NLP Exercise

Part-of-Speech Tagging
Medical Named Entity Recognition

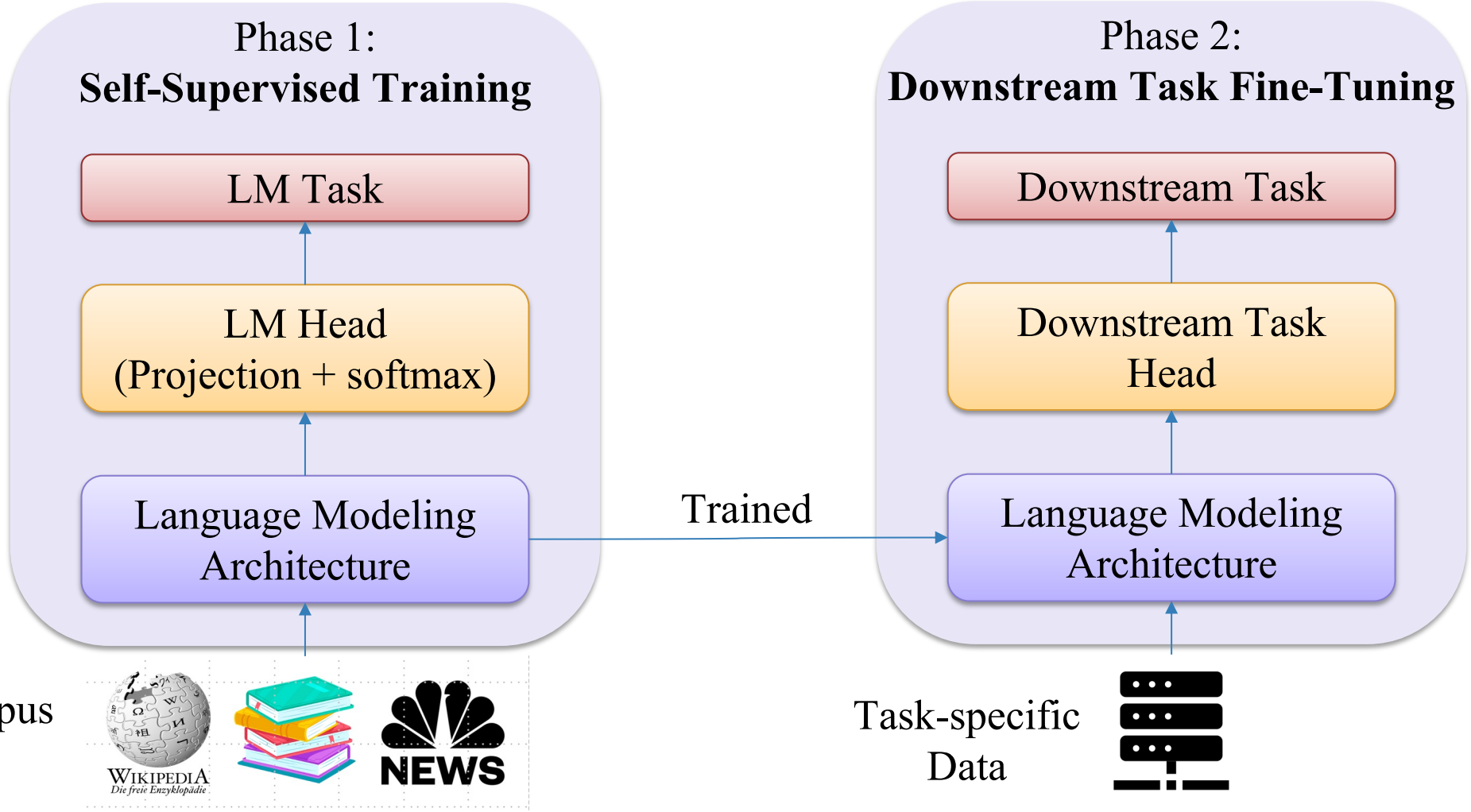
AI VIET NAM
Nguyen Quoc Thai

Outline

- Introduction
- Part-of-Speech Tagging (POS)
- Named Entity Recognition (NER)
- Medical Named Entity Recognition

Introduction

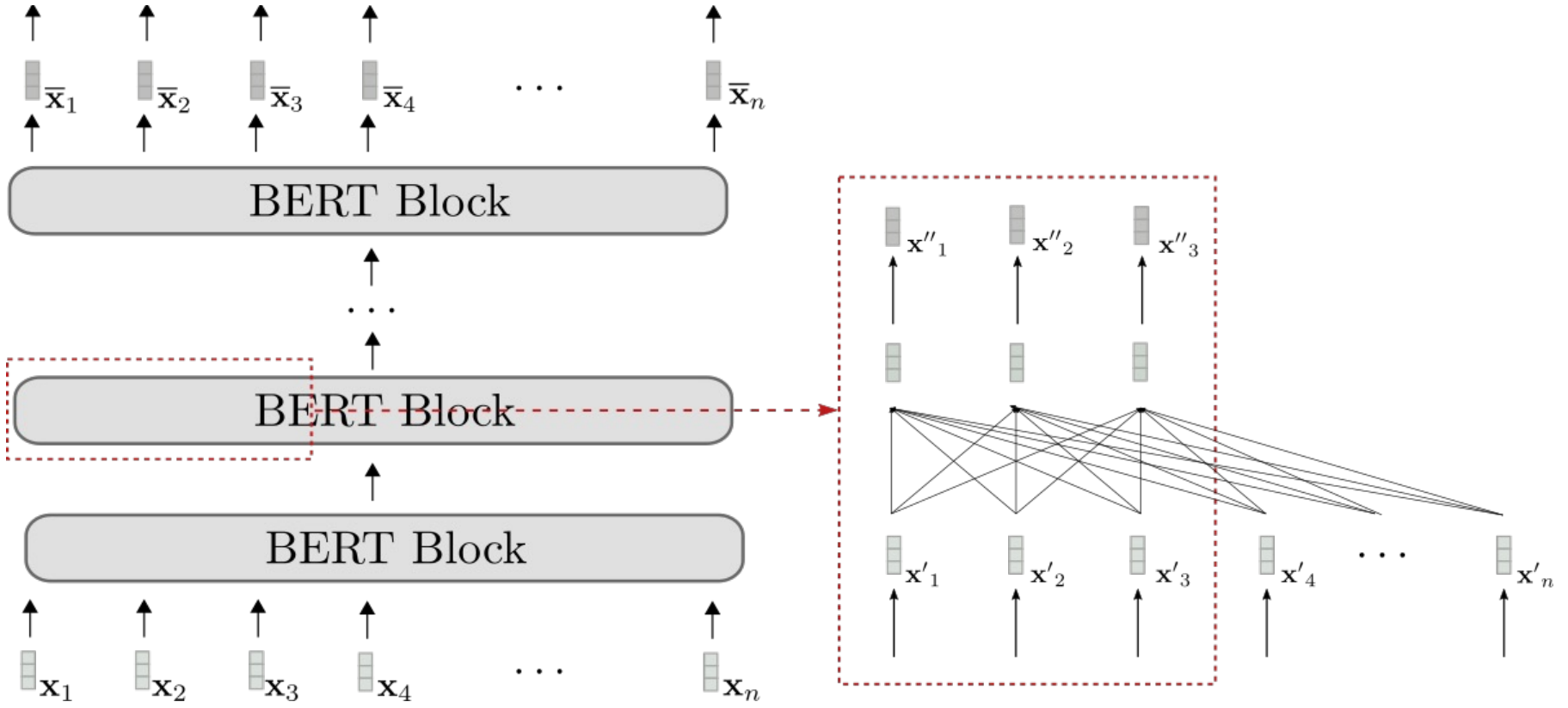
Pre-trained Models for Text



Introduction

BERT: Encoder Model

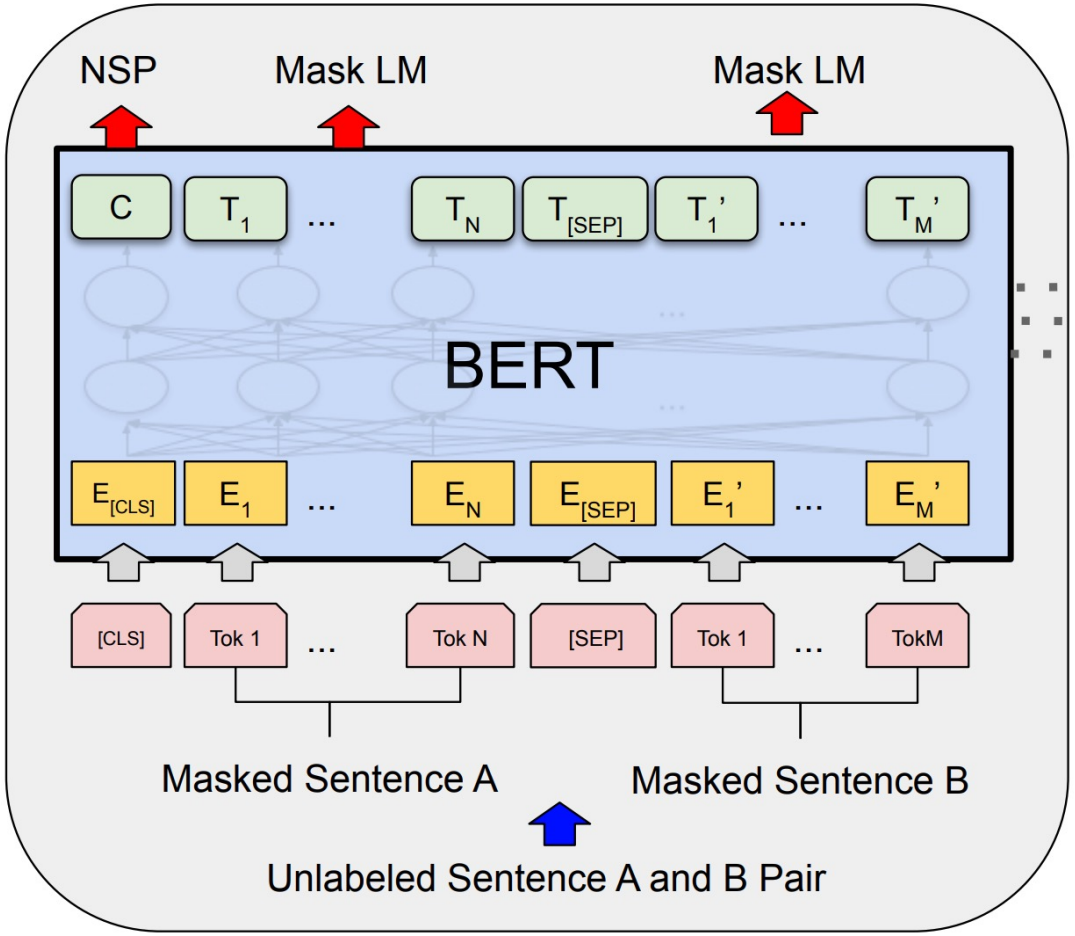
- Maps an input sequence to a contextualized sequence: $f_{\theta_{BERT}}: X_{1:n} \rightarrow \overline{X}_{1:n}$



Introduction

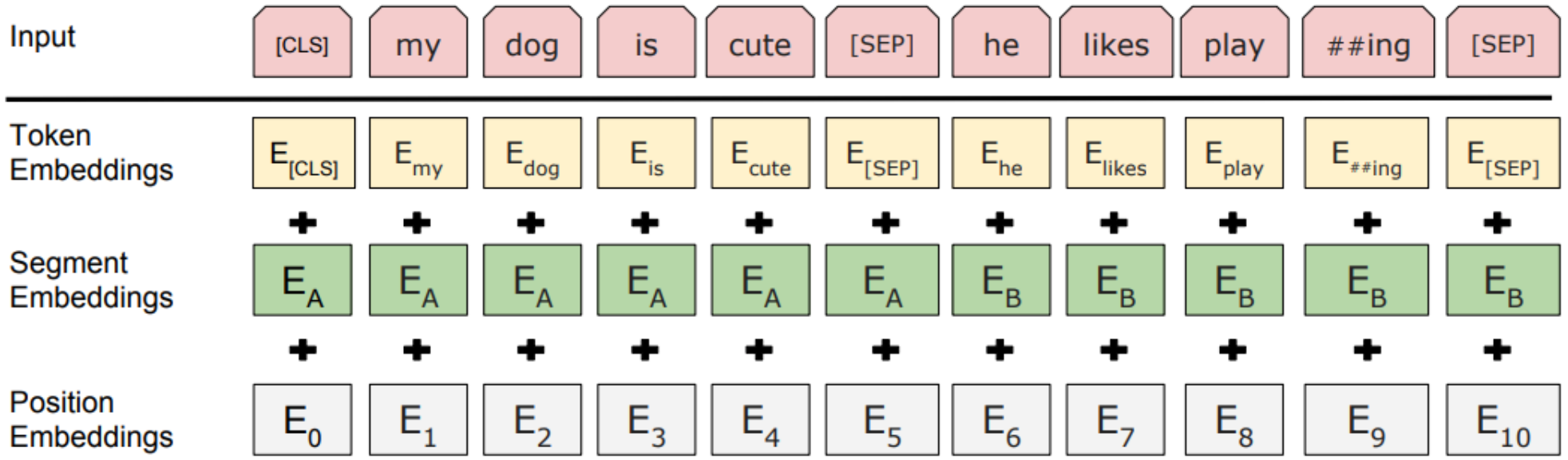
BERT: Pre-Training

- Masked LM (15% token):
  - 80%: replace with [MASK]
  - 10%: replace with a random word
  - 10%: keep unchanged
- Next Sentence Prediction (NSP)
  - Classification Task
  - 2 Labels: IsNext and NotNext
  - Use [SEP] [CLS] token



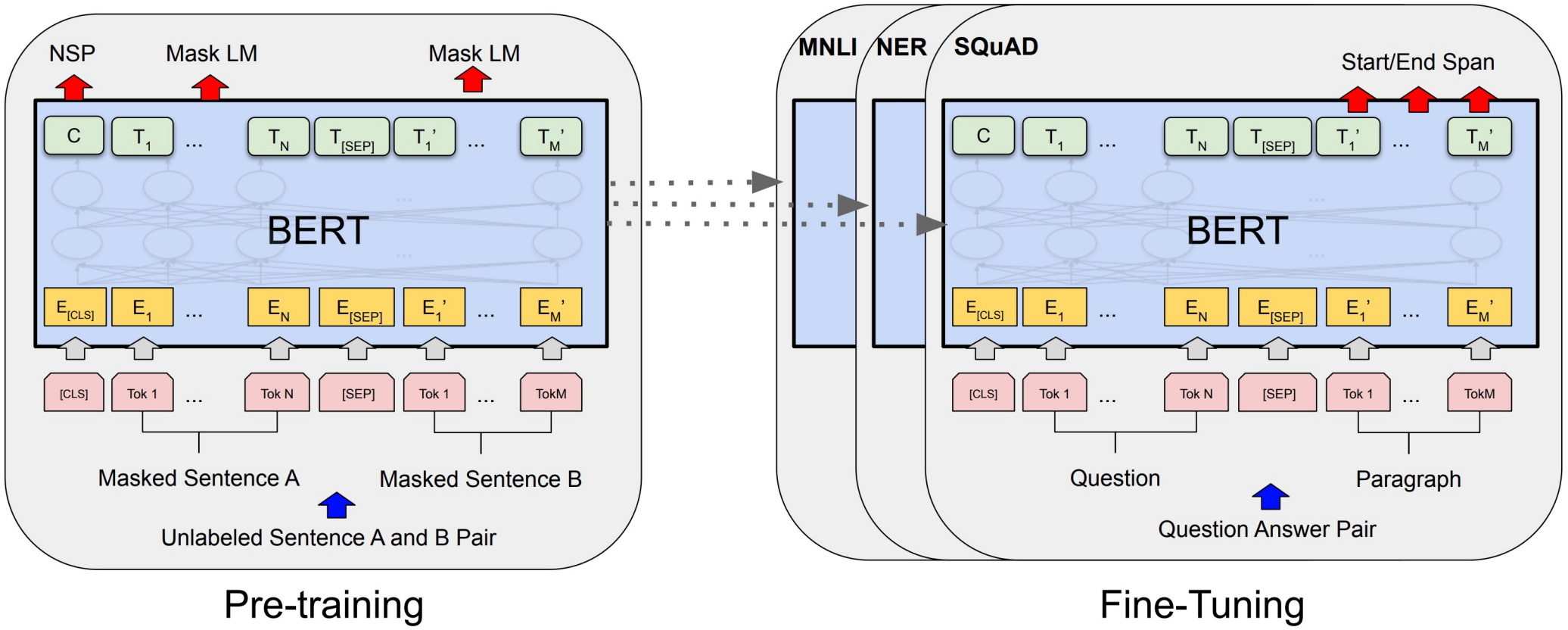
Introduction

BERT: Input Representation



Introduction

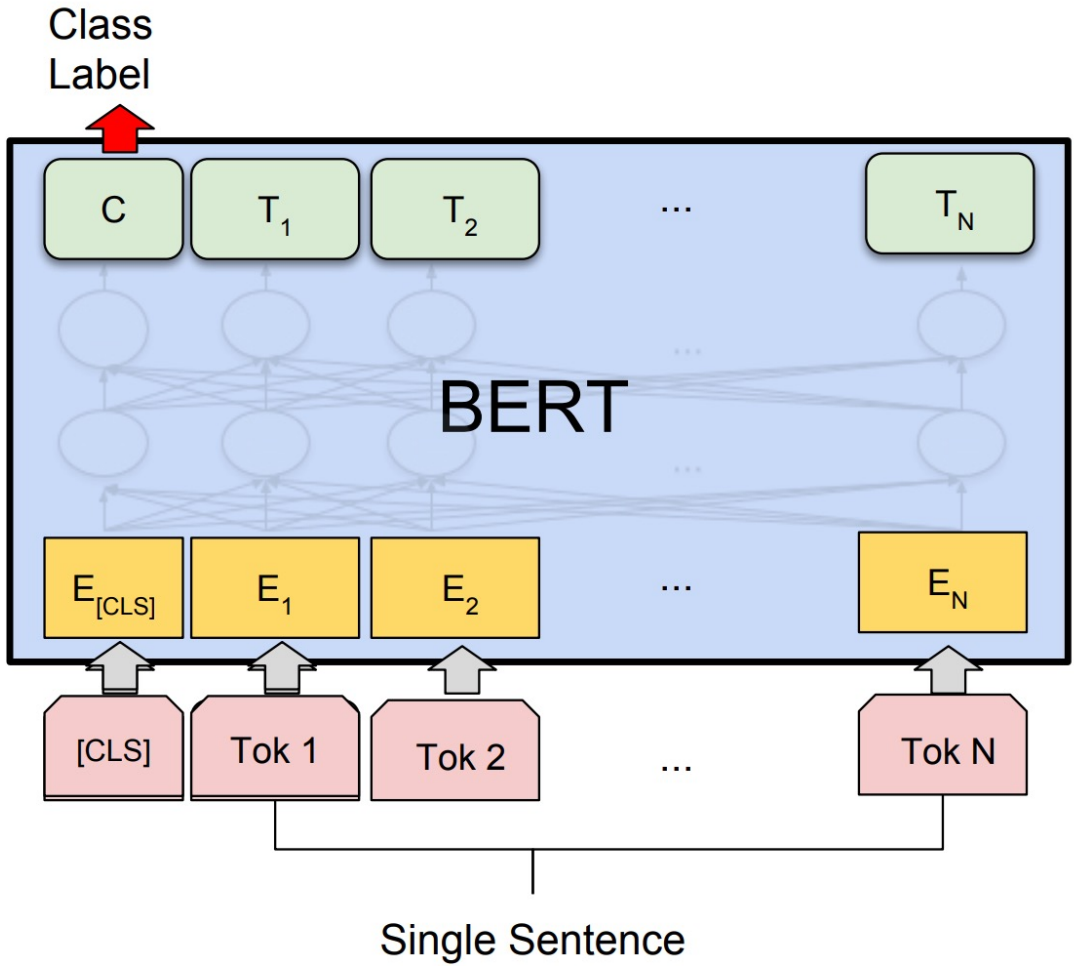
BERT: Fine-Tuning

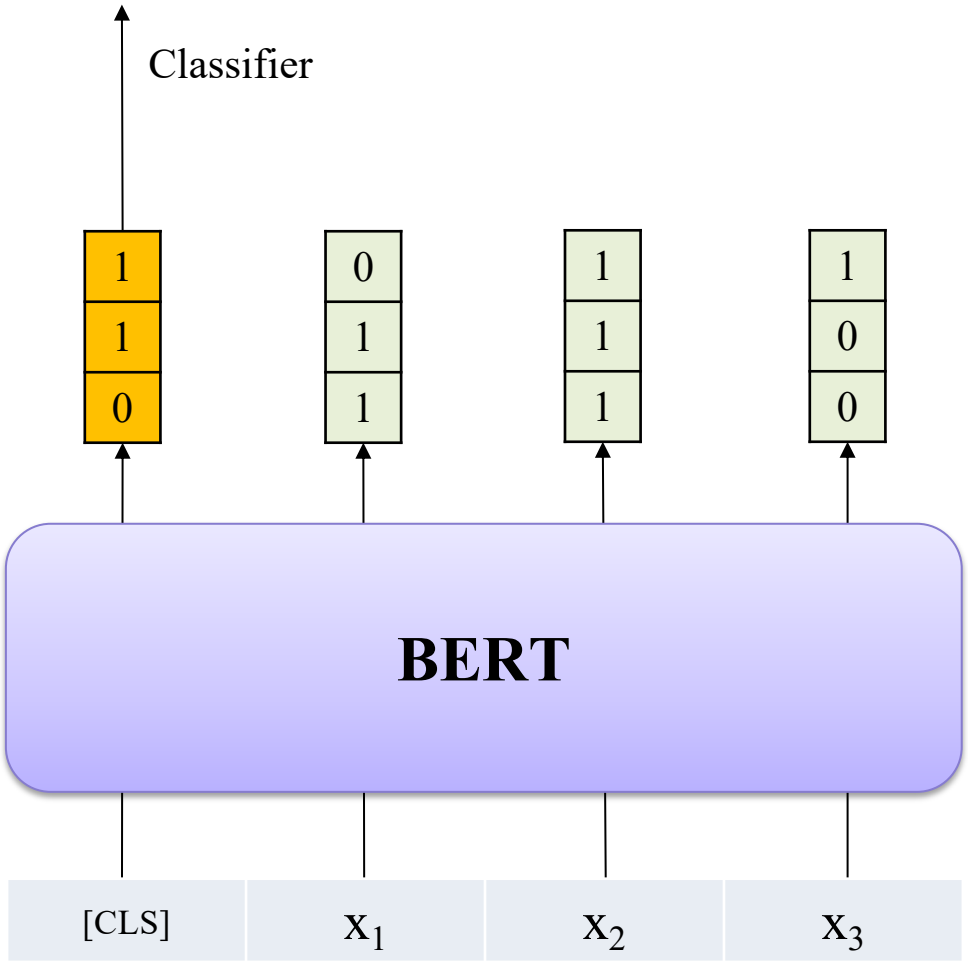


Introduction

BERT: Text Classification

- Document-Level Text Classification

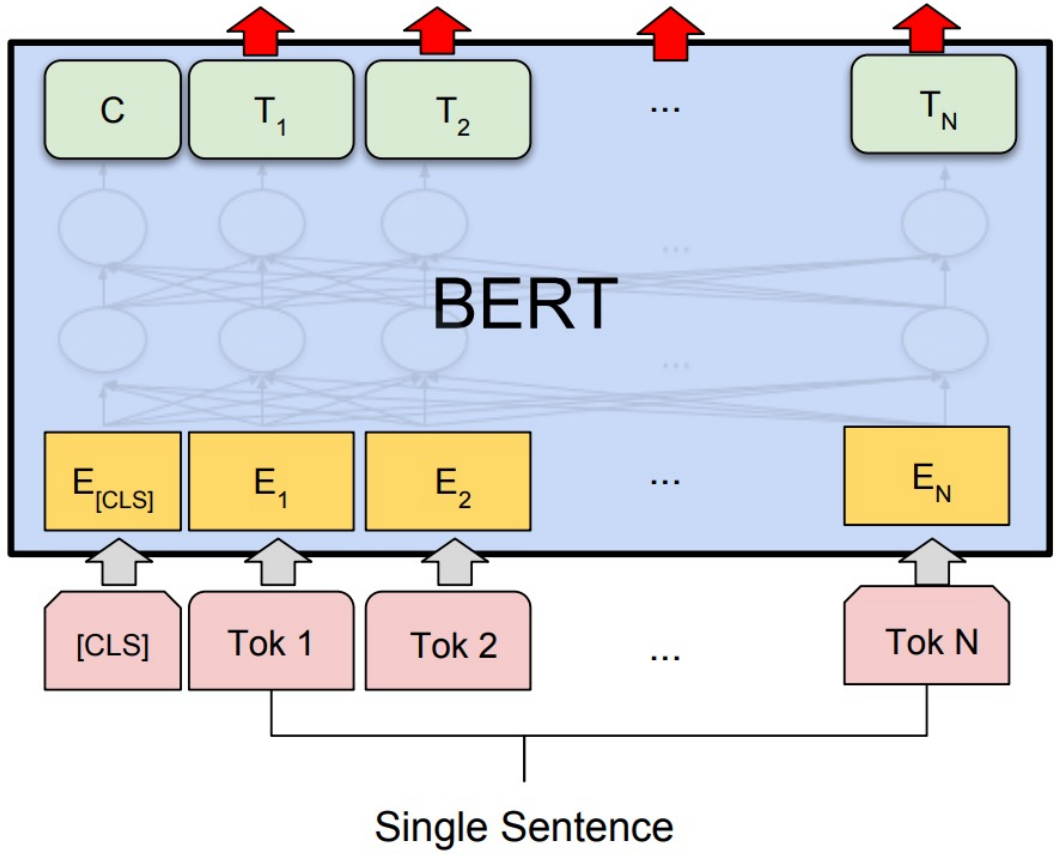


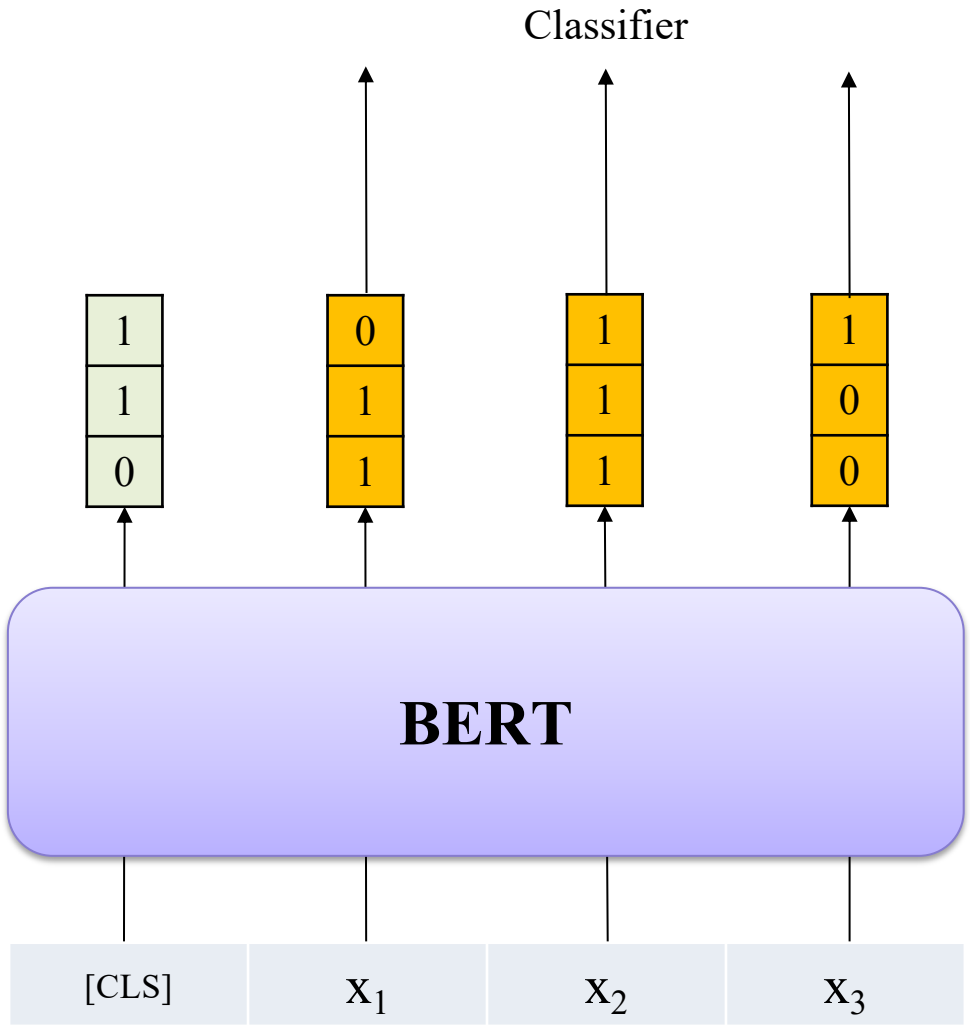


Introduction

BERT: Text Classification

- Token-Level Text Classification





Introduction

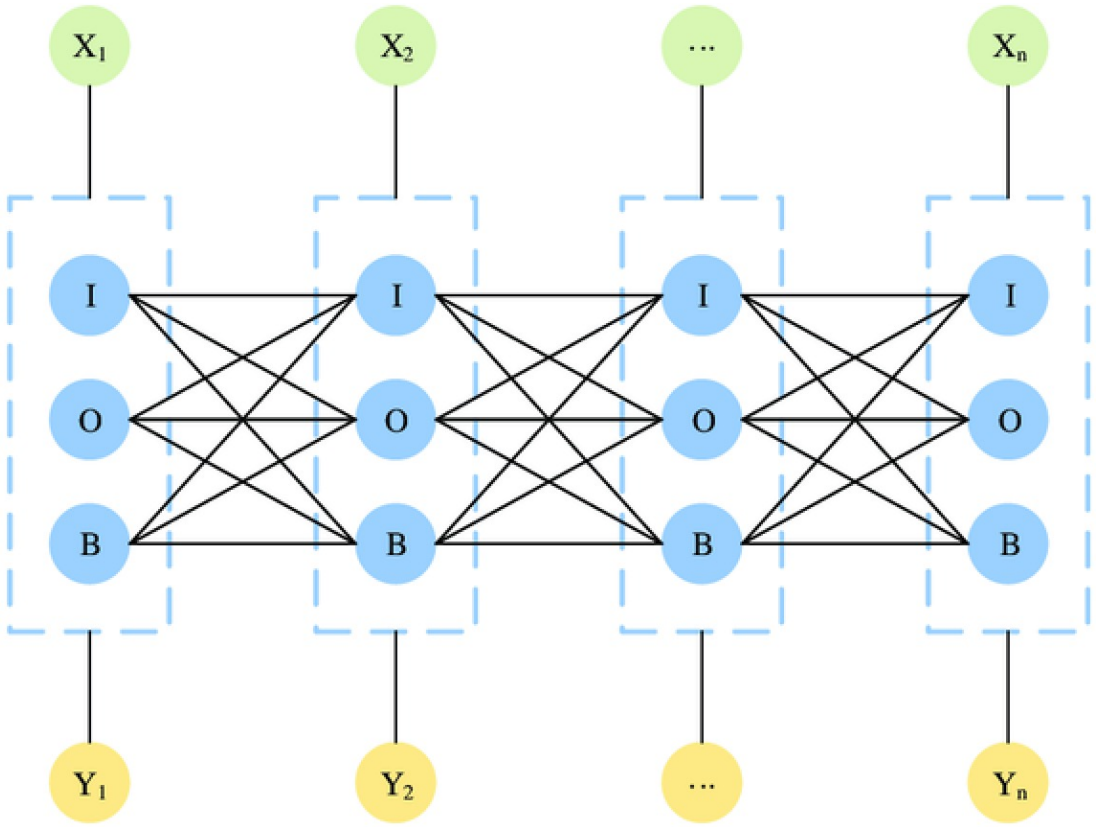
Token-Level Text Classification

- Input:

Sequence n tokens: $\{w_1, w_2, \ldots, w_n\}$

- Ouput:

Sequence n tokens: $\{y_1, y_2, \ldots, y_n\}$



Introduction

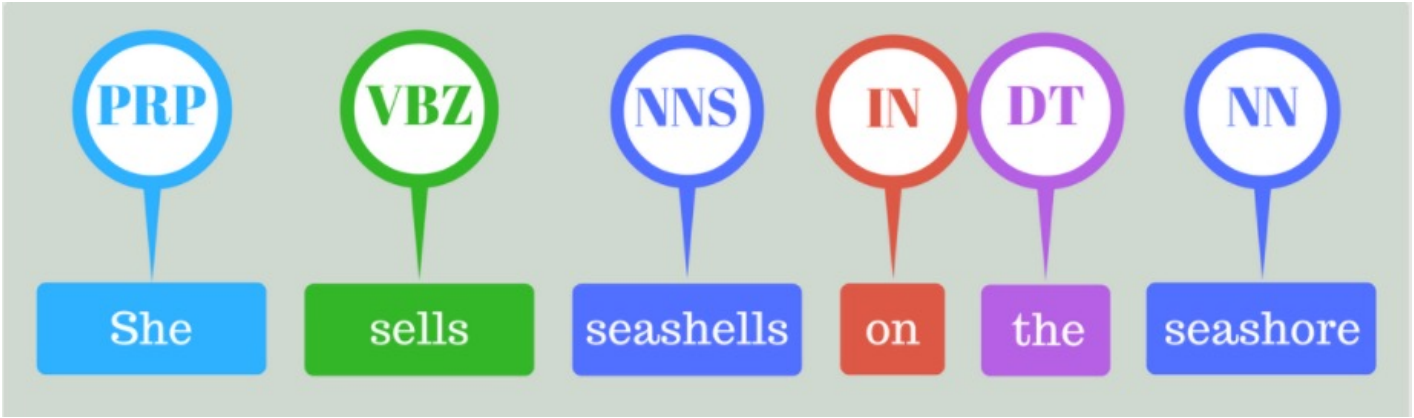
Token-Level Text Classification

- Input:

Sequence n tokens: $\{w_1, w_2, \ldots, w_n\}$

- Ouput:

Sequence n tokens: $\{y_1, y_2, \ldots, y_n\}$



Part-of-Speech (POS) Tagging

Introduction

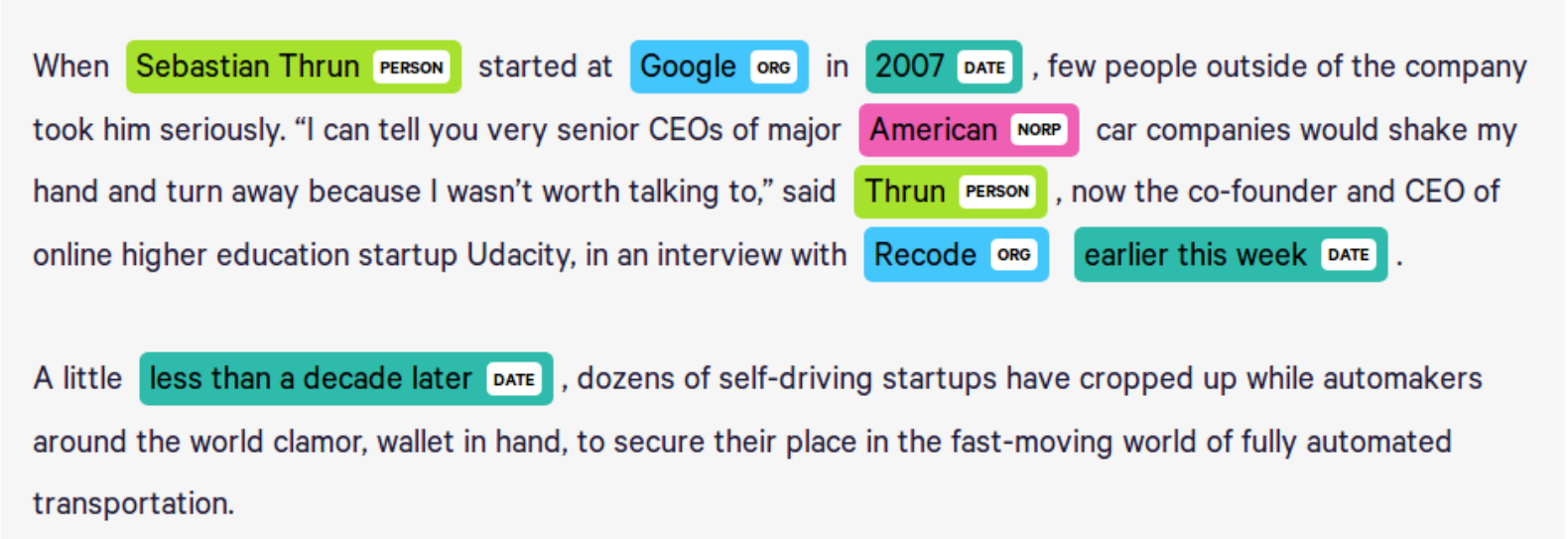
Token-Level Text Classification

- Input:

Sequence n tokens: $\{w_1, w_2, \ldots, w_n\}$

- Ouput:

Sequence n tokens: $\{y_1, y_2, \ldots, y_n\}$



Named Entity Recognition (NER)

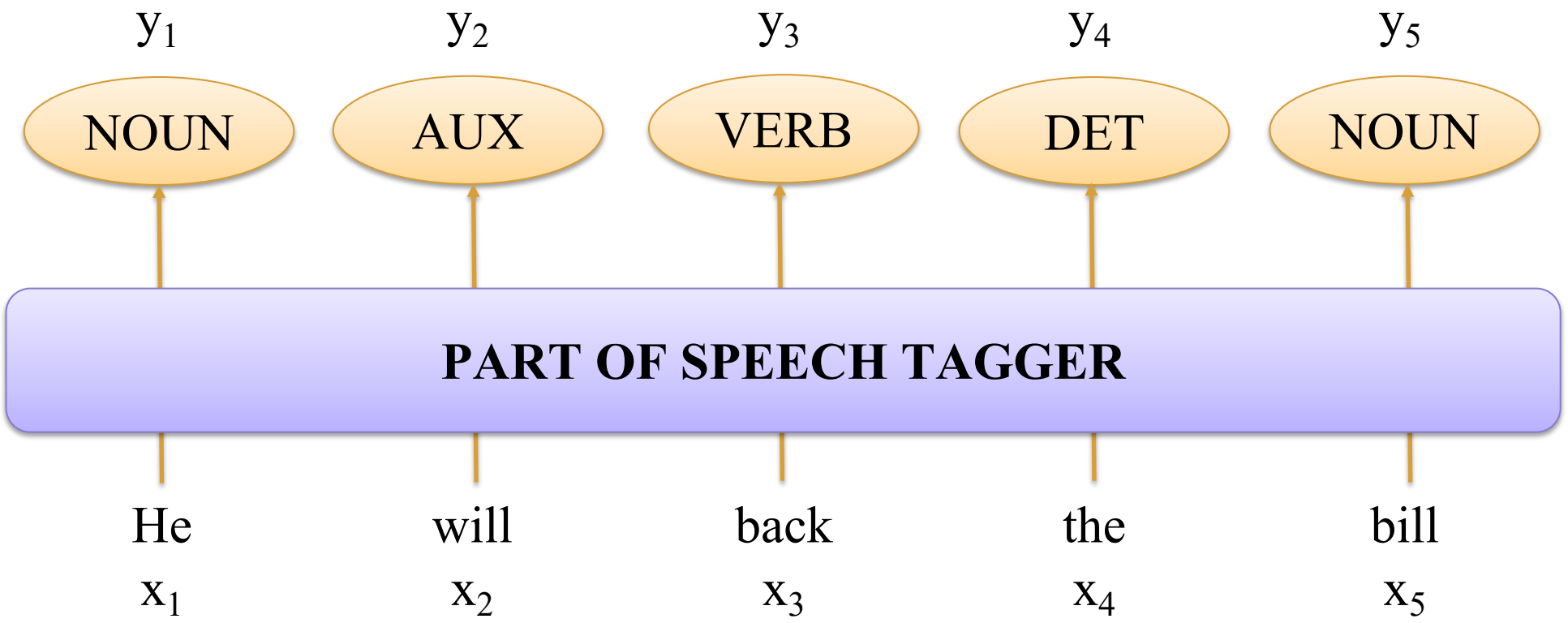
Outline

- Introduction
- Part-of-Speech Tagging (POS)
- Named Entity Recognition (NER)
- Medical Named Entity Recognition

POS Tagging

Token-Level Text Classification

- Assign a POS tag to each token in text



POS Tagging

"Universal Dependencies" Tagset

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red*, *young*, *awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very*, *slowly*, *home*, *yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm*, *cat*, *mango*, *beauty*
	VERB	words for actions and processes	*draw*, *provide*, *go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina*, *IBM*, *Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and*, *or*, *but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that*, *which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

POS Tagging

"Penn Treebank" Tagset

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	"to"	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+,%, &	WRB	wh-adverb	*how, where*

POS Tagging

Penn Tree Bank Dataset

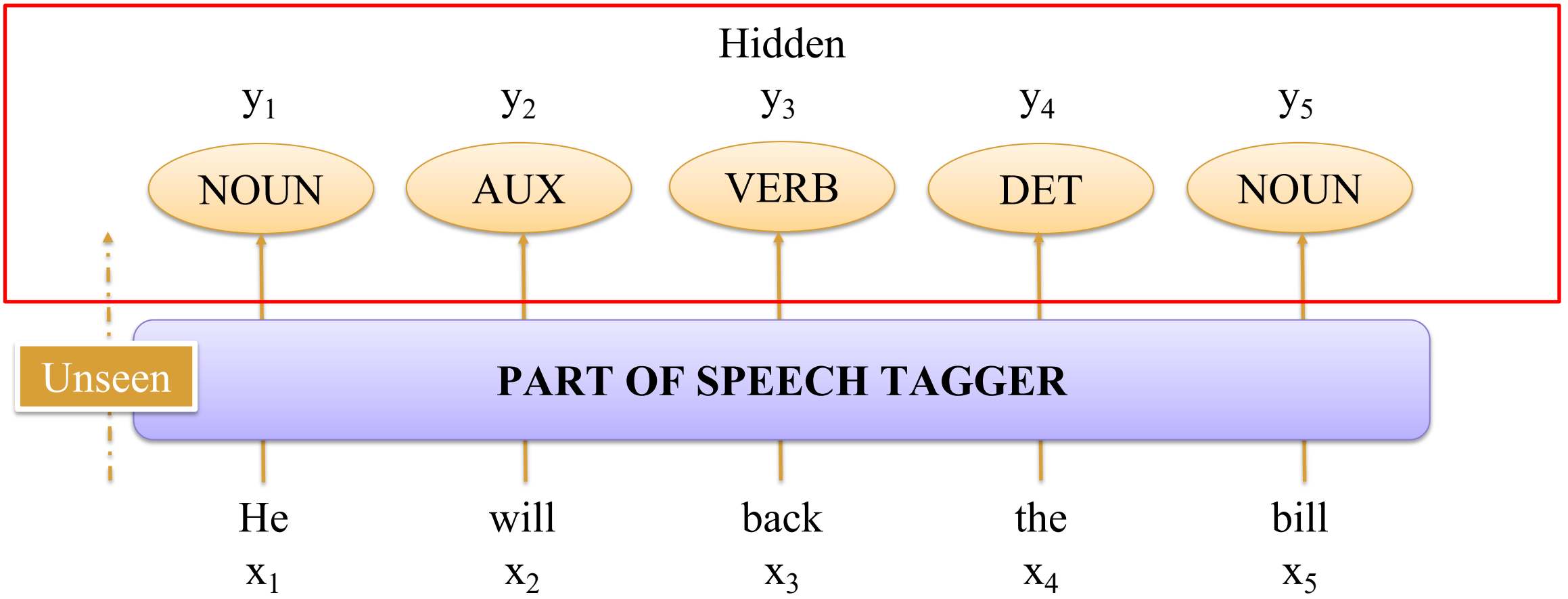
- Samples: 3914

A	lorillard	spokewoman	said	,	this	is	an	old	story	.
DT	NNP	NN	VBD	,	DT	VBZ	DT	JJ	NN	.

There	is	no	asvestos	in	our	products	now	.
EX	VBZ	DT	NN	IN	PRP$	NNS	RB	.

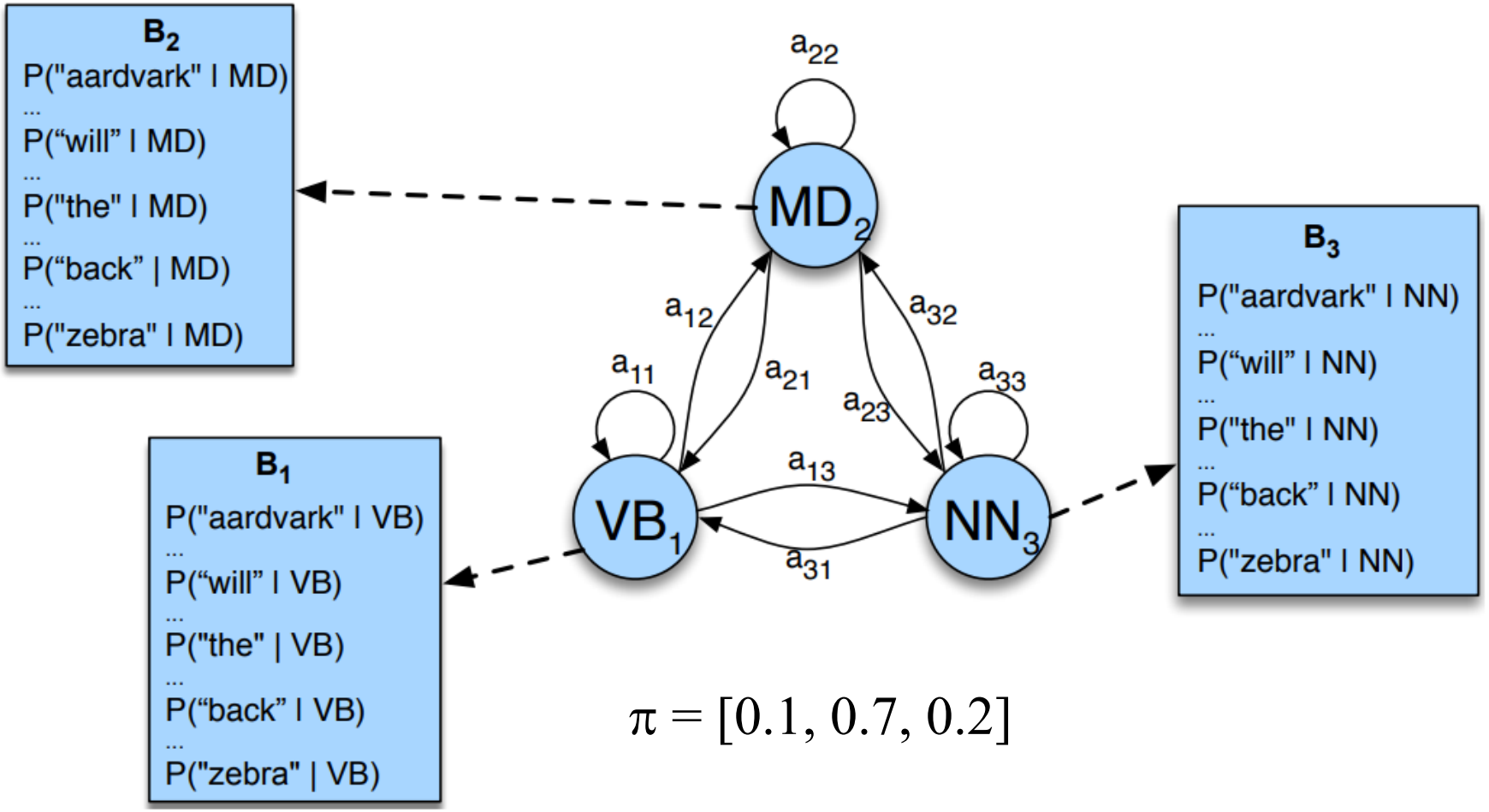
POS Tagging

POS Tagging using Hidden Markov Model



POS Tagging

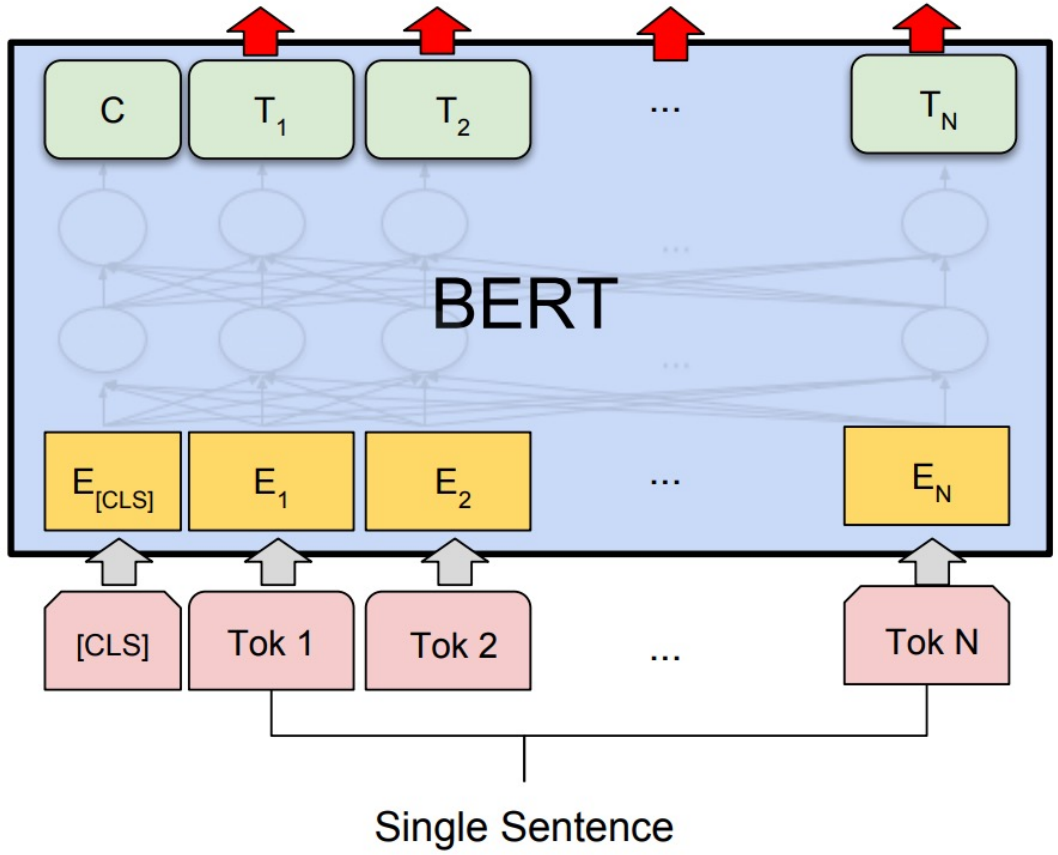
POS Tagging using Hidden Markov Model

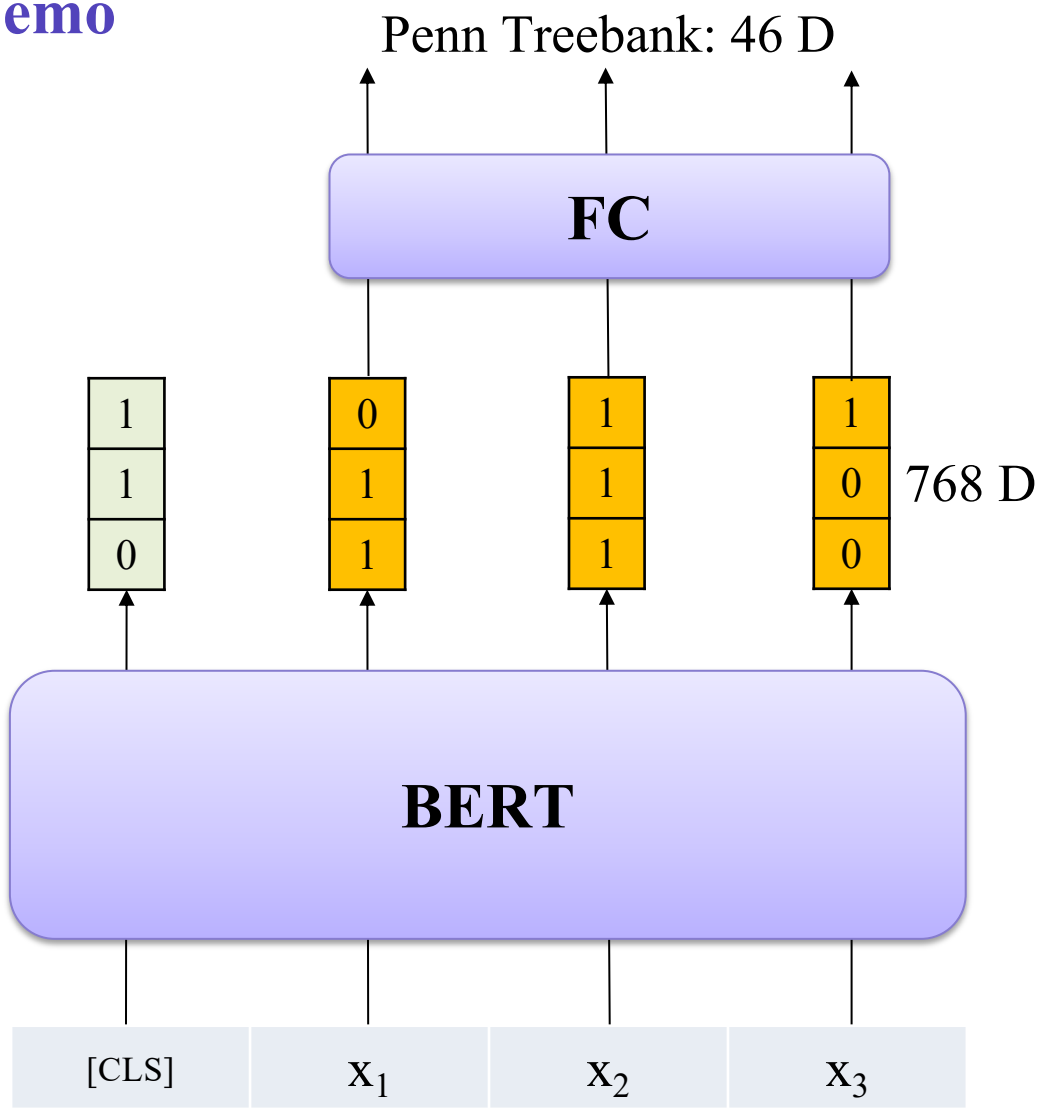


$$\pi = [0.1, 0.7, 0.2]$$

POS Tagging

POS Tagging using Pre-trained Model - Demo





POS Tagging

POS Tagging using Spacy

I go to school.

text	POS	TAG	POS explained
I	PRON	PRP	pronoun
go	VERB	VBP	verb
to	ADP	IN	adposition
school	NOUN	NN	noun
.	PUNCT	.	punctuation

He will back the bill.

text	POS	TAG	POS explained
He	PRON	PRP	pronoun
will	AUX	MD	auxiliary
back	VERB	VB	verb
the	DET	DT	determiner
bill	NOUN	NN	noun
.	PUNCT	.	punctuation

Outline

- Introduction
- Part-of-Speech Tagging (POS)
- Named Entity Recognition (NER)
- Medical Named Entity Recognition

Named Entity Recognition

Example

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Named Entity Recognition

Named Entity Recognition (NER) – Information Extraction

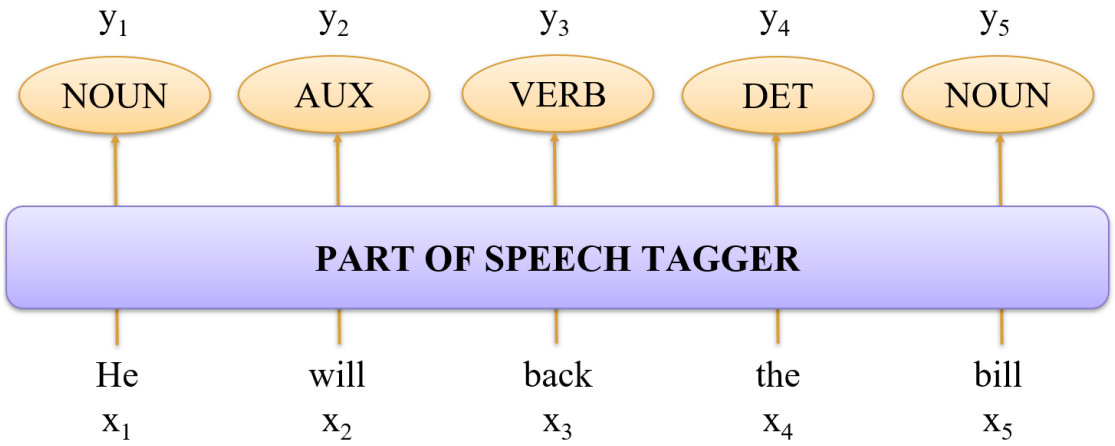
- Mapping span of text into entity tag
- Four entity tags are most common

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	**Mt. Sanitas** is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states	**Palo Alto** is raising the fees for parking.

- Other entity: Date, Time, Price,…

Named Entity Recognition

Span Recognition Problem



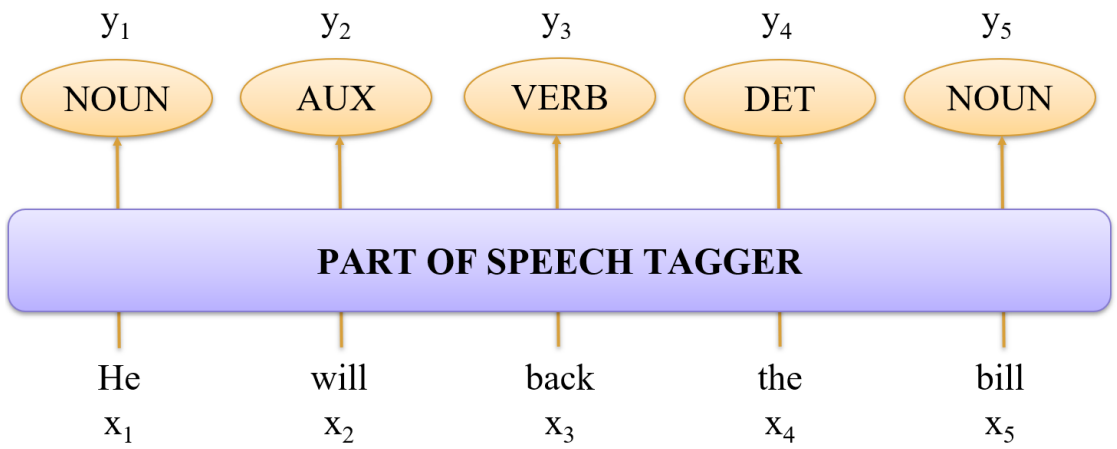
How to segment ?

Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million

[organization] [person] [location] [monetary value]

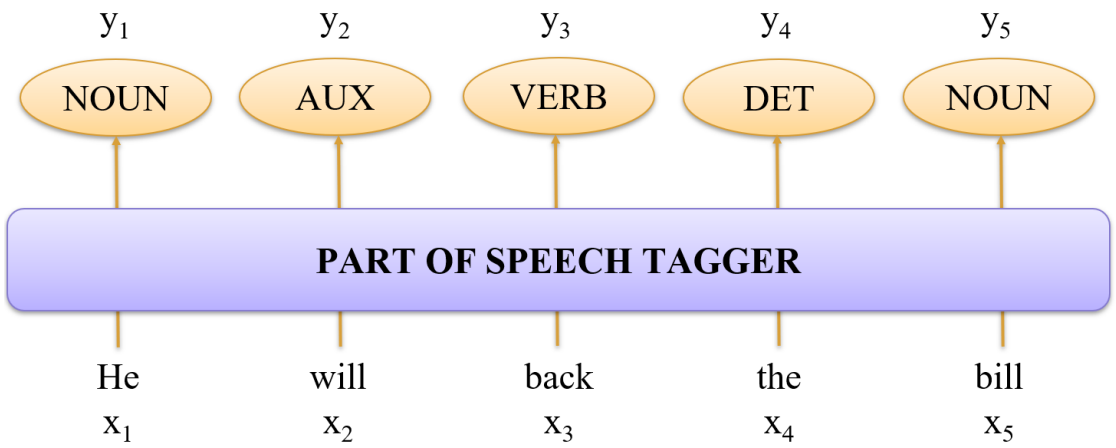
Named Entity Recognition

Span Recognition Problem



Named Entity Recognition

BIO Tagger



Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

Named Entity Recognition

BIO Tagger

B: token that *begins* a span
I: tokens *inside* a span
O: tokens outside of any span

# of tags (where n is #entity types):
1 O tag,
n B tags,
n I tags
=> 2n+1 tags

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

Named Entity Recognition

BIO Tagger Variants: IO, BIOES

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Named Entity Recognition

BIO Tagger Variants: IO, BIOES

I-Pr	I-Pr	O	O	O	I-Or	O	O	O	I-Lo	I-Lo	I-Lo

Sundar Pichai is CEO of Google having headquarter in Mountain View, California

B-Pr	I-Pr	O	O	O	B-Or	O	O	O	B-Lo	I-Lo	I-Lo

Sundar Pichai is CEO of Google having headquarter in Mountain View, California

B-Pr	E-Pr	O	O	O	B-Or	O	O	O	B-Lo	I-Lo	E-Lo

Sundar Pichai is CEO of Google having headquarter in Mountain View, California

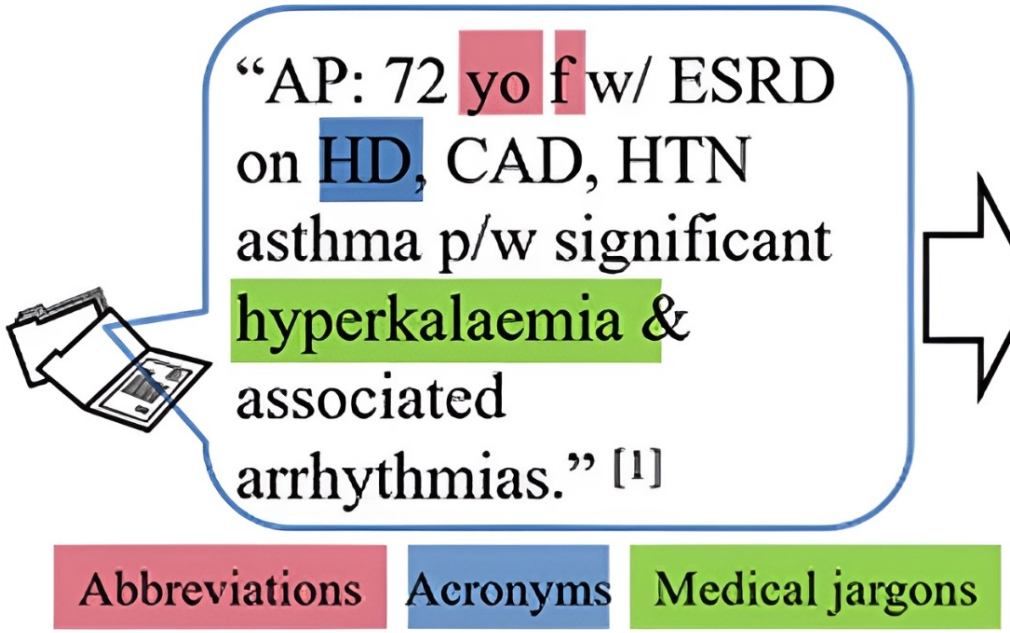
Outline

- Introduction
- Part-of-Speech Tagging (POS)
- Named Entity Recognition (NER)
- Medical Named Entity Recognition

Medical NER

NER Application for Medical Information Extraction

Unstructured data



Meaningful information

Attributes	Medical text
Age	72 years old (yo)
Gender	female (f)
Condition	hypertensive disease (HD)
Symptom	hyperkalemia

Medical NER

MACCROBAT2018

15939911.ann
15939911.txt
16778410.ann
16778410.txt
17803823.ann
17803823.txt
18236639.ann
18236639.txt

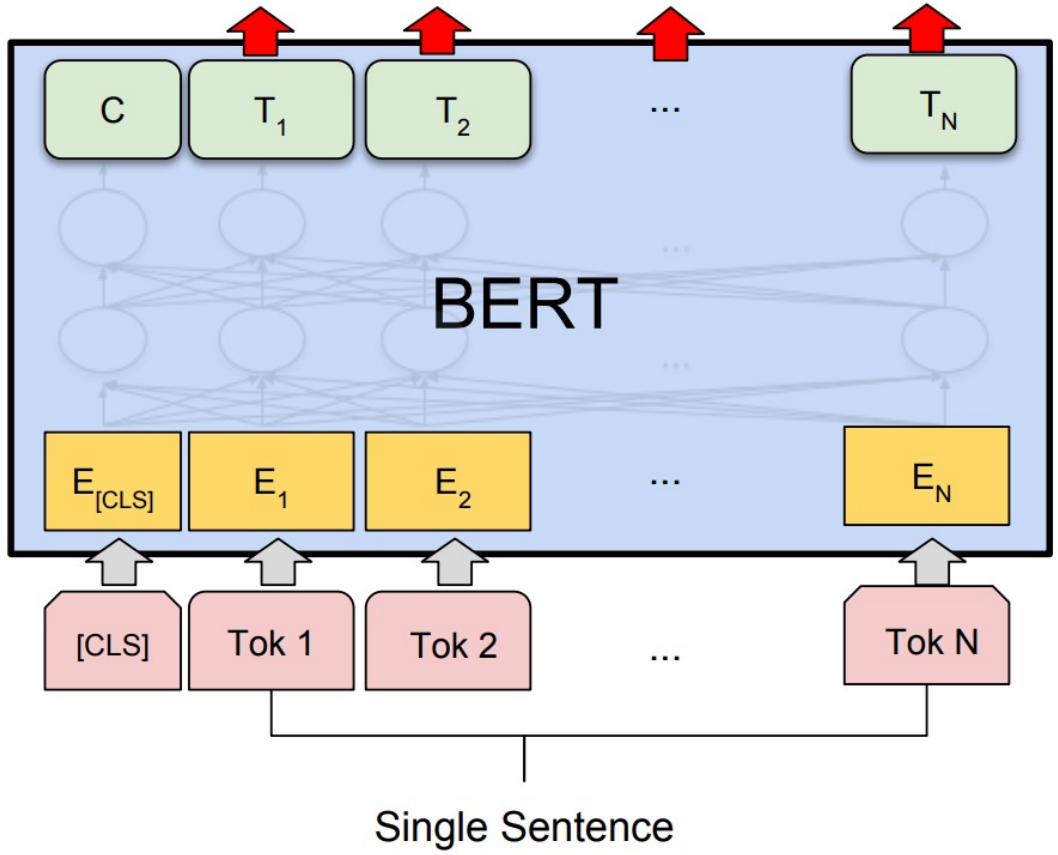
```
Sony formed a joint venture with Ericsson, a mobile phone
company based in Sweden.
Sony announced today that ...
```

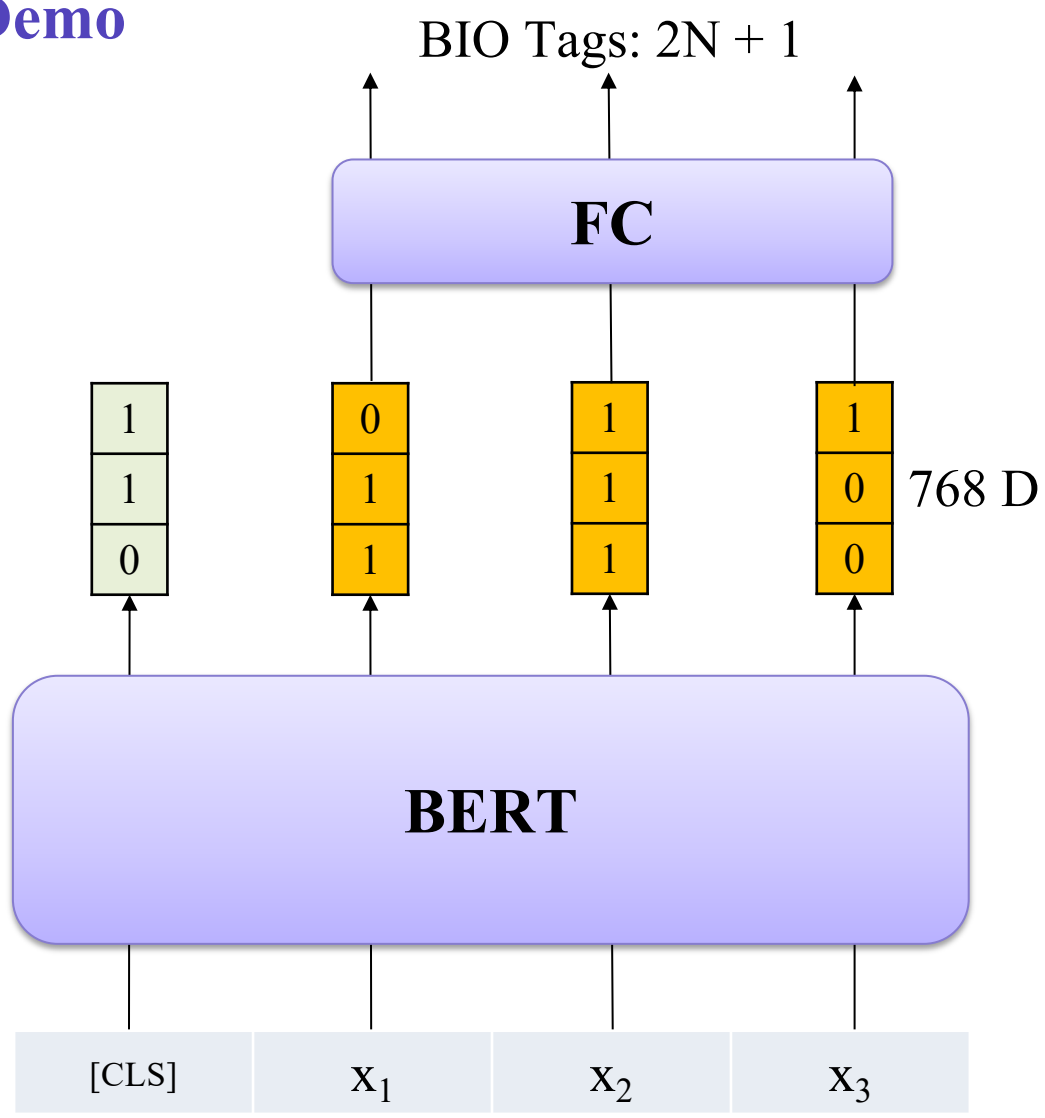
```
T1          Organization 0 4                    Sony
T2          MERGE-ORG 14 27                     joint venture
T3          Organization 33 41                  Ericsson
E1          MERGE-ORG:T2 Org1:T1 Org2:T3
T4          Country 75 81                       Sweden
R1          Origin Arg1:T3 Arg2:T4
```

Standoff Format

Medical NER

Medical NER using Pre-trained Model - Demo





Thanks!

Any questions?