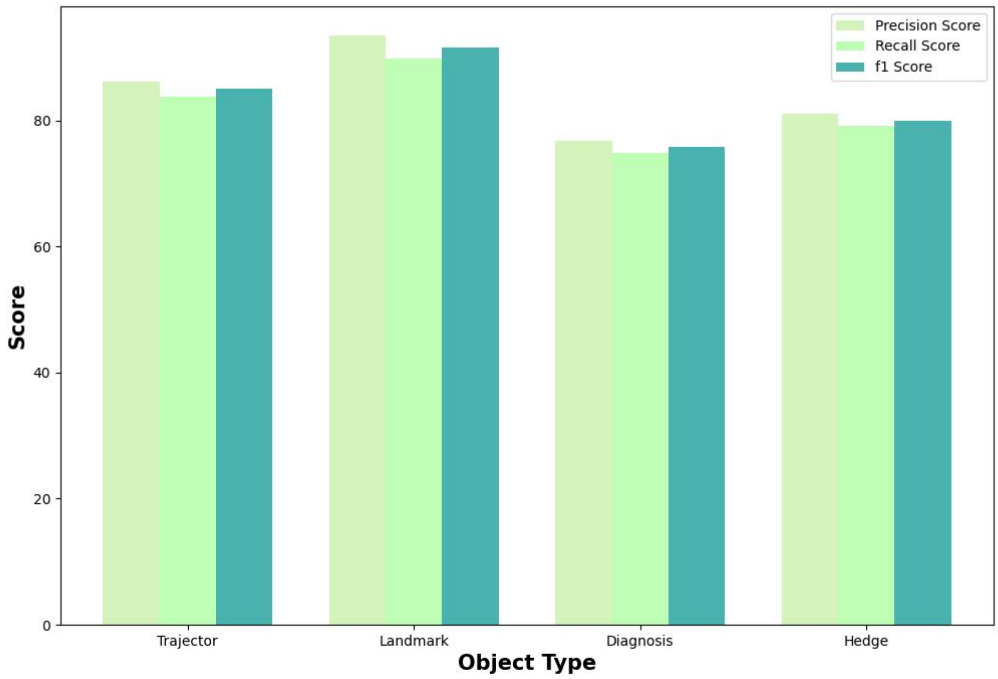


R-BERT

P	R	F1	P	R	F1	P	R
86.2	83.8	85	93.4	89.8	91.5	76.8	74.9

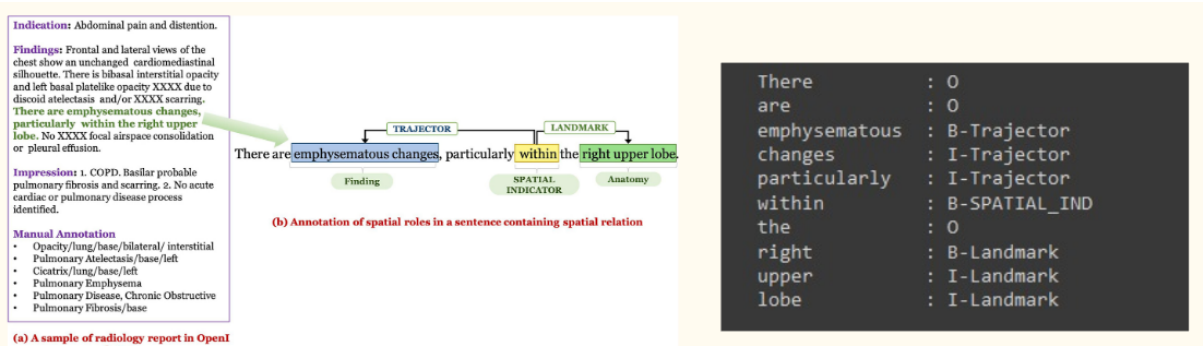
The above scores are for the following categories:

- ▼ Trajector: 85
- ▼ Landmark: 91.5
- ▼ Diagnosis: 75.8
- ▼ Hedge: 80

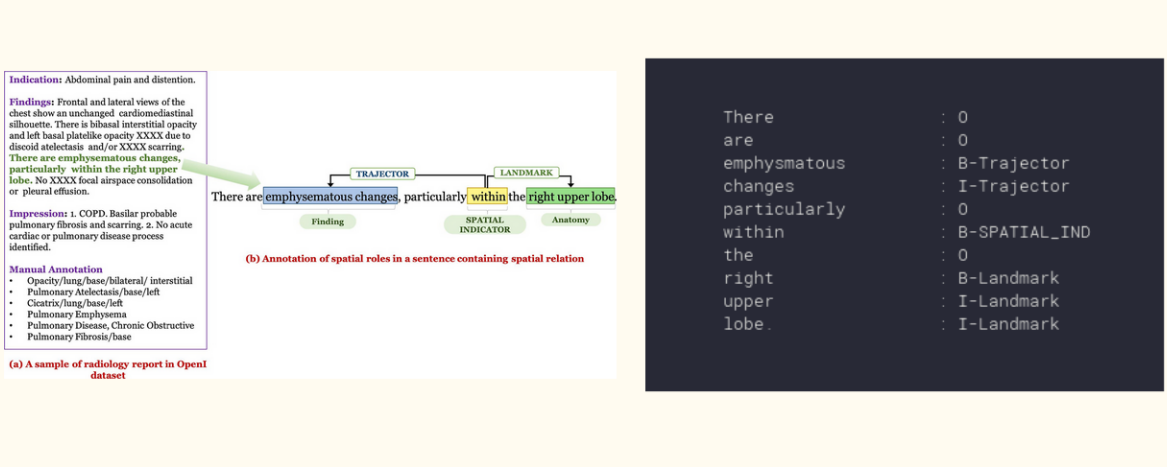


Analysis

Our baseline model performs quite well, as can be seen in one of the examples below:



However, this fails to accurately capture the fact that the word “particularly” has not been tagged correctly. It has been tagged as an *I-Trajector* when it should only be given an *O* tag. This is fixed by the R-BERT implementation.



The R-BERT model is able to leverage the semantic information to create a better understanding of the labels and therefore has managed to increase the accuracy.

References

- NLP-progress Relation Extraction
- Huggingface Transformers
- https://github.com/wang-h/bert-relation-classification