APPLYING NER IN MEDICINE RECORDS

GROUP 4

BI12-452 Tran The Trung
BI12-059 Phan Thanh Binh
BA11-093 Tran Minh Trung
BI12-263 Chau Phan Phuong Mai

TABLE OF CONTENTS

01.

INTRODUCTION

02.

DATASET

03.

METHODS

04.

PERFORMANCE EVALUATION

05.

FUTURE WORKS

01.

INTRODUCTION

MOTIVATION

Electronic Heath Records

- Primary source of information for clinicians tracking
- Bring significant advancement for the downstream task

EXAMPLES

- The reason for administering drugs
- Previous disorders of the patient
- The outcome of past treatments

MANUALLY ABSTRACTION

- Highly expensive
- Time-consuming
- Error prone process

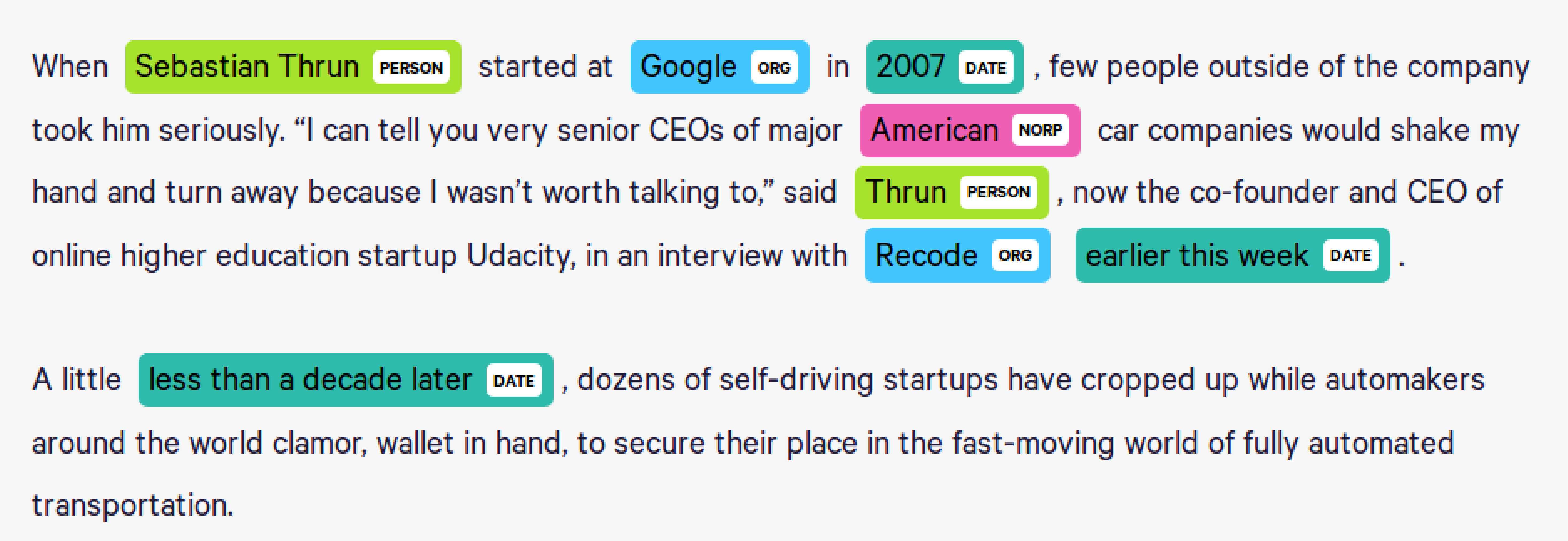
OBJECTIVES

Implement a model which can automate extract the medical information from EHRs from two Name Entity Recognition approaches:

- Sequence labelling-based
- Span-based

↳ Save effort, time & money

OBJECTIVES



When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

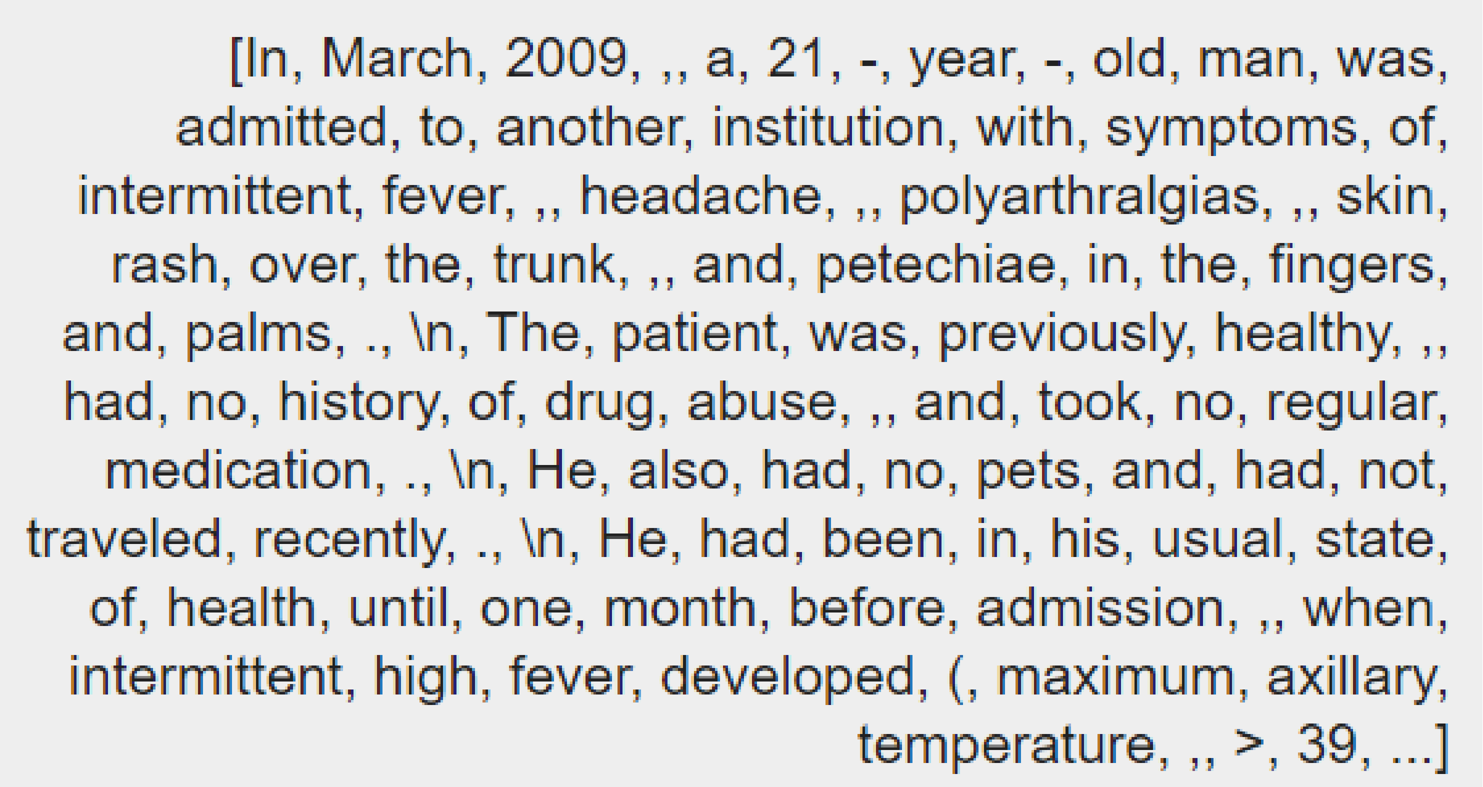
Example for the model’s output

02.
DATASET

MACCROBAT DATASETS

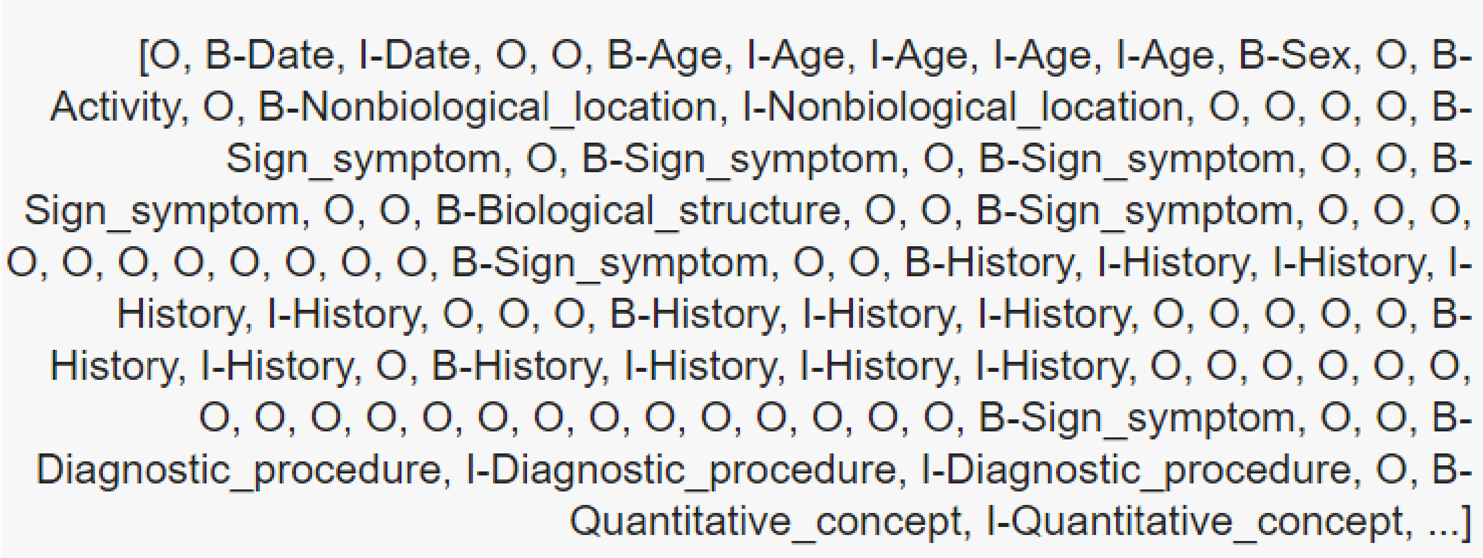
- Source: Huggingface
- 200 source documents
- Tag labels: 41 special terms

- Input: List of tokenized words
- Output: Label of words in BIO POS tagging
- Train set: 90%, Test set: 10%



[In, March, 2009, ,, a, 21, -, year, -, old, man, was, admitted, to, another, institution, with, symptoms, of, intermittent, fever, ,, headache, ,, polyarthralgias, ,, skin, rash, over, the, trunk, ,, and, petechiae, in, the, fingers, and, palms, ., \n, The, patient, was, previously, healthy, ,, had, no, history, of, drug, abuse, ,, and, took, no, regular, medication, ., \n, He, also, had, no, pets, and, had, not, traveled, recently, ., \n, He, had, been, in, his, usual, state, of, health, until, one, month, before, admission, ,, when, intermittent, high, fever, developed, (, maximum, axillary, temperature, ,, >, 39, ...]

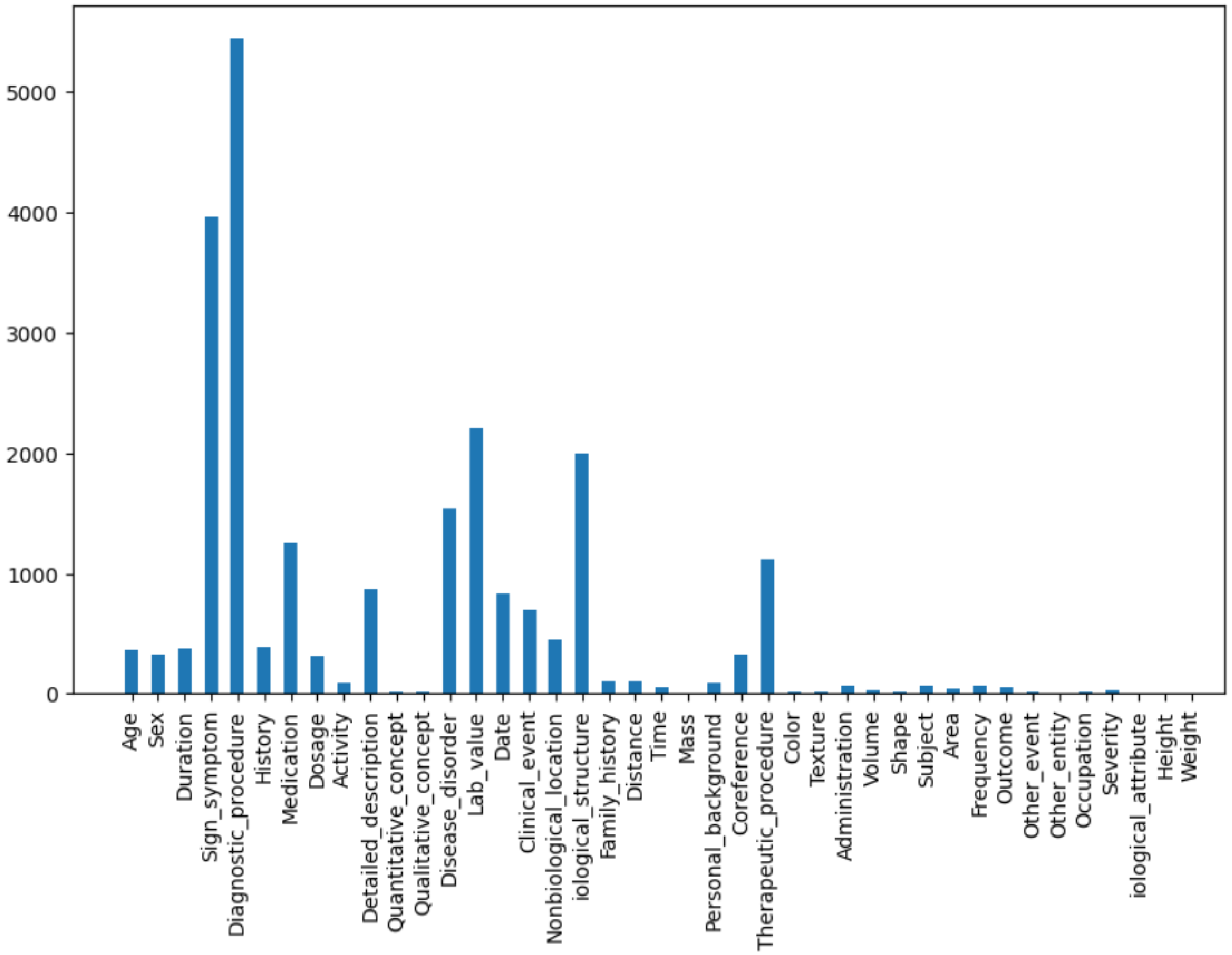
Example of the tokenized document



[O, B-Date, I-Date, O, O, B-Age, I-Age, I-Age, I-Age, I-Age, B-Sex, O, B-Activity, O, B-Nonbiological_location, I-Nonbiological_location, O, O, O, O, B-Sign_symptom, O, B-Sign_symptom, O, B-Sign_symptom, O, O, B-Sign_symptom, O, O, B-Biological_structure, O, O, B-Sign_symptom, O, O, O, O, O, O, O, O, O, O, O, B-Sign_symptom, O, O, B-History, I-History, I-History, I-History, I-History, O, O, O, B-History, I-History, I-History, O, O, O, O, O, B-History, I-History, O, B-History, I-History, I-History, I-History, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, B-Sign_symptom, O, O, B-Diagnostic_procedure, I-Diagnostic_procedure, I-Diagnostic_procedure, O, B-Quantitative_concept, I-Quantitative_concept, ...]

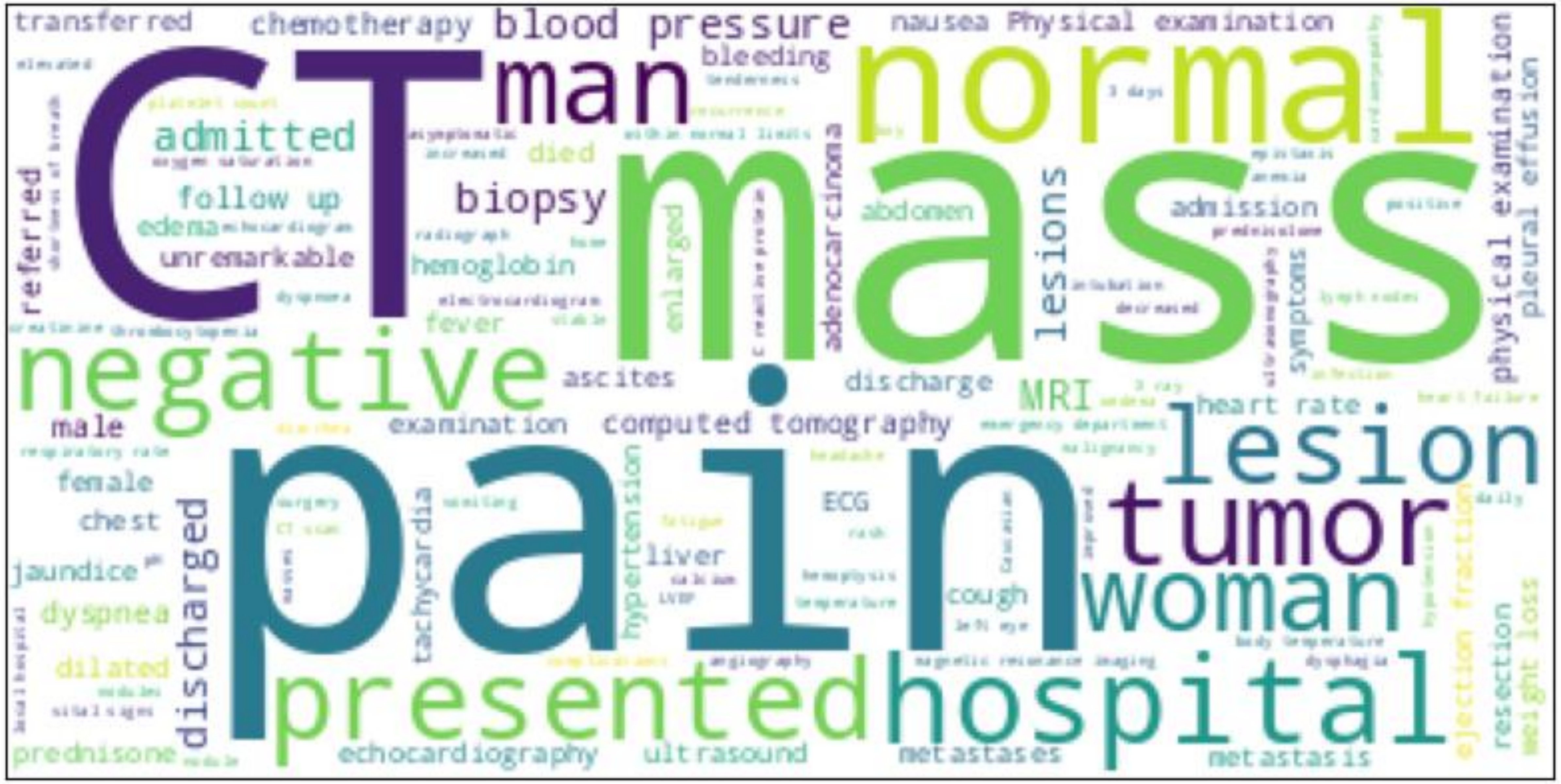
Example of the POS tagging

MACCROBAT DATASETS



Histogram of entities in the dataset

MACCROBAT DATASETS

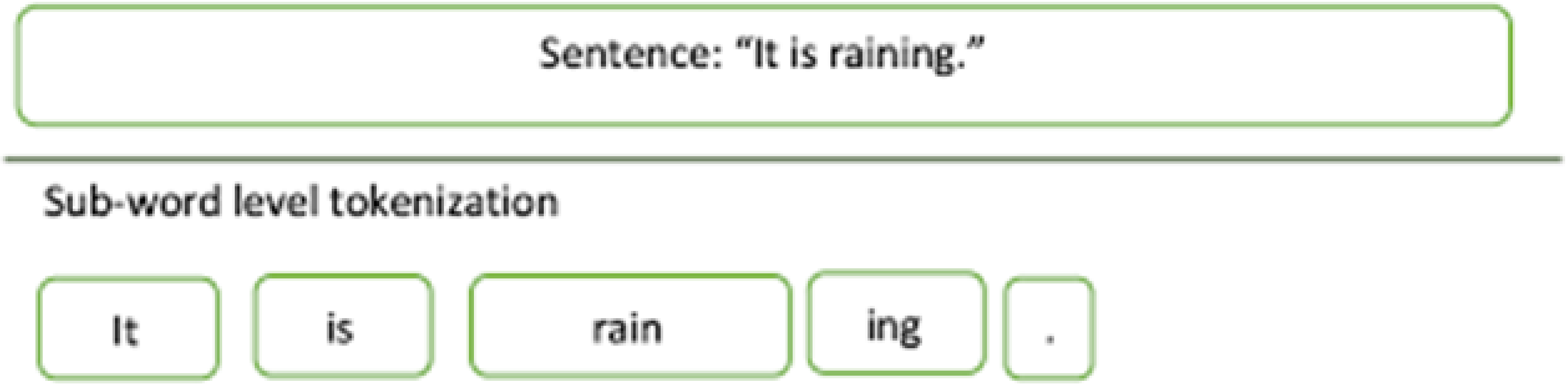


WordClouds of words in the dataset

DATA PREPROCESSING: Tokenize

Byte Pair Encoding (BPE)

- Operates on character or byte level
- Resolve the problem of Unknown tokens

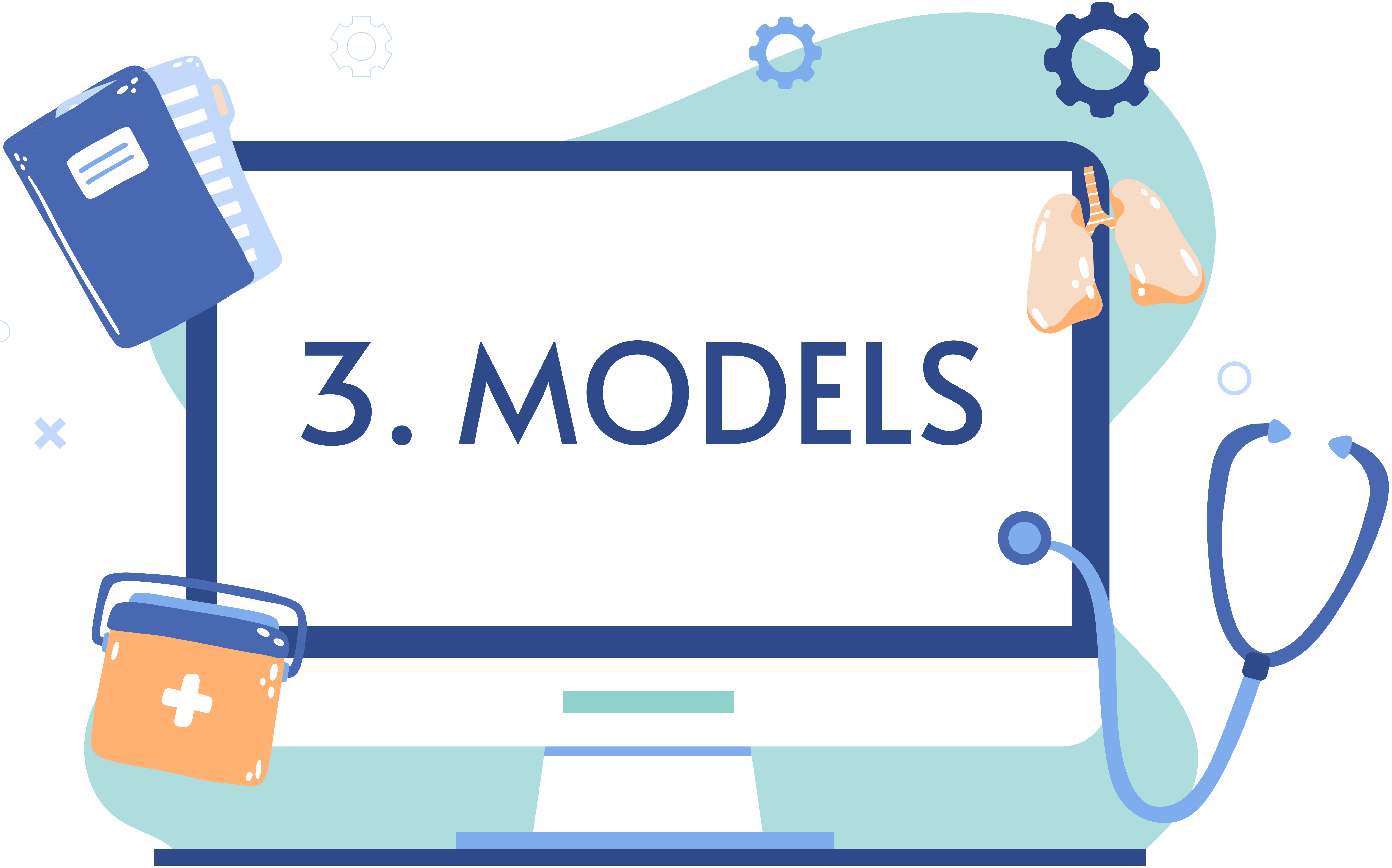


Sentence: "It is raining."

Sub-word level tokenization

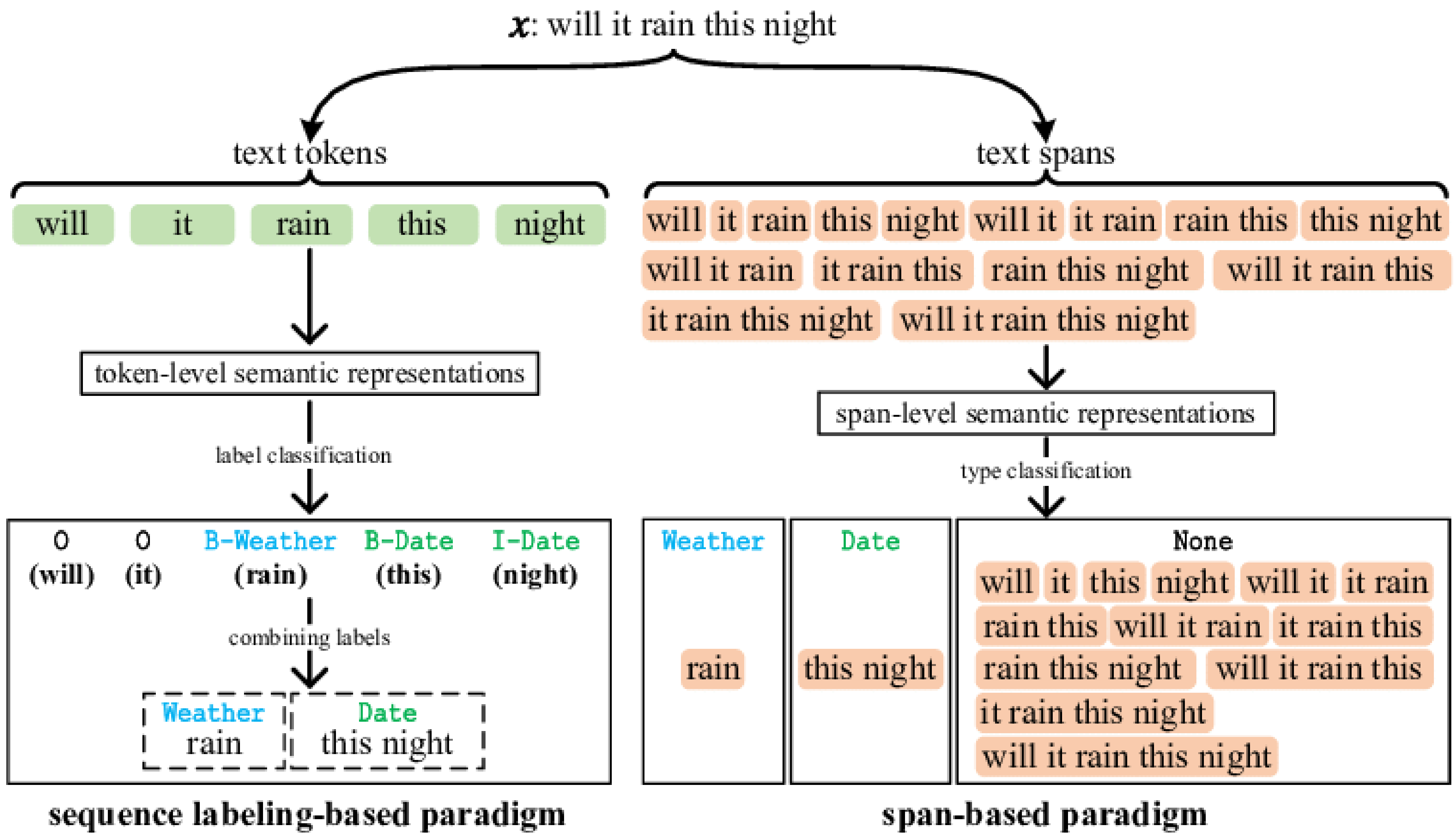
It | is | rain | ing | .

Example of word tokenization



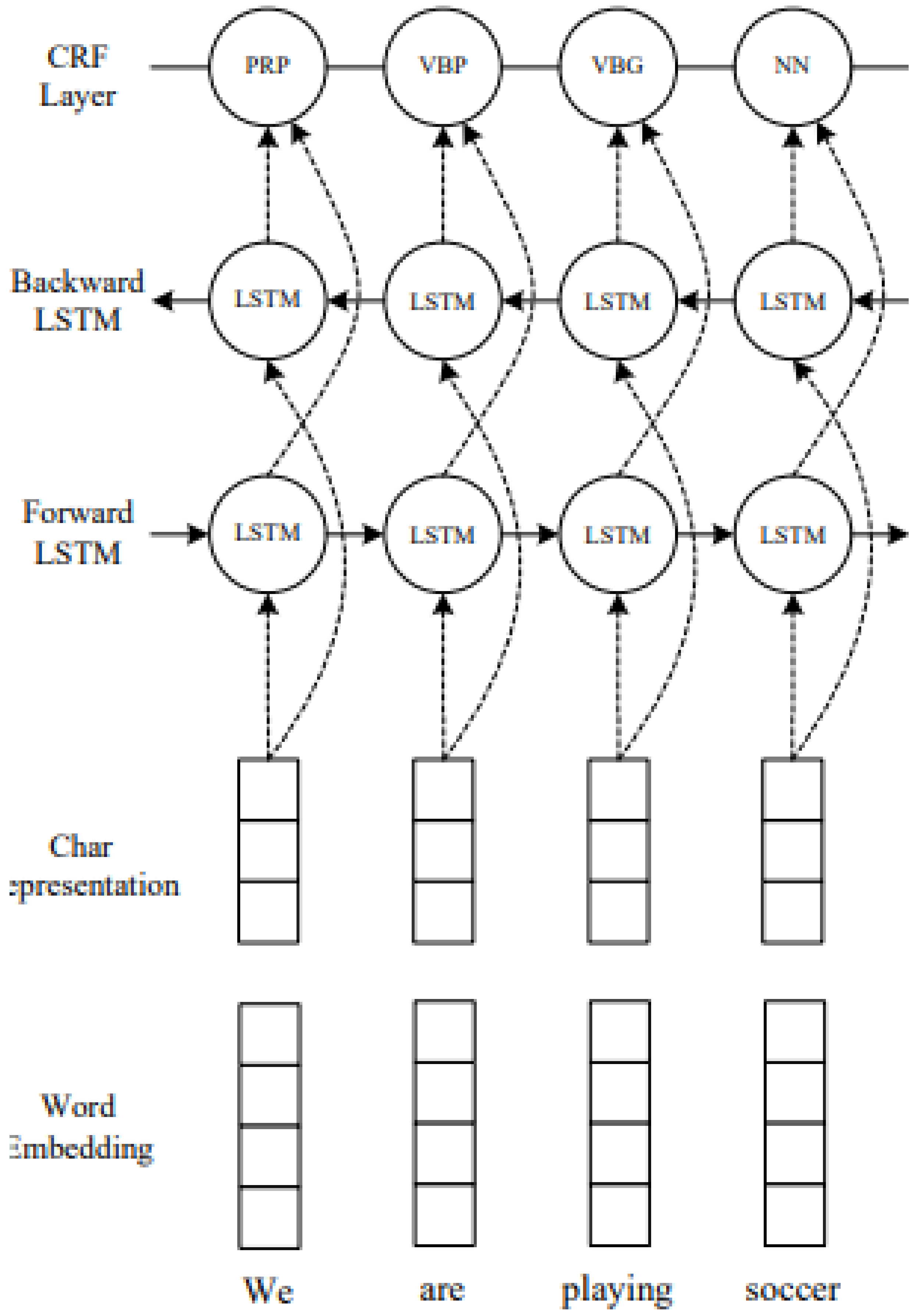
3. MODELS

Name Entity Recognition Approaches



Sequence labelling-based

Bi-LSTM-CNN-CRF architecture



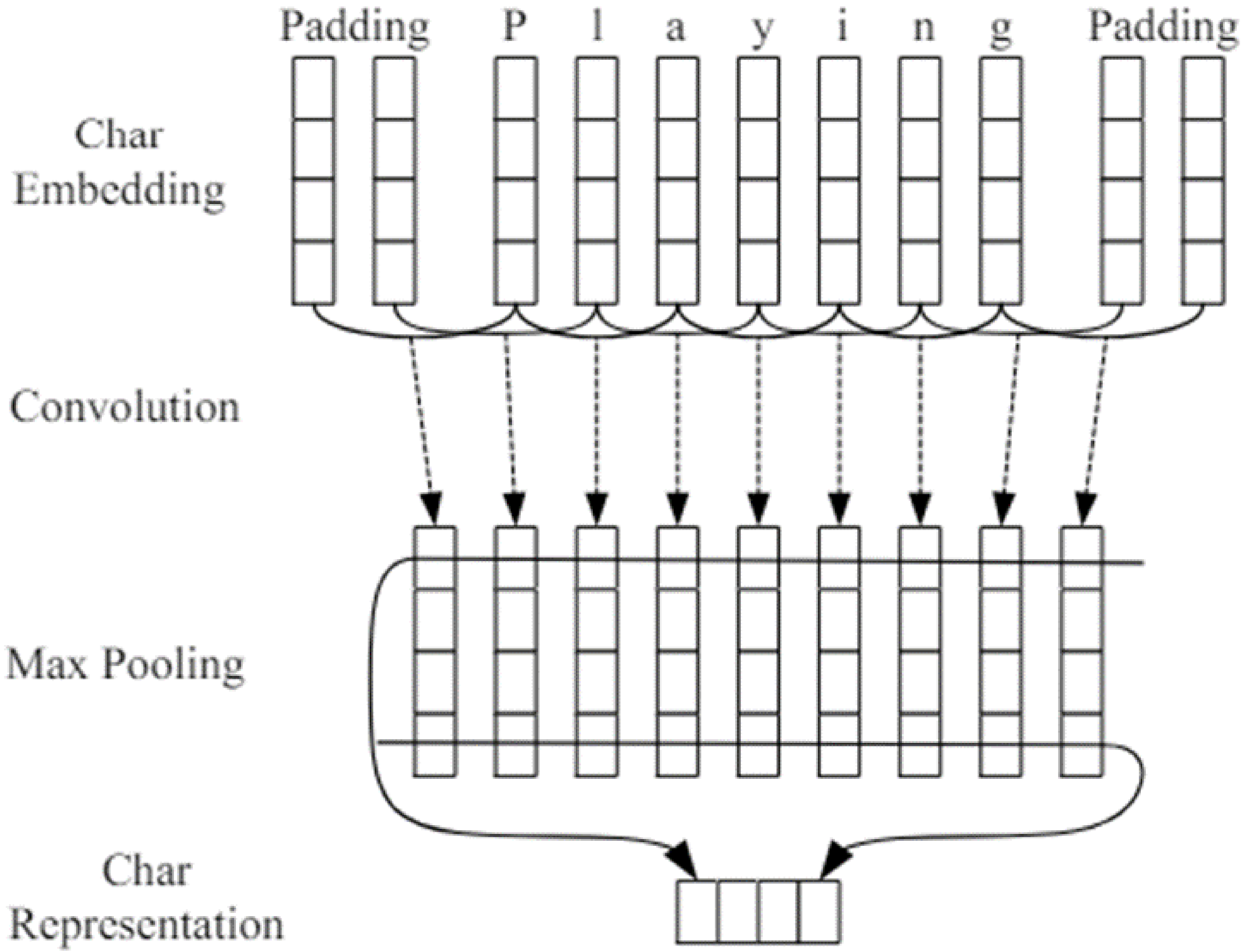
The architechture of Bi-LSTM-CNN-CRF

Input: A squence of words
Output: Tag lables for each words

Sequence labeling based:

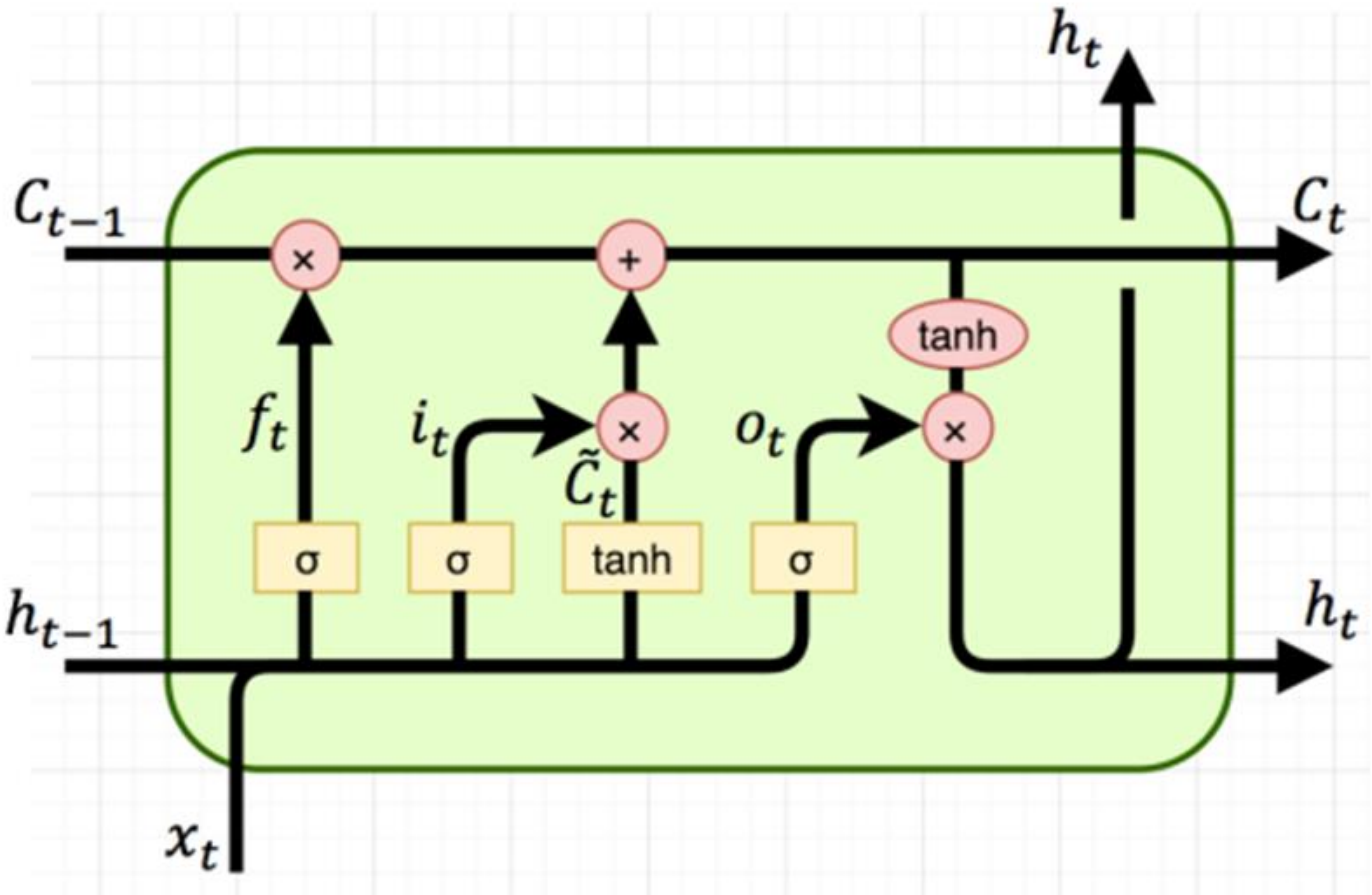
- Assigns a label to each word or token indicating its entity type in BIO format
- Processes input sequentially

Convolution Neural Network



Char Representation process

Long Short-Term Memory



Gaining model context information of each word

Conditional Random Field

$$p(Y|X) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)}{\sum_{Y' \in \mathscr{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)}$$

Jointly decode labels for the whole sentence

Span-based

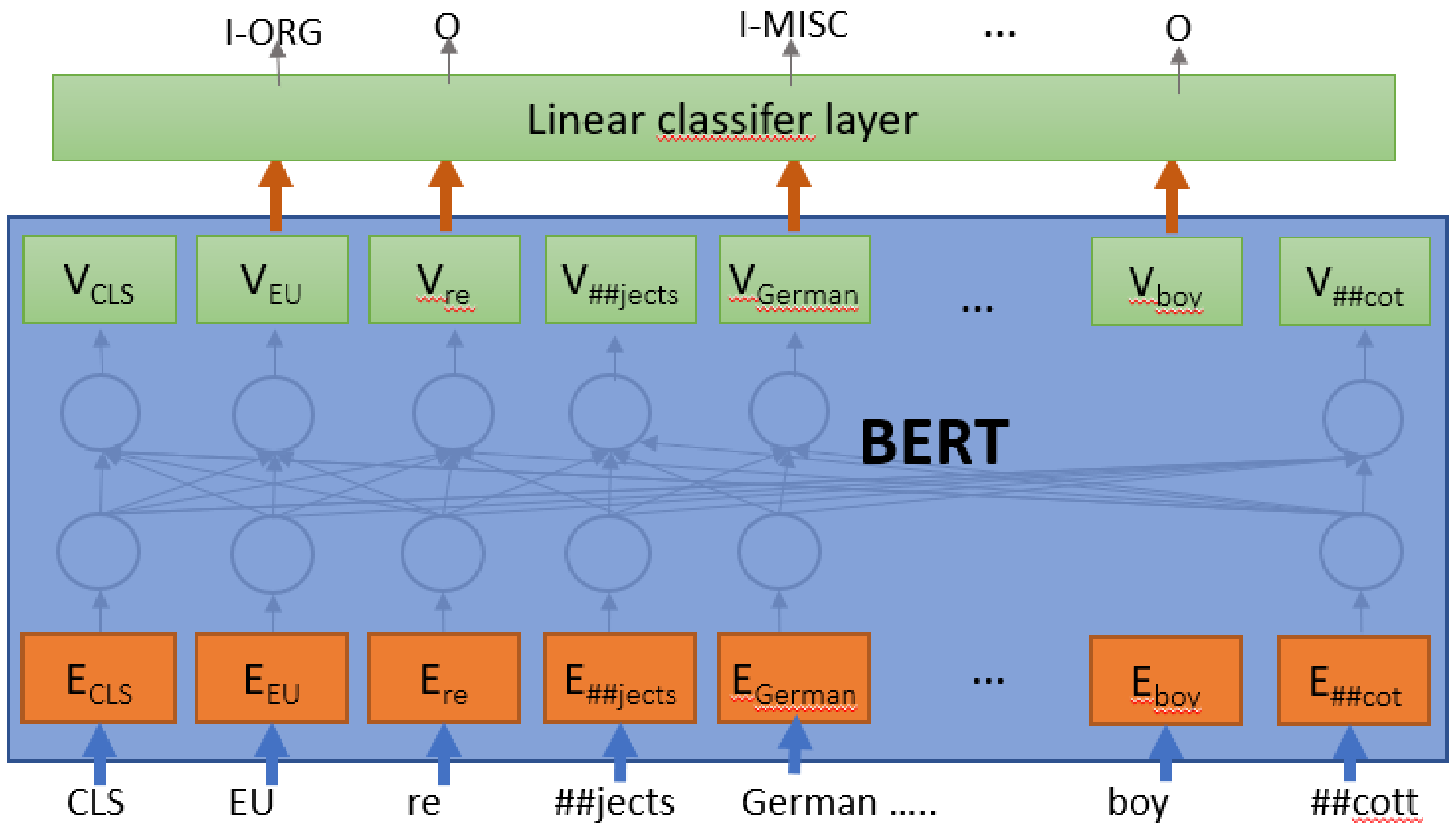
BERT & its variants

BERT

Input: A squence of words
Output: Label Entity of whole span

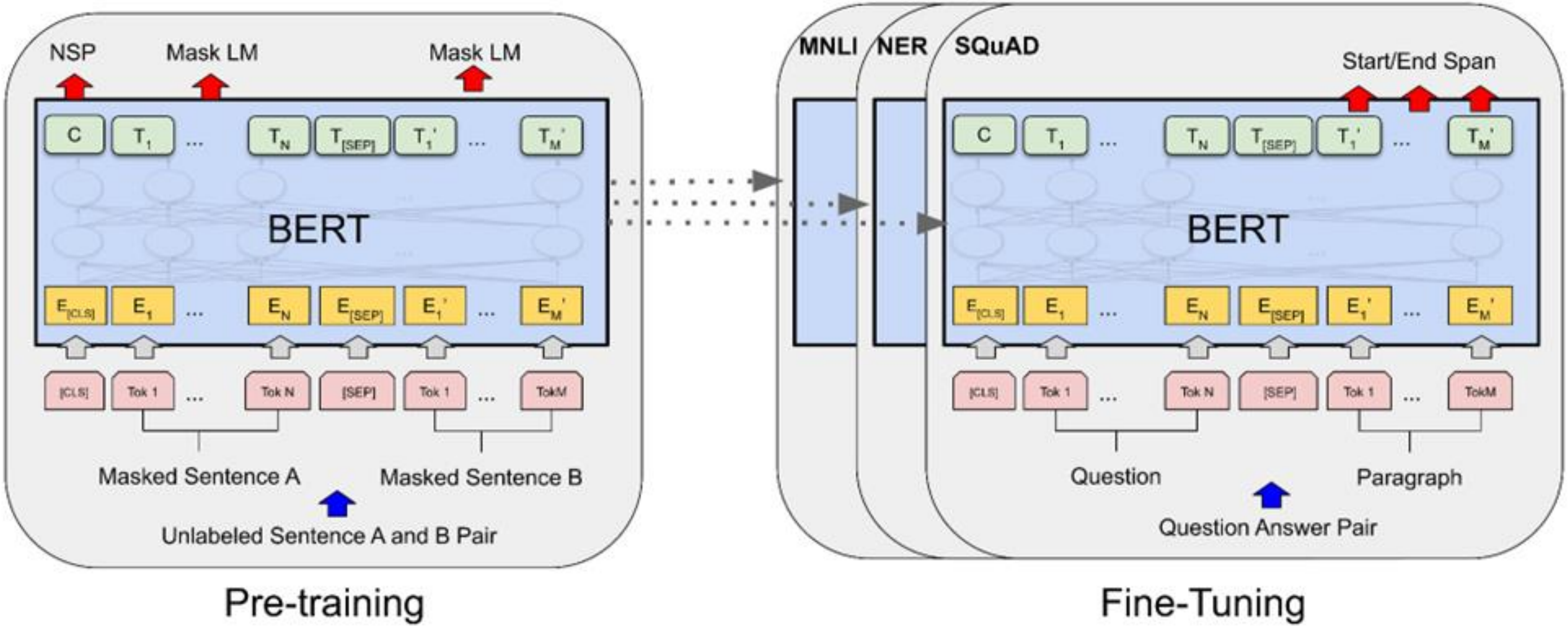
Span-based:

- Predicts the start and end positions of each entity directly.
- All spans or sentences are executed in parallel

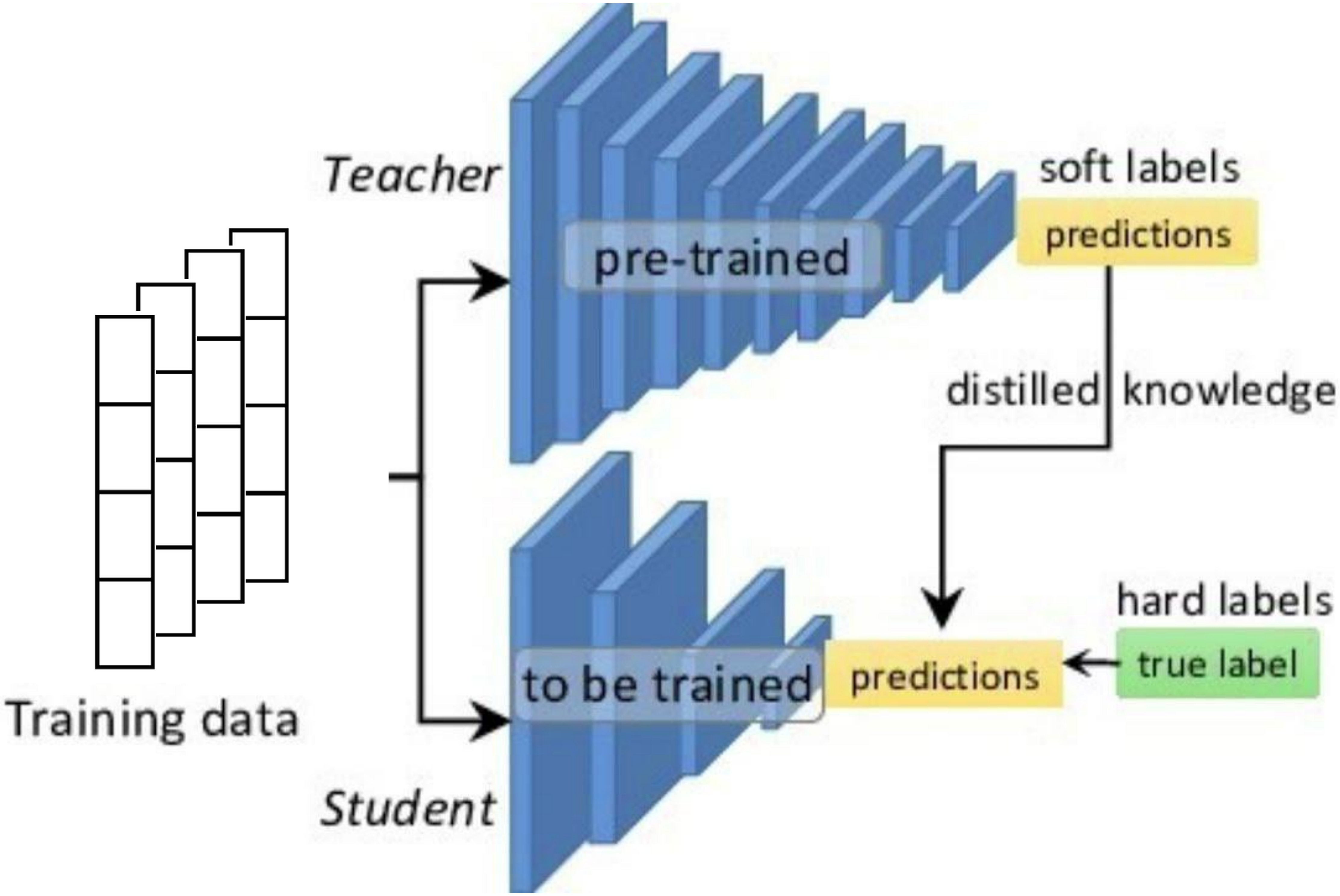


BERT architechture

BERT


NSP
Mask LM
Mask LM
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Question
Paragraph
Question Answer Pair
Fine-Tuning

DistilBERT


Teacher
pre-trained
soft labels
predictions
distilled knowledge
Training data
hard labels
true label
to be trained
predictions
Student

RoBERTa

- Changes some hyperparameters (peak learning rate, batch sizes, Adam epsilon, ...)
- Implements some other optimizations (dynamic masking, modification in NSP task, ...)

DeBERTa

- Introduces a disentangled self-attention mechanism
- Embeds the absolute position information
- Implements dynamic masking

04.

Performance Evaluation

Performance Evaluation of different models

	CNN-LSTM-CRF	BERT	DistilBERT	RoBERTa	DeBERTa
F1 score	**0.865**	**0.818**	**0.782**	**0.82**	**0.84**

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F1 score formula

INFERENCE

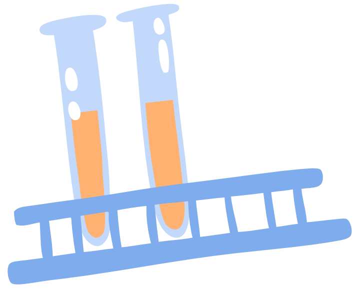
Input_text:

"A 63-year-old woman with no known cardiac history presented with a sudden onset of dyspnea requiring intubation and ventilatory support out of the hospital. She denied preceding symptoms of chest discomfort, palpitations, syncope, or infection. The patient was afebrile and normotensive, with a sinus tachycardia of 140 beats/min."

→

Output_text:

```
[('63 year old', 'Age'),
 ('woman', 'Sex'),
 ('no known cardiac history', 'History'),
 ('presented', 'Clinical_event'),
 ('dyspnea', 'Sign_symptom'),
 ('intubation', 'Therapeutic_procedure'),
 ('ventilatory support', 'Therapeutic_procedure'),
 ('hospital', 'Nonbiological_location'),
 ('discomfort', 'Sign_symptom'),
 ('palpitations', 'Sign_symptom'),
 ('syncope', 'Sign_symptom'),
 ('infection', 'Sign_symptom'),
 ('afebrile', 'Sign_symptom'),
 ('normotensive', 'Sign_symptom'),
 ('tachycardia', 'Sign_symptom')]
```



05.

FUTURE WORKS

01 Find out different methods apart from LLMs.

02 Spend more time on training models.

03 Increase the number of detected entities by training with others datasets.

04 Focus more on in-depth data in the field of medicine.

THANKS!