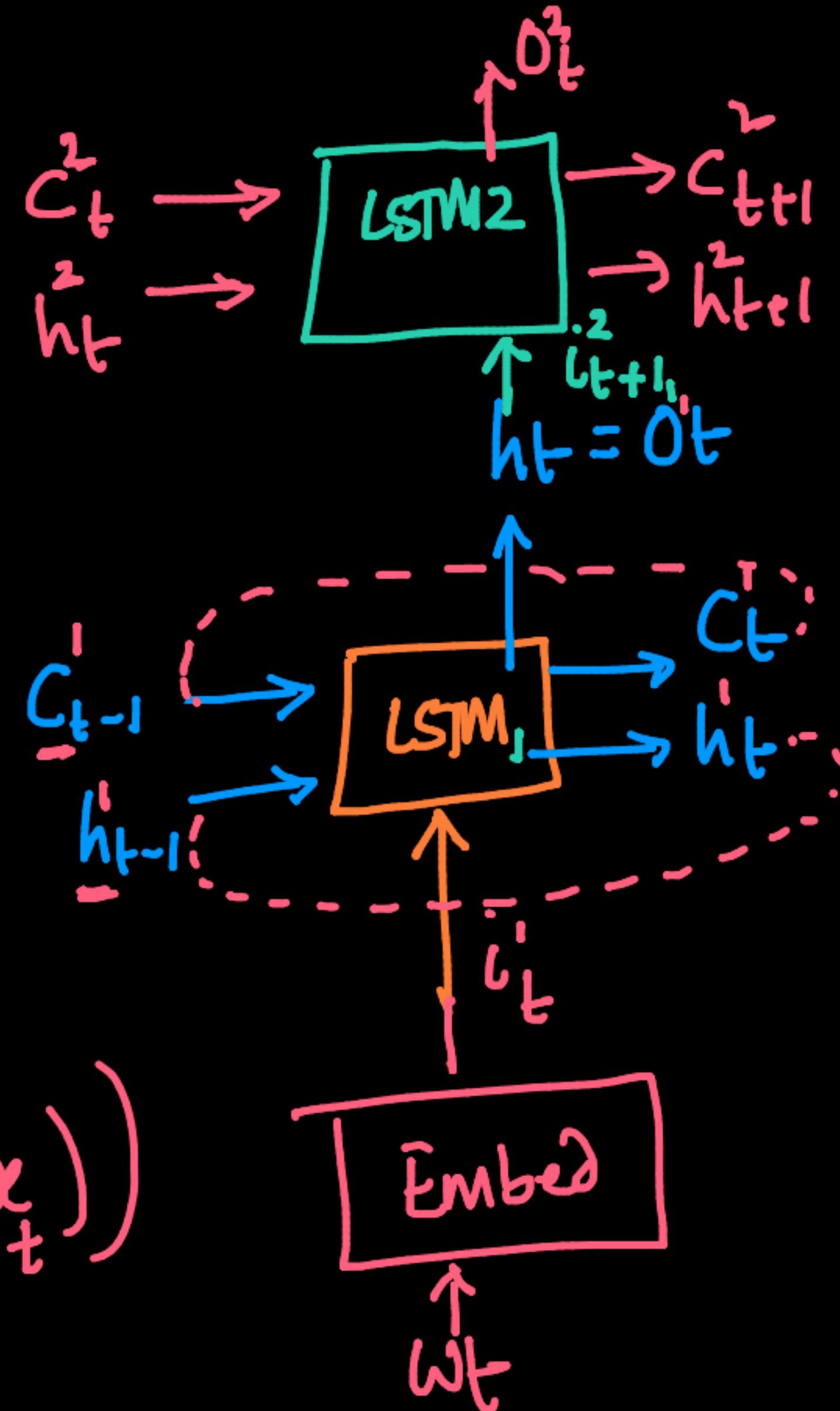


- powerful
- complex
- risk of  
overfitting



$$\underline{f}_1 \left( \underline{f}_2(x_t) \right)$$

{ Matrix mul  
   $\sigma$ , tanh  
  (non-lin)

pefk

MLP

fb

f<sub>2</sub>

1

1

1

f

Matrix-Mul  
fēlū (non-lin)

+ Code + Text Last edited on 11 November

```
[ ] recurrent_dropout=0.4)
(encoder_outputs, state_h, state_c) = encoder_lstm3(encoder_output2)

# Set up the decoder, using encoder_states as the initial state
decoder_inputs = Input(shape=(None,))

# Embedding layer
dec_emb_layer = Embedding(y_voc, embedding_dim, trainable=True)
dec_emb = dec_emb_layer(decoder_inputs)

# Decoder LSTM
decoder_lstm = LSTM(latent_dim, return_sequences=True,
                     return_state=True, dropout=0.4,
                     recurrent_dropout=0.2)
(decoder_outputs, decoder_fwd_state, decoder_back_state) = \
decoder_lstm(dec_emb, initial_state=[state_h, state_c])

# Dense layer
decoder_dense = TimeDistributed(Dense(y_voc, activation='softmax'))
decoder_outputs = decoder_dense(decoder_outputs)

# Define the model
model = Model([encoder_inputs, decoder_inputs], decoder_outputs)

model.summary()
```

The diagram illustrates a neural network architecture. On the left, there is a vertical stack of three LSTM layers labeled **LSTM1**, **LSTM2**, and **LSTM3**. Above **LSTM3** is an embedding layer labeled **E**. An input vector  $\vec{e}$  is shown entering the embedding layer. The output of the embedding layer is labeled **embed**. To the right of the embedding layer, there is a sequence of hidden states labeled  $h_1, h_2, \dots, h_t$ . A green bracket on the left groups the embedding layer and the three LSTM layers as the **encoder**. A yellow bracket on the right groups the sequence of hidden states as the **decoder**. The final output is labeled  $O_t$ . A blue bracket at the bottom groups the entire sequence of hidden states as the **outputs**.

LSTM\_V1.ipynb - Colaboratory | ROUGE (metric) - Wikipedia | L08\_NER\_re.ipynb - Colaboratory | +

Update

+ Code + Text Last edited on 11 November

Connect ▾ |   | ▾

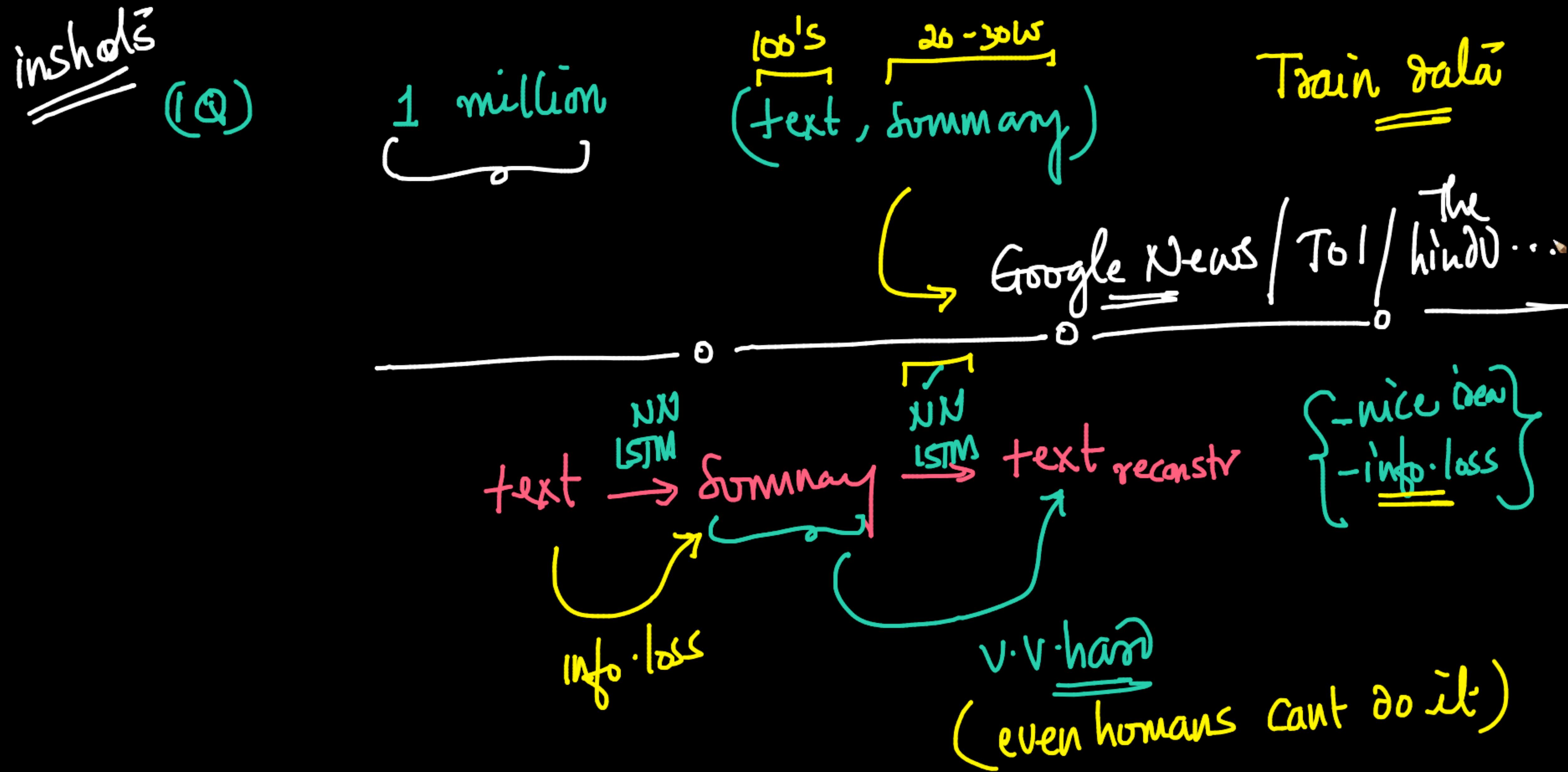
```
=====] - ETA: 0s - loss: 1.0648 - accuracy: 0.7901  
l_loss did not improve from 1.67808  
=====] - 57s 1s/step - loss: 1.0648 - accuracy: 0.7901 - val_loss: 1.6851 - val_accuracy: 0.7422  
  
=====] - ETA: 0s - loss: 1.0543 - accuracy: 0.7915  
l_loss did not improve from 1.67808  
=====] - 57s 1s/step - loss: 1.0543 - accuracy: 0.7915 - val_loss: 1.6880 - val_accuracy: 0.7423  
  
=====] - ETA: 0s - loss: 1.0455 - accuracy: 0.7927  
l_loss did not improve from 1.67808  
=====] - 57s 1s/step - loss: 1.0455 - accuracy: 0.7927 - val_loss: 1.6882 - val_accuracy: 0.7428  
  
=====] - ETA: 0s - loss: 1.0357 - accuracy: 0.7942  
l_loss did not improve from 1.67808  
=====] - 57s 1s/step - loss: 1.0357 - accuracy: 0.7942 - val_loss: 1.6912 - val_accuracy: 0.7424  
Early stopping
```

- Plot the train and validation loss

inPSen

outsent → - - -

```
[ ] from matplotlib import pyplot  
  
pyplot.plot(history.history['loss'], label='train')
```



ROUGE - 3 }  
f-score: 0.8  
hard

- ① Larger data → scrape it (respectful website policies)
- ② { encoder → { decoder  
  (3) }
- ③ let's increase enc & dec layers ...  
    → overfit
- add more depth  
    → GRU's

+ Code + Text

Connect



HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a puncture wound with a flap injury. The patient has a flap wound SYMPTOM also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The wound SYMPTOM is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the wound SYMPTOM was cleaned and wrapped, and then he was referred to us for further assessment.

PAST MEDICAL HISTORY: Significant for history of pulmonary fibrosis DISEASE and atrial fibrillation DISEASE . He is status post bilateral lung transplant back in 2004 because of the pulmonary fibrosis DISEASE .

ALLERGIES: There are no known allergies.

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, CellCept CHEMICAL , prednisone CHEMICAL , omeprazole CHEMICAL , Bactrim CHEMICAL which he is on chronically, folic acid CHEMICAL , vitamin D CHEMICAL , Mag-Ox, Toprol-XL, calcium CHEMICAL 500 mg DOSAGE , vitamin B1, Centrum Silver, verapamil CHEMICAL , and digoxin CHEMICAL .

FAMILY HISTORY: Consistent with a sister of his has ovarian cancer DISEASE and his father had liver cancer DISEASE . Heart disease DISEASE in the patient's mother and father, and father also has diabetes DISEASE .

+ Code + Text

Connect



HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a **puncture wound** with a flap injury. The patient has a flap **wound SYMPTOM** also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The **wound SYMPTOM** is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the **wound SYMPTOM** was cleaned and wrapped, and then he was referred to us for further assessment.

Named entities recognition

PAST MEDICAL HISTORY: Significant for history of **pulmonary fibrosis DISEASE** and **atrial fibrillation DISEASE**. He is status post bilateral lung transplant back in 2004 because of the **pulmonary fibrosis DISEASE**.

ALLERGIES: There are no known allergies.

✓✓  
NER

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, **CellCept CHEMICAL**, **prednisone CHEMICAL**, **omeprazole CHEMICAL**, **Bactrim CHEMICAL** which he is on chronically, **folic acid CHEMICAL**, **vitamin D CHEMICAL**, Mag-Ox, Toprol-XL, **calcium CHEMICAL**, **500 mg DOSAGE**, vitamin B1, Centrum Silver, **verapamil CHEMICAL**, and **digoxin CHEMICAL**.

✓

FAMILY HISTORY: Consistent with a sister of his has **ovarian cancer DISEASE** and his father had **liver cancer DISEASE**. **Heart disease DISEASE** in the patient's mother and father, and father also has **diabetes DISEASE**.

LSTM\_V1.ipynb - Colaboratory ROUGE (metric) - Wikipedia L08\_NER\_re.ipynb - Colaboratory F-score - Wikipedia

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

# NER: Domain Specific

HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a puncture wound with a flap injury. The patient has a flap wound SYMPTOM also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The wound SYMPTOM is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the wound SYMPTOM was cleaned and wrapped, and then he was referred to us for further assessment.

PAST MEDICAL HISTORY: Significant for history of pulmonary fibrosis DISEASE and atrial fibrillation DISEASE. He is status post bilateral lung transplant back in 2004 because of the pulmonary fibrosis DISEASE.

ALLERGIES: There are no known allergies.

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, CellCept CHEMICAL, prednisone CHEMICAL, omeprazole CHEMICAL, Bactrim CHEMICAL which he is on chronically, folic acid CHEMICAL, vitamin D CHEMICAL, Mag-Ox, Toprol-XL, calcium CHEMICAL, 500 mg DOSAGE, vitamin B1, Centrum Silver, verapamil CHEMICAL, and digoxin CHEMICAL.

FAMILY HISTORY: Consistent with a sister of his has ovarian cancer DISEASE and his father had liver cancer DISEASE. Heart disease DISEASE in the patient's mother and father, and father also has diabetes DISEASE.

Large Corpus → labelled  
text → NER → ...

LSTM\_V1.ipynb - Colaboratory x ROUGE (metric) - Wikipedia x L08\_NER\_re.ipynb - Colaboratory x F-score - Wikipedia x + colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

+ Code + Text Connect | Update

Dog Bite  
CHIEF COMPLAINT: Dog bite to his right lower leg.

(Q) NER vs POS

HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a puncture wound with a flap injury. The patient has a flap wound SYMPTOM also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The wound SYMPTOM is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the wound SYMPTOM was cleaned and wrapped, and then he was referred to us for further assessment.

PAST MEDICAL HISTORY: Significant for history of pulmonary fibrosis DISEASE and atrial fibrillation DISEASE. He is status post bilateral lung transplant back in 2004 because of the pulmonary fibrosis DISEASE.

ALLERGIES: There are no known allergies.

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, CellCept CHEMICAL, prednisone CHEMICAL, omeprazole CHEMICAL, Bactrim CHEMICAL which he is on chronically, folic acid CHEMICAL, vitamin D CHEMICAL, Mag-Ox, Toprol-XL, calcium CHEMICAL 500 mg DOSAGE, vitamin B1, Centrum Silver, verapamil CHEMICAL, and digoxin CHEMICAL.

+ Code + Text

Connect



Dog Bite

CHIEF COMPLAINT: Dog bite to his right lower leg.

City  
New York

HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a puncture wound with a flap injury. The patient has a flap wound SYMPTOM also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The wound SYMPTOM is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the wound SYMPTOM was cleaned and wrapped, and then he was referred to us for further assessment.

PAST MEDICAL HISTORY: Significant for history of pulmonary fibrosis DISEASE and atrial fibrillation DISEASE. He is status post bilateral lung transplant back in 2004 because of the pulmonary fibrosis DISEASE.

ALLERGIES: There are no known allergies.

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, CellCept CHEMICAL, prednisone CHEMICAL, omeprazole CHEMICAL, Bactrim CHEMICAL which he is on chronically, folic acid CHEMICAL, vitamin D CHEMICAL, Mag-Ox, Toprol-XL, calcium CHEMICAL, 500 mg DOSAGE, vitamin B1, Centrum Silver, verapamil CHEMICAL, and digoxin CHEMICAL.

+ Code + Text

Connect



## 1. how email service provider, suggest a meeting by detecting a time frame in the mail body.

Our next meeting will be held on Friday September 18th from 13:15 to 14:00 at Someroom or via Zoom

Suggested Meetings

day date ST ET Conf place

We think we've found an event

Our next meeting will be held on Friday September 18th from 13:15 to 14:00 a...

When: We think this event time has passed  
Who:  
Where: Enter location

There are no conflicts with any events on your calendar.

We can't automatically schedule your meeting.  
Please schedule it using Edit Details.

Edit details

Gmail

- Now, if you pass it through the Named Entity Recognition API, it pulls out the entities Bandra (location) and Fitbit (Product).
  - This can be then used to categorize the complaint and assign it to the relevant department within the organization that should be handling this.

+ Code + Text

Connect



- Person's name (Ramu, Raja, Seeta, etc.),
- Countries (India, Sri Lanka, etc),
- Organization (Google, Facebook, etc.)
- or anything that has been given a specific name.

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	The Mt. Sanitas loop is in Sunshine Canyon.
Geo-Political Entity	GPE	countries, states, provinces	Palo Alto is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the Golden Gate Bridge.
Vehicles	VEH	planes, trains, automobiles	It was a classic Ford Falcon.

## Types of NER

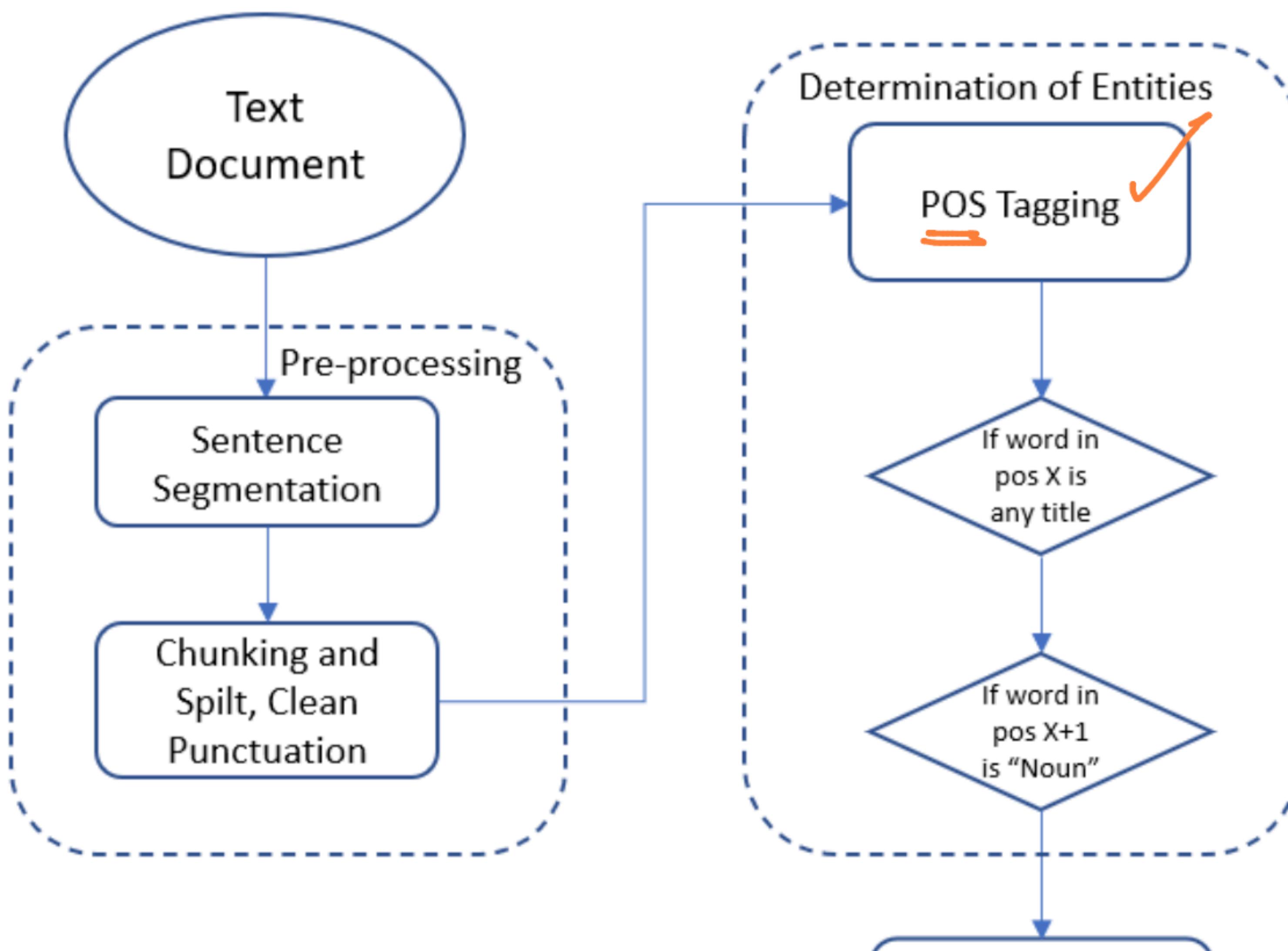
- Classical Approaches (rule-based or Dictionary)
- ML Approaches

[+ Code](#) [+ Text](#)[Connect](#)

Facility	FAC	bridges, buildings, airports	Consider the <b>Golden Gate Bridge</b> .
Vehicles	VEH	planes, trains, automobiles	It was a classic <b>Ford Falcon</b> .

- {x}
- Types of NER
- Classical Approaches (rule-based or Dictionary)
  - ML Approaches
    - Multi-class classification
    - **Conditional Random Field (CRF)**
  - DL Approaches
    - **Bidirectional LSTM-CRF**
    - Bidirectional LSTM-CNNs
    - Bidirectional LSTM-CNNs-CRF
    - Pre-trained language models (Elmo and BERT)
  - Hybrid Approaches (DL + ML)

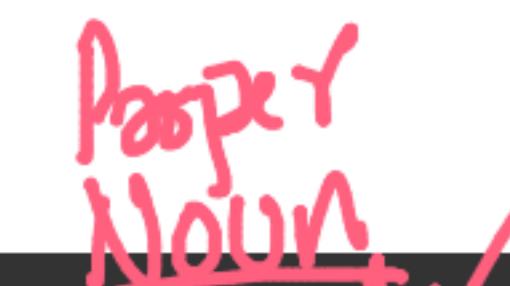
Classical Approaches:



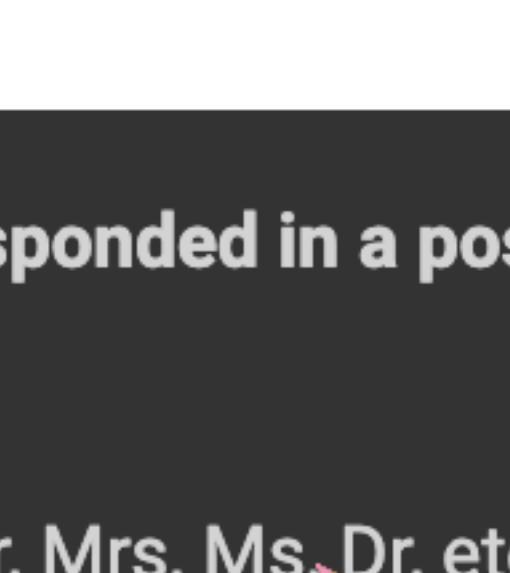
LSTM\_V1.ipynb - Colaboratory x ROUGE (metric) - Wikipedia x L08\_NER\_re.ipynb - Colaboratory x F-score - Wikipedia x + colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

+ Code + Text Connect |   

Mark word in pos X+1 as "NAME"

Sentence = "The medicine was prescribed by Dr. Sharma and the patient responded in a positive manner." 

- In this case Dr. Sharma will be labelled as name.

Cons: But we will miss on a lot of names, if it doesn't start with a title like Mr, Mrs, Ms, Dr, etc. 

Machine Learning:

Multi-class Classification

Named entities are the labels so we can apply different classification algorithms.

There are mainly two phases while we use an ML-based solution for NER.

1. Training the ML model on the annotated documents.
2. The trained model can be used to annotate the raw documents.

17 / 17

+ Code + Text

Sentence - The medicine was prescribed by Dr. Sharma and the patient responded in a positive manner.

- In this case Dr. Sharma will be labelled as name.

**Cons:** But we will miss on a lot of names, if it doesn't start with a title like Mr, Mrs, Ms, Dr, etc.

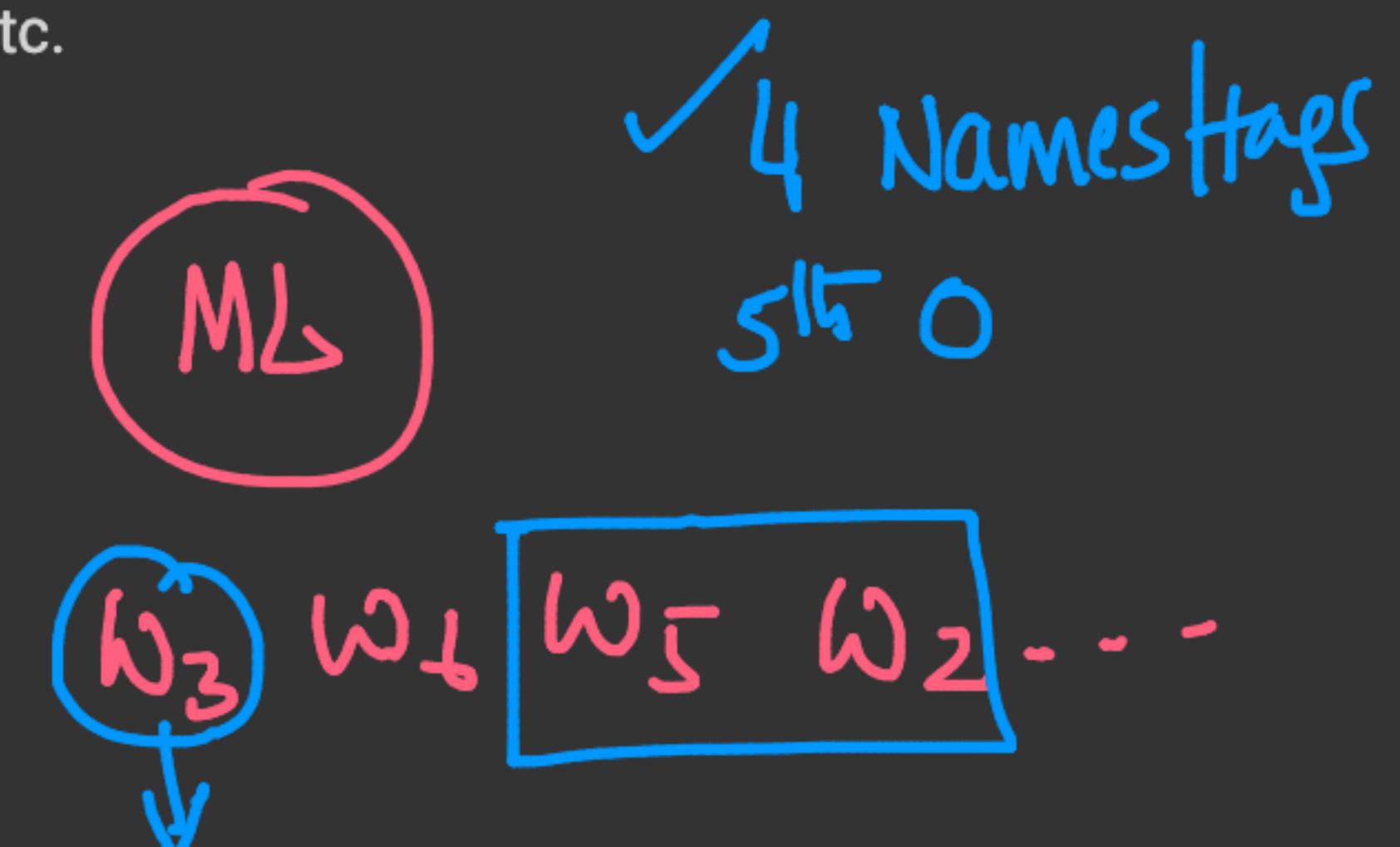
## Machine Learning:

### Multi-class Classification

Named entities are the labels so we can apply different classification algorithms.

There are mainly two phases while we use an ML-based solution for NER.

1. Training the ML model on the annotated documents.
2. The trained model can be used to annotate the raw documents.



+ Code + Text

Connect



## Problems in NER:

There exist some issues which need to be addressed.

- **Segmentation(Detection) 'ambiguity'.**

Consider 'New York'. This can be detected as 'New' & 'York' as separate entities.

Hence, deciding over the boundary is crucial whether to consider 'New York' as a single entity or 'New' and 'York' 2 different entities.

- **Tag assignment(Recognition) 'ambiguity'.**

We have 'Nirma' as a girl's name (PERSON) & as a detergent brand(Organization in India).

- In order to overcome these ambiguities, we introduce a new concept called **Sequence Labelling**

## What is Sequence labelling?

It refers to assigning labels/tags to each element of a sequence being passed as an input using an algorithm or machine learning model.

This sequence can be **words of a sentence passed in the same order as in the sentence**.

A LSTM can be taken as a Sequence labeller.

NER can be done using a number of Sequence Labelling methods listed below alongside Rule-Based methods:

1. Linear Chain Conditional Random Fields (Linear Chain CRF)

+ Code + Text

Connect

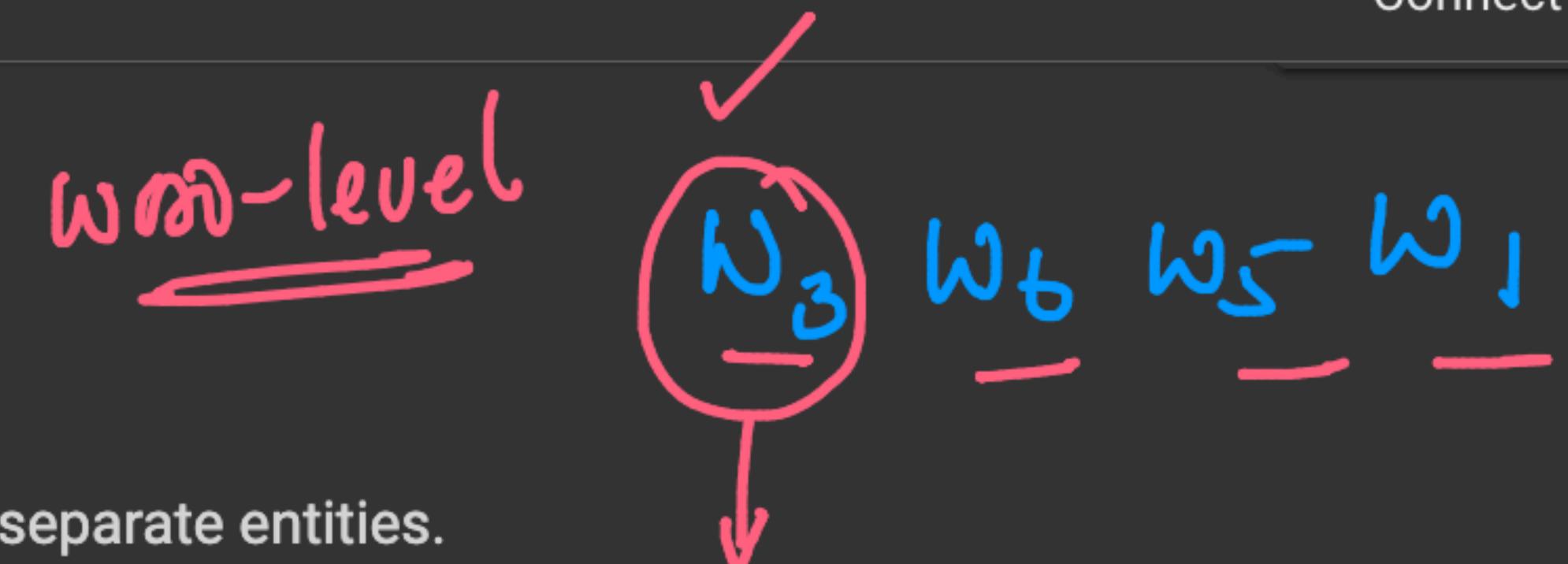


## Problems in NER:

There exist some issues which need to be addressed.

- ✓ Segmentation(Detection) 'ambiguity'.

Consider 'New York'. This can be detected as New & York as separate entities.



Hence, deciding over the boundary is crucial whether to consider 'New York' as a single entity or 'New' and 'York' 2 different entities.

- Tag assignment(Recognition) 'ambiguity'.

We have 'Nirma' as a girl's name (PERSON )& as a detergent brand(Organization in India).

- In order to overcome these ambiguities, we introduce a new concept called **Sequence Labelling**



## What is Sequence labelling?

It refers to assigning labels/tags to each element of a sequence being passed as an input using an algorithm or machine learning model.

This sequence can be **words of a sentence passed in the same order as in the sentence**.

A LSTM can be taken as a Sequence labeller.

NER can be done using a number of Sequence Labelling methods listed below alongside Rule-Based methods:

1. Linear Chain Conditional Random Fields (Linear Chain CRF)

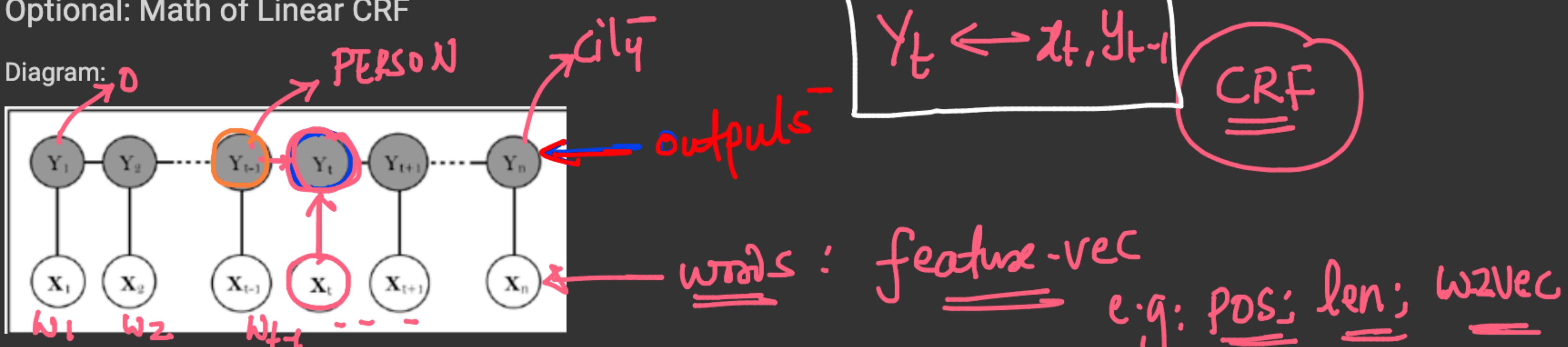
+ Code + Text

Connect ▾ |   | ✓

called as feature functions that will assist in generating unique features

- These features function can consider any logic(depends on programmer) but the output has to be either True:1 or False:0.

Optional: Math of Linear CRF



## Fundamental Equation:

Consider the below formula

$$P(\underline{y_t = \text{loc}} | \underline{x_t}, \underline{y_{t-1}}) \quad \left| \quad P(y_t = \text{loc} | x_t)$$

CRF:

LSTM\_V1.ipynb - Colaboratory | ROUGE (metric) - Wikipedia

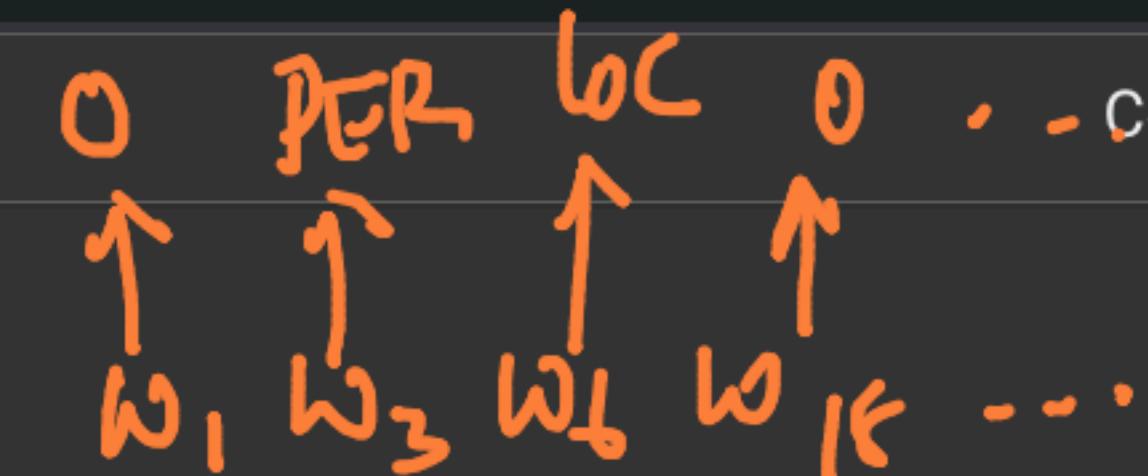
L08\_NER\_re.ipynb - Colaboratory | F-score - Wikipedia

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

Update

+ Code + Text

$$\left( \mathbf{x}_1 \right) \left( \mathbf{x}_2 \right) \dots \left( \mathbf{x}_{t-1} \right) \left( \mathbf{x}_t \right) \left( \mathbf{x}_{t+1} \right) \dots \left( \mathbf{x}_n \right)$$



Fundamental Equation:

Consider the below formula:

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

op sent

ip sent

Softmax

The formula shows the probability of a sequence  $y$  given input  $x$  using a CRF model. The numerator is the product of weights  $w_j$  and feature functions  $F_j(x, y)$  for all  $j$ . The denominator is the sum of the same over all possible sequences  $y'$ . A red brace groups the numerator and denominator, with an arrow pointing to the word "Softmax". Handwritten annotations "op sent" and "ip sent" with arrows point to the two terms in the numerator.

$$, \text{ where } F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, x, i)$$

Here,

+ Code + Text

Fundamental equation.

Connect



Consider the below formula:

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

, where  $F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, \underline{x}_i)$



Here,

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

LSTM\_V1.ipynb - Colaboratory

ROUGE (metric) - Wikipedia

L08\_NER\_re.ipynb - Colaboratory

F-score - Wikipedia

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

Update

+ Code + Text

fundamental equation.

Consider the below formula:

ip sent  $\rightarrow$  L

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

, where  $F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, \underline{x}, i)$

Here,

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

+ Code + Text

Fundamental equation.

Connect



Consider the below formula:

CRF:

feature

, where

$F_j(x, y)$

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$
$$F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, x, i)$$

is NOUN

:

Here,

SUM over  
all words

PER  
J  
 $w_3 w_6$   
 $w_4$   
 $w_3$   
nun

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

+ Code + Text

Fundamental equation.

Connect



Consider the below formula:

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

, where  $F_j(x, y) = \sum_{i=1}^L f_j(\underline{y_{i-1}}, \underline{y_i}, \underline{x}, i)$

Here,

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

+ Code + Text

Fundamental equation.

Consider the below formula:

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

, where  $F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, \underline{x}, \underline{i})$

Here,

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

logistic regression  
softmax regression

+ Code + Text

Fundamental equation.

Connect



Consider the below formula:

CRF:

$$p_{\theta}(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

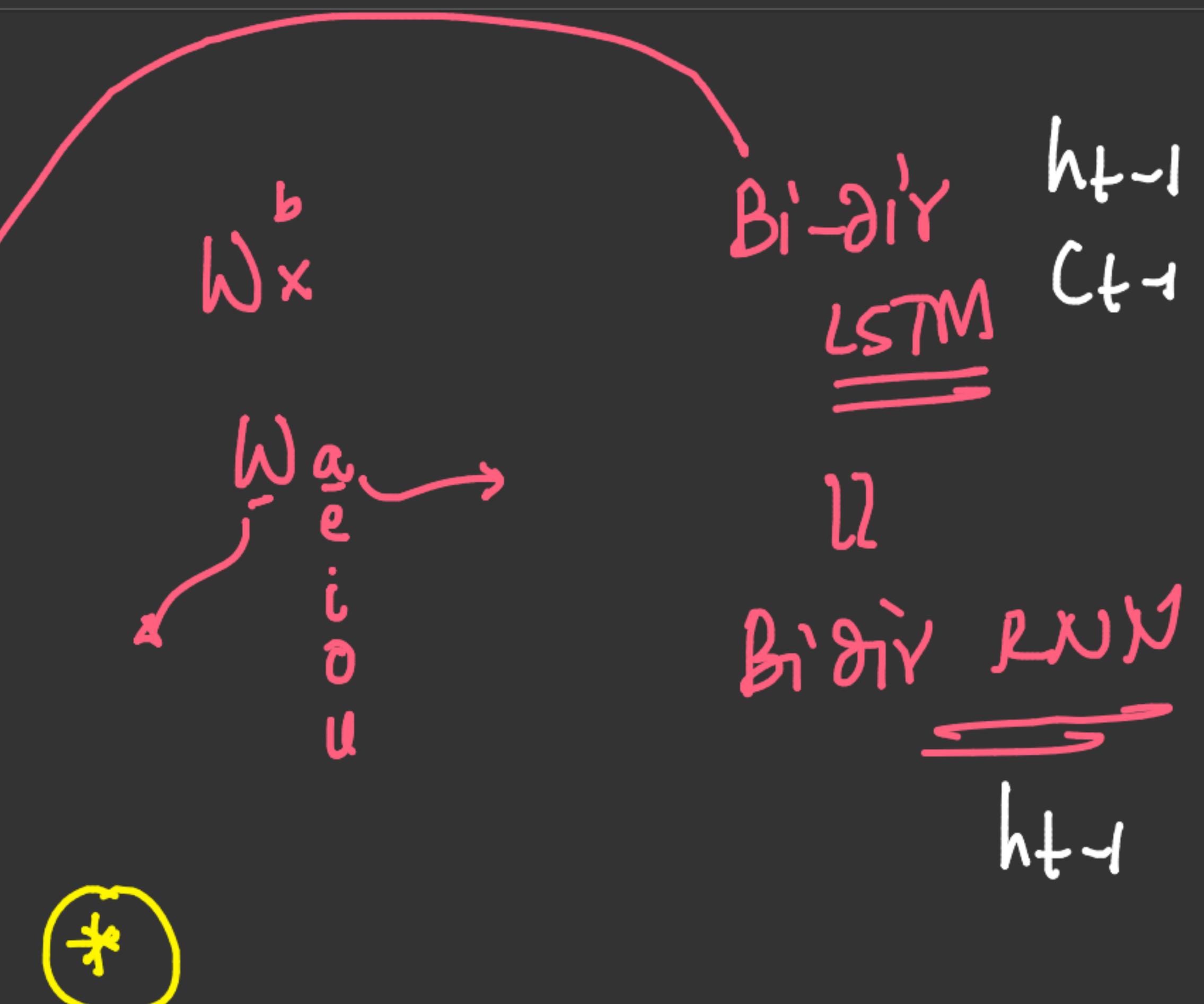
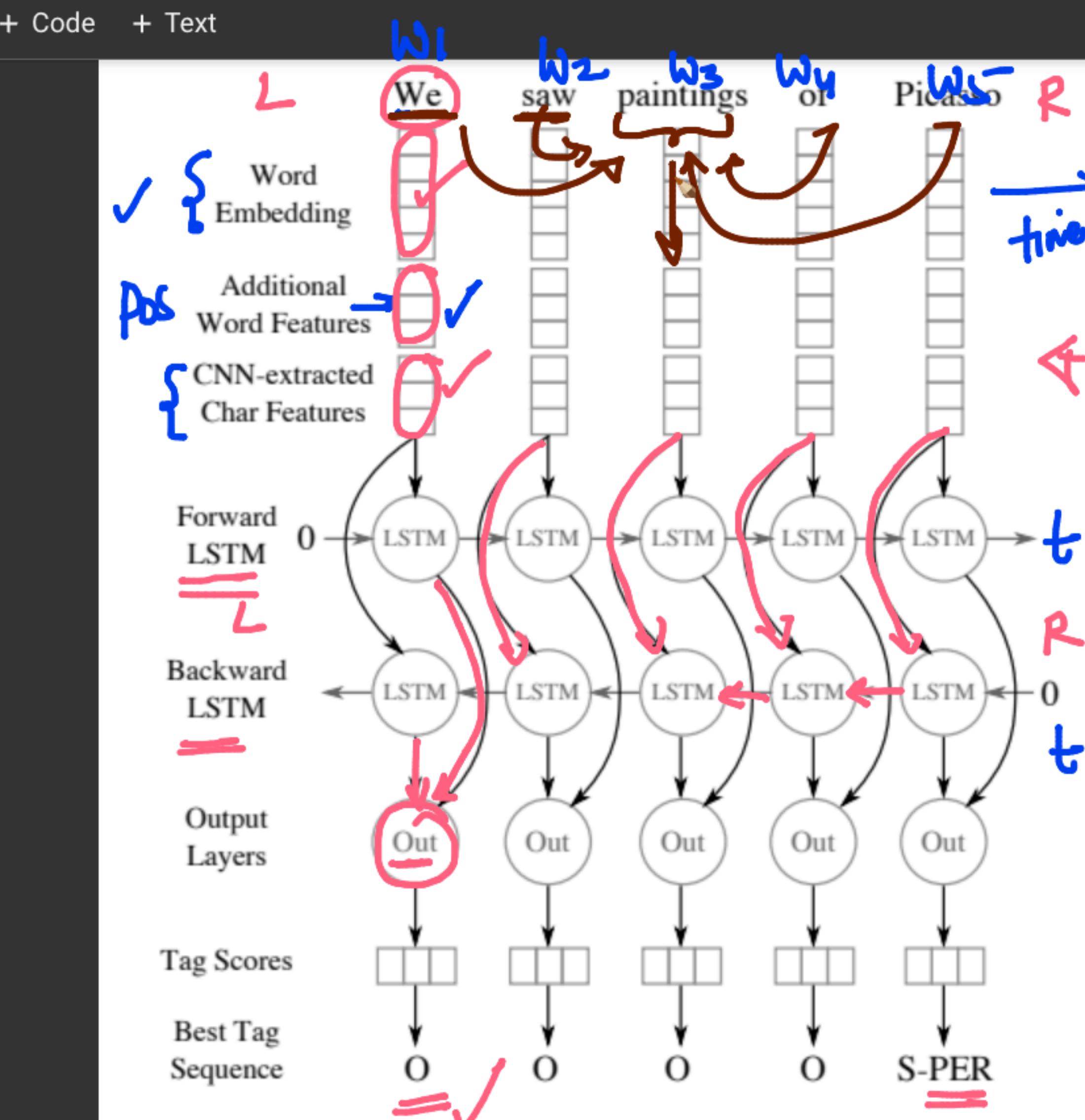
, where  $F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, x, i)$

MLE

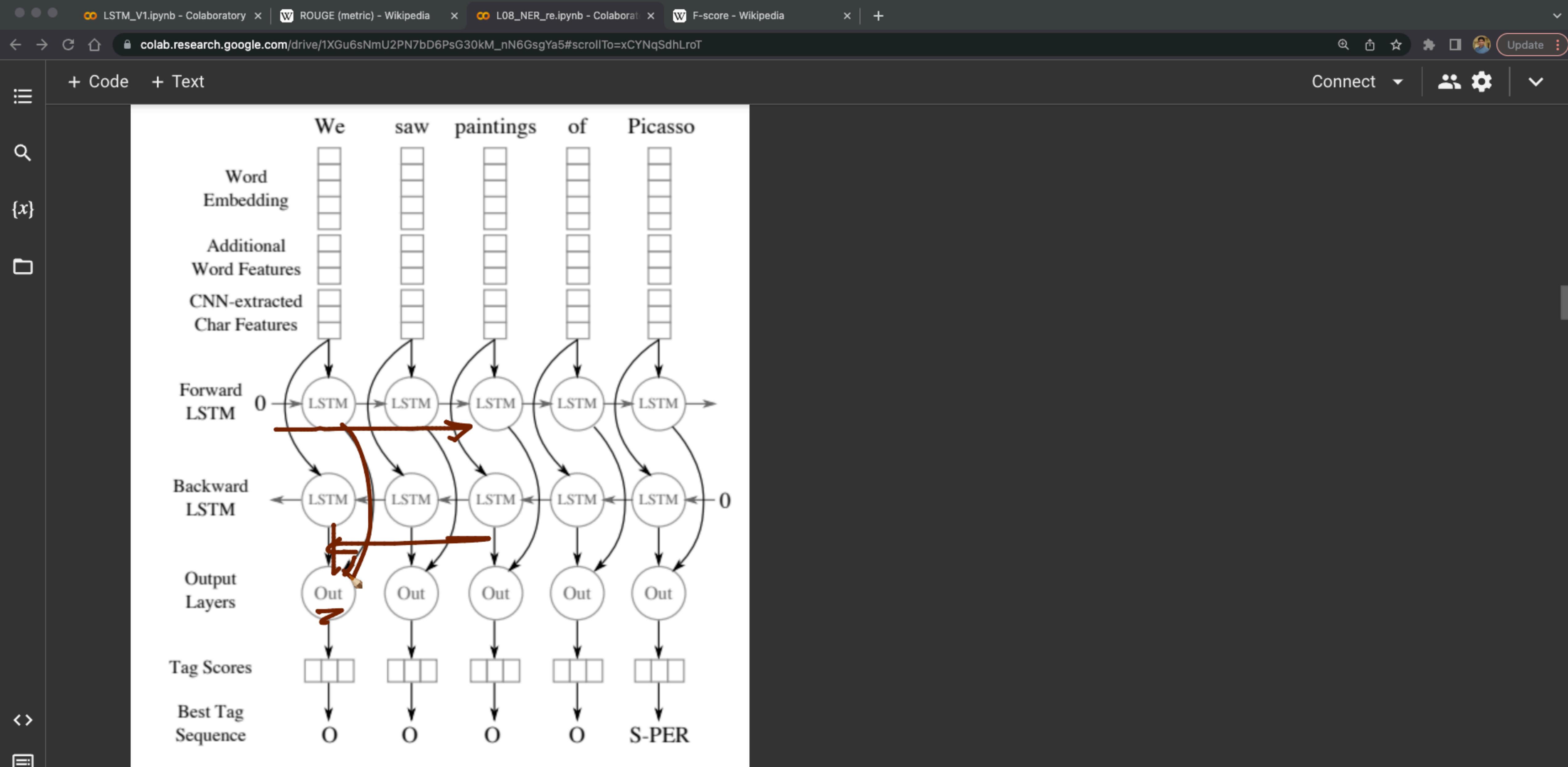
Here,

$p_{\theta}(y|x)$  refers to the probability of calculating a Label

+ Code + Text



- Convolutional neural networks (CNN) have also been investigated for modelling character-level information, among other NLP tasks. So



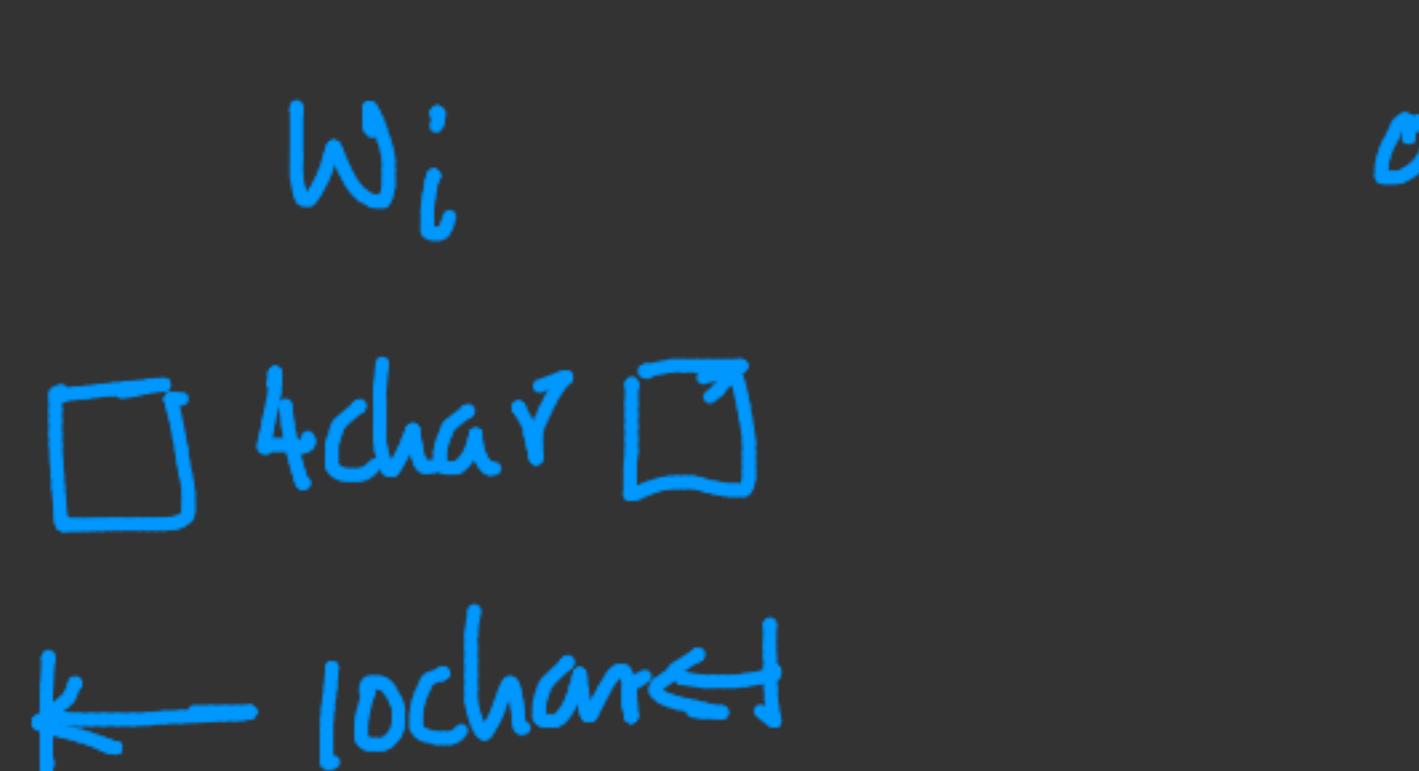
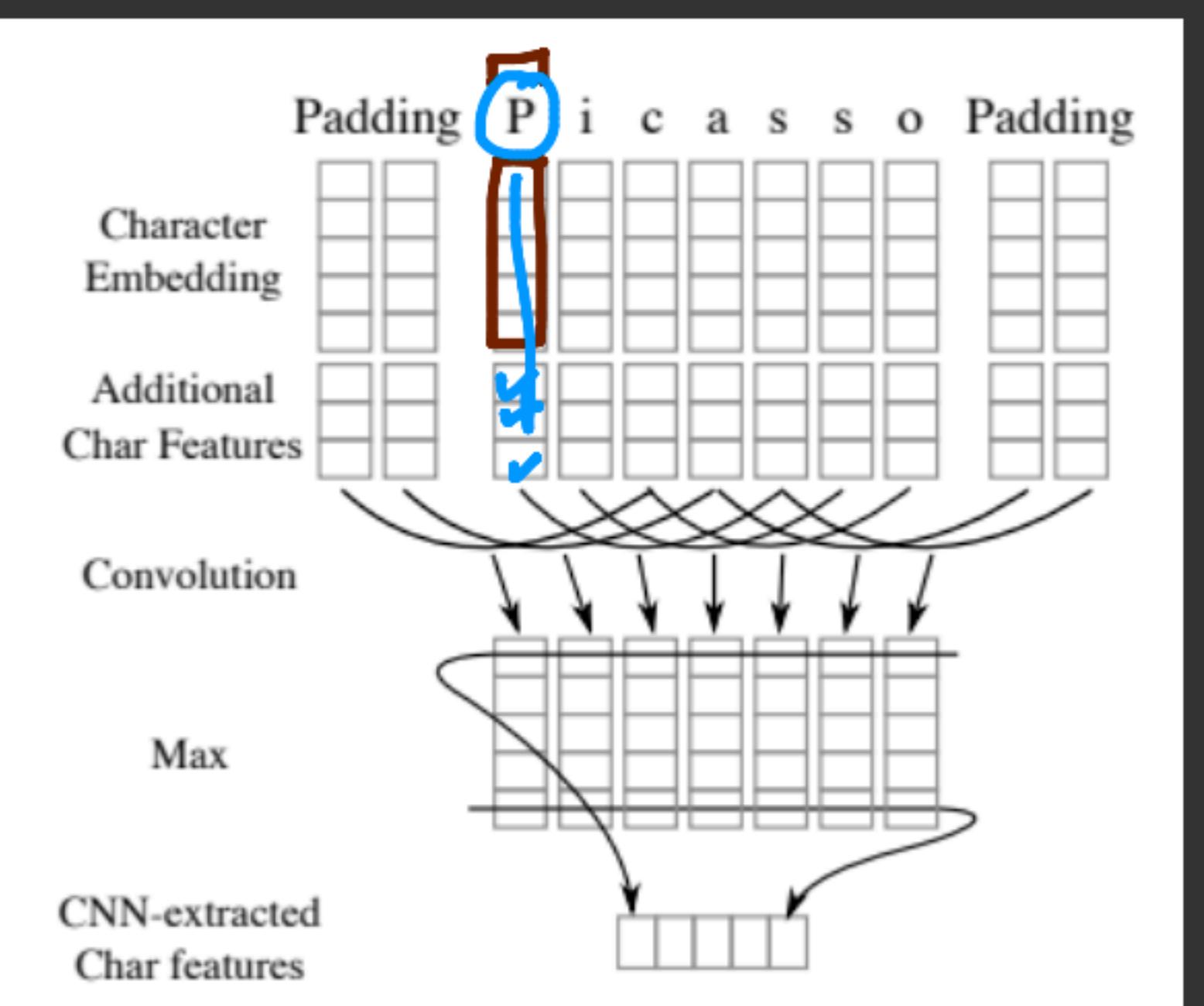
- Convolutional neural networks (CNN) have also been investigated for modelling character-level information, among other NLP tasks. So

+ Code + Text

Connect |  

Best Tag Sequence  
O O O O S-PER

- Convolutional neural networks (**CNN**) have also been investigated for **modelling character-level information**, among other NLP tasks. So in NER CNNs are used to extract character-level features for and POS-tagging respectively



## Bi-directional LSTM CRF

We have how a Bi-directional LSTM model works with another layer of CNN, but for sequence labelling we can consider **CRF**, as

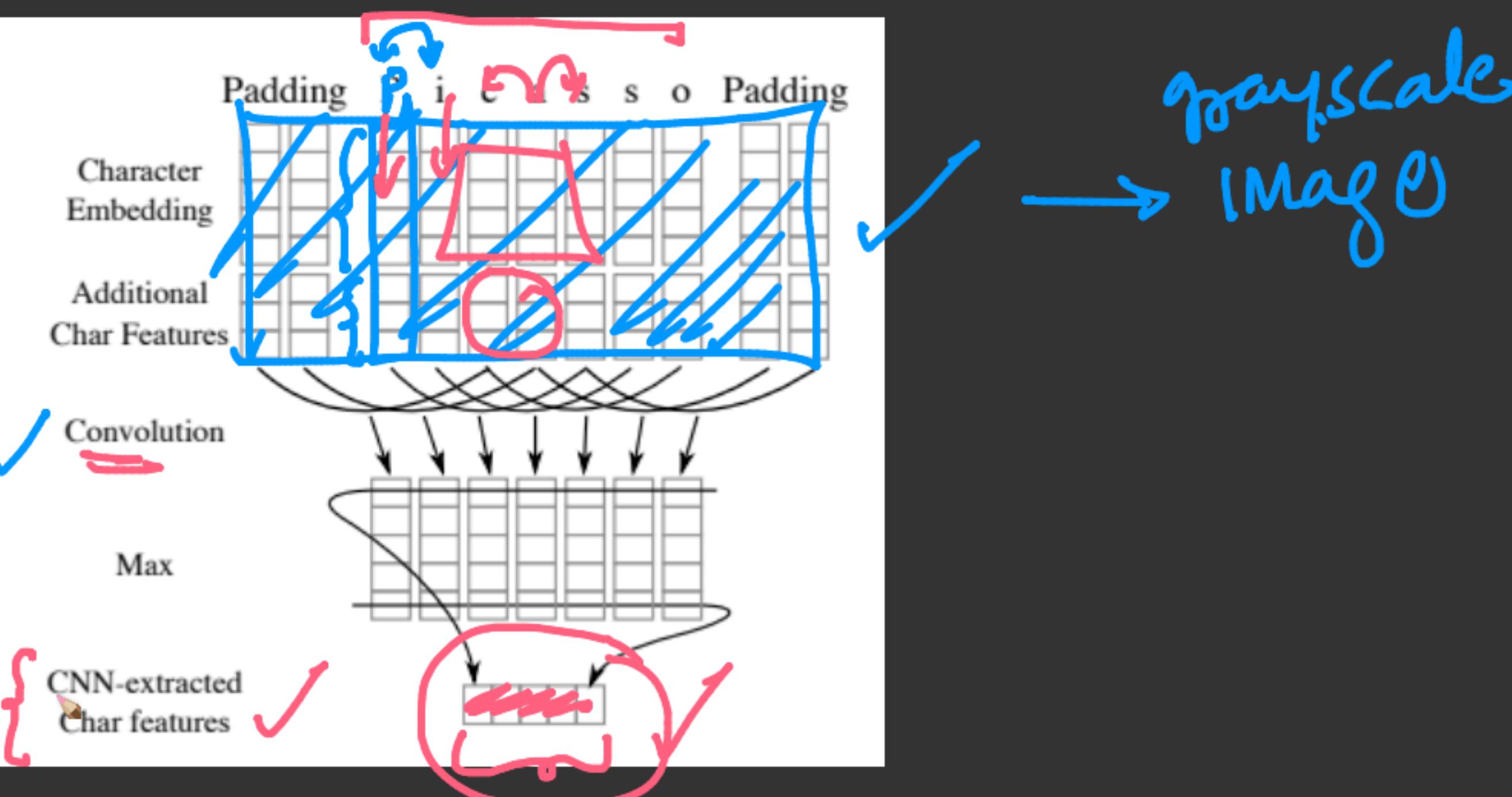
- it is beneficial to consider the **correlations between labels in neighbourhoods** and jointly decode the best chain of labels for a given input

+ Code + Text

Connect |  

Best Tag Sequence  
O O O O S-PER

- Convolutional neural networks (**CNN**) have also been investigated for **modelling character-level information**, among other NLP tasks. So in NER CNNs are used to extract character-level features for and POS-tagging respectively



## Bi-directional LSTM CRF

We have how a Bi-directional LSTM model works with another layer of CNN, but for sequence labelling we can consider **CRF**, as

- it is beneficial to consider the **correlations between labels in neighbourhoods** and jointly decode the best chain of labels for a given input

LSTM\_V1.ipynb - Colaboratory ROUGE (metric) - Wikipedia L08\_NER\_re.ipynb - Colaboratory F-score - Wikipedia

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

+ Code + Text Connect |

Therefore, we model label sequence jointly using a conditional random field (CRF) instead of decoding each label independently.

CRF Layer

Backward LSTM

Forward LSTM

CNN

Char Representation

Word Embedding

We are playing soccer

PRP VBP VBG NN

LSTM\_V1.ipynb - Colaboratory x ROUGE (metric) - Wikipedia x

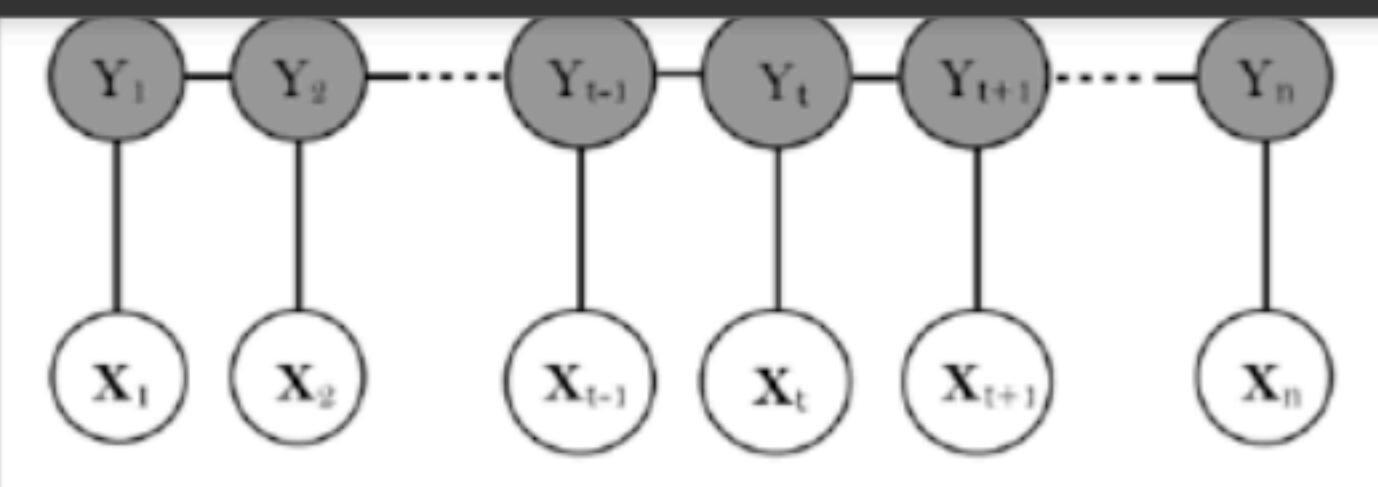
L08\_NER\_re.ipynb - Colaboratory x F-score - Wikipedia x +

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=xCYNqSdhLroT

Update

+ Code + Text

Connect |



Fundamental Equation:

**Consider the below formula:**

CRF:

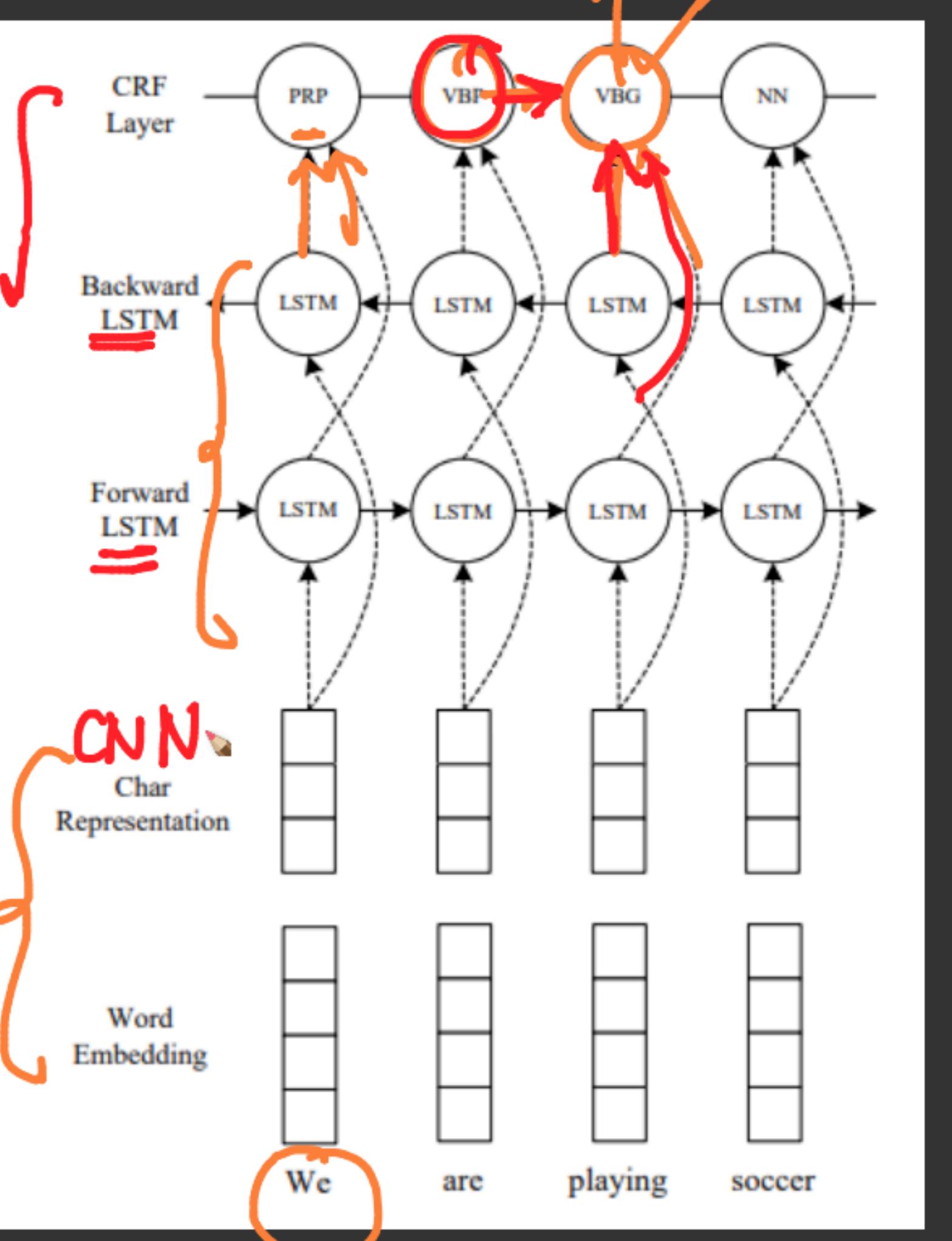
$$p_\theta(y|x) = \frac{\exp(\sum_j w_j F_j(x, y))}{\sum_{y'} \exp(\sum_j w_j F_j(x, y'))}$$

, where  $F_j(x, y) = \sum_{i=1}^L f_j(y_{i-1}, y_i, x, i)$

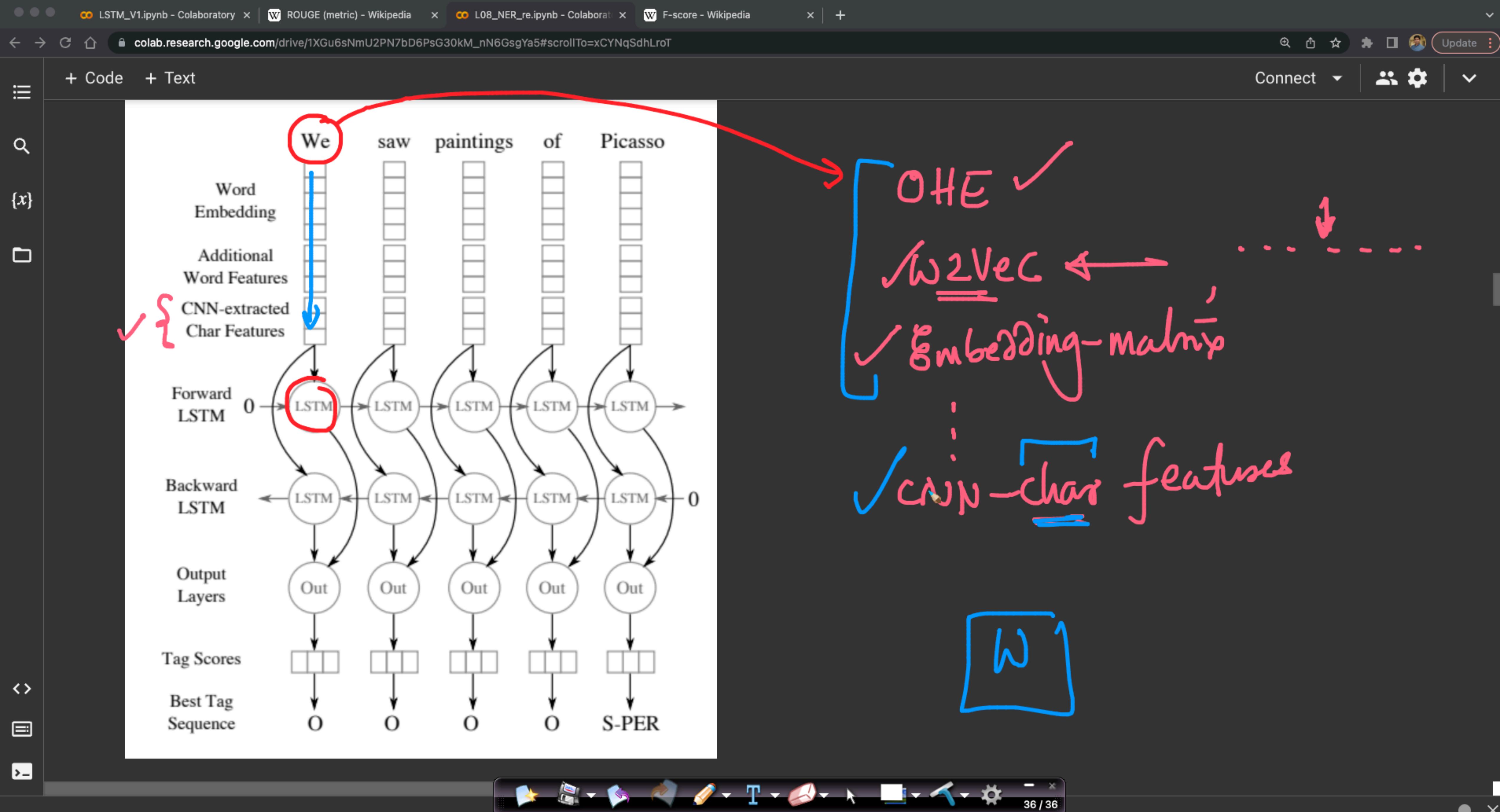
Here,

[+ Code](#)

- Therefore, we model label sequence jointly using a conditional random field (CRF) instead of decoding each label independently.



CRF + BiLSTM



For Preprocessing, we will first put each sentence through a function which will **put each of the words along with its POS and Tag together in a tuple**. Each sentence will form a list of tuples.

```
class SentenceGetter(object):
    """Class to Get the sentence in this format:
    [(Token_1, Part_of_Speech_1, Tag_1), ..., (Token_n, Part_of_Speech_n, Tag_n)]"""
    def __init__(self, data):
        """Args:
            data is the pandas.DataFrame which contains the above dataset"""
        self.n_sent = 1
        self.data = data
        self.empty = False
        agg_func = lambda s: [(w, p, t) for w, p, t in zip(s["Word"].values.tolist(),
                                                          s["POS"].values.tolist(),
                                                          s["Tag"].values.tolist())]
        self.grouped = self.data.groupby("Sentence #").apply(agg_func)
        self.sentences = [s for s in self.grouped]

    def get_next(self):
        """Return one sentence"""
        try:
            s = self.grouped["Sentence: {}".format(self.n_sent)]
            self.n_sent += 1
            return s
        except:
            return None
```

colab.research.google.com/drive/1XGu6sNmU2PN7bD6PsG30kM\_nN6GsgYa5#scrollTo=ec3P9FEALrof

+ Code + Text Connect | |

```
def get_next(self):
    """Return one sentence"""
    try:
        s = self.grouped["Sentence: {}".format(self.n_sent)]
        self.n_sent += 1
        return s
    except:
        return None

getter = SentenceGetter(data)
sent = getter.get_next()
print('This is what a sentence looks like:')
print(sent)
```

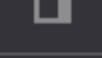
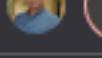
This is what a sentence looks like:  
[('Thousands', 'NNS', 'O'), ('of', 'IN', 'O'), ('demonstrators', 'NNS', 'O'), ('have', 'VBP', 'O'), ('marched', 'VBN', 'O'), ('through', 'O'), ('the', 'DT', 'O'), ('march', 'VBD', 'O'), ('in', 'IN', 'O'), ('the', 'DT', 'O'), ('streets', 'NNS', 'O'), ('.', '.', 'O')]

# Get all the sentences  
sentences = getter.sentences  
# Plot sentence by lenght  
plt.hist([len(s) for s in sentences], bins=50)  
plt.title('Token per sentence')  
plt.xlabel('Len (number of token)')  
plt.ylabel('# samples')  
plt.show()

Token per sentence

38 / 38

LSTM\_V1.ipynb - Colaboratory x ROUGE (metric) - Wikipedia x L08\_NER\_re.ipynb - Colaboratory x F-score - Wikipedia x + colab.research.google.com/drive/1jSDd8nTYGpPmDD4JsXF-FeWJGOK1boX#scrollTo=e0dd5839

+ Code + Text Last edited on 11 November Connect |  

rouge\_f.append(fmeas)

```
prediction_df['rouge_pr'] = rouge_pr
prediction_df['rouge_rc'] = rouge_rc
prediction_df['rouge_f'] = rouge_f
prediction_df.head(10)
```

Actual Predicted rouge\_pr rouge\_rc rouge\_f

	Actual	Predicted	rouge_pr	rouge_rc	rouge_f
0	start deepika denies starring as amrita in bha...	start deepika to play in biopic on padmavati ...	0.363636	0.363636	0.363636
1	start shahid takes care of ishaan me on being ...	start my son taimur is a dream to be a on my ...	0.285714	0.333333	0.307692
2	start pilot dies on board airways plane mid ai...	start pilot flyers to fly plane with emergenc...	0.400000	0.400000	0.400000
3	start celebrations begin at ram nath kovind s ...	start pm modi to be called by a dictator in u...	0.166667	0.200000	0.181818
4	start guest their heads in protest in mp end	start women wear skirts in protest against end	0.500000	0.444444	0.470588
5	start can t tolerate weinstein like behaviour ...	start i don t know how to be used to stop fac...	0.266667	0.400000	0.320000
6	start gujarat poll panel orders probe in contr...	start ec orders removal of using fake news on...	0.250000	0.272727	0.260870
7	start facebook suspends canadian firm amid dat...	start facebook suspends facebook over fake ne...	0.600000	0.600000	0.600000
8	start iraq s first non arab president passes a...	start iraq s 1st state citizen to be held in ...	0.333333	0.400000	0.363636
9	start us scientists propose new organ in human...	start scientists propose new type of human sy...	0.666667	0.600000	0.631579

ROUGE-gram

## Scope for improvement

As we can see that the predicted summary may not be accurate we can try to improve it



+ Code + Text Last edited on 11 November

Connect ▾ |   | 

rouge f.append(fmeas)

```
prediction_df['rouge_pr'] = rouge_pr
prediction_df['rouge_rc'] = rouge_rc
prediction_df['rouge_f'] = rouge_f
prediction_df.head(10)
```

A set of small, light-gray navigation icons located at the bottom right of the slide.

	Actual	Predicted	rouge_pr	rouge_rc	rouge_f
0	start <u>deepika</u> denies starring as amrita in bha...	start <u>deepika</u> to play in biopic on padmavati ...	0.363636	0.363636	0.363636
1	start shahid takes care of ishaan me on being ...	start my son taimur is a dream to be a on my ...	0.285714	0.333333	0.307692
2	start pilot dies on board airways plane mid ai...	start pilot flyers to fly plane with emergenc...	0.400000	0.400000	0.400000
3	start celebrations begin at ram nath kovind s ...	start pm modi to be called by a dictator in u...	0.166667	0.200000	0.181818
4	start guest their heads in protest in mp end	start women wear skirts in protest against end	0.500000	0.444444	0.470588
5	start can t tolerate weinstein like behaviour ...	start i don t know how to be used to stop fac...	0.266667	0.400000	0.320000
6	start gujarat poll panel orders probe in contr...	start ec orders removal of using fake news on...	0.250000	0.272727	0.260870
7	start facebook suspends canadian firm amid dat...	start facebook suspends facebook over fake ne...	0.600000	0.600000	0.600000
8	start iraq s first non arab president passes a...	start iraq s 1st state citizen to be held in ...	0.333333	0.400000	0.363636
9	start us scientists propose new organ in human...	start scientists propose new type of human sy...	0.666667	0.600000	0.631579

## → Scope for improvement

As we can see that the predicted summary may not be good enough, we can try to improve it.

LSTM\_V1.ipynb - Colaboratory x ROUGE (metric) - Wikipedia x L08\_NER\_re.ipynb - Colaboratory x F-score - Wikipedia x + colab.research.google.com/drive/1jSDd8nTYGpPmDD4JsXF-1FeWJGOK1boX#scrollTo=e0dd5839

+ Code + Text Last edited on 11 November Connect |

rouge\_f.append(fmeas)

```
prediction_df['rouge_pr'] = rouge_pr
prediction_df['rouge_rc'] = rouge_rc
prediction_df['rouge_f'] = rouge_f
prediction_df.head(10)
```

	Actual	Predicted	rouge_pr	rouge_rc	rouge_f
0	start deepika denies starring as amrita in bha...	start deepika to play in biopic on padmavati ...	0.363636	0.363636	0.363636
1	start shahid takes care of ishaan me on being ...	start my son taimur is a dream to be a on my ...	0.285714	0.333333	0.307692
2	start pilot dies on board airways plane mid ai...	start pilot flyers to fly plane with emergenc...	0.400000	0.400000	0.400000
3	start celebrations begin at ram nath kovind s ...	start pm modi to be called by a dictator in u...	0.166667	0.200000	0.181818
4	start guest their heads in protest in mp end	start women wear skirts in protest against end	0.500000	0.444444	0.470588
5	start can t tolerate weinstein like behaviour ...	start i don t know how to be used to stop fac...	0.266667	0.400000	0.320000
6	start gujarat poll panel orders probe in contr...	start ec orders removal of using fake news on...	0.250000	0.272727	0.260870
7	start facebook suspends canadian firm amid dat...	start facebook suspends facebook over fake ne...	0.600000	0.600000	0.600000
8	start iraq s first non arab president passes a...	start iraq s 1st state citizen to be held in ...	0.333333	0.400000	0.363636
9	start us scientists propose new organ in human...	start scientists propose new type of human sy...	0.666667	0.600000	0.631579

## Scope for improvement

As we can see that the predicted summary may not always be accurate, we can try to improve it.





+ Code + Text

