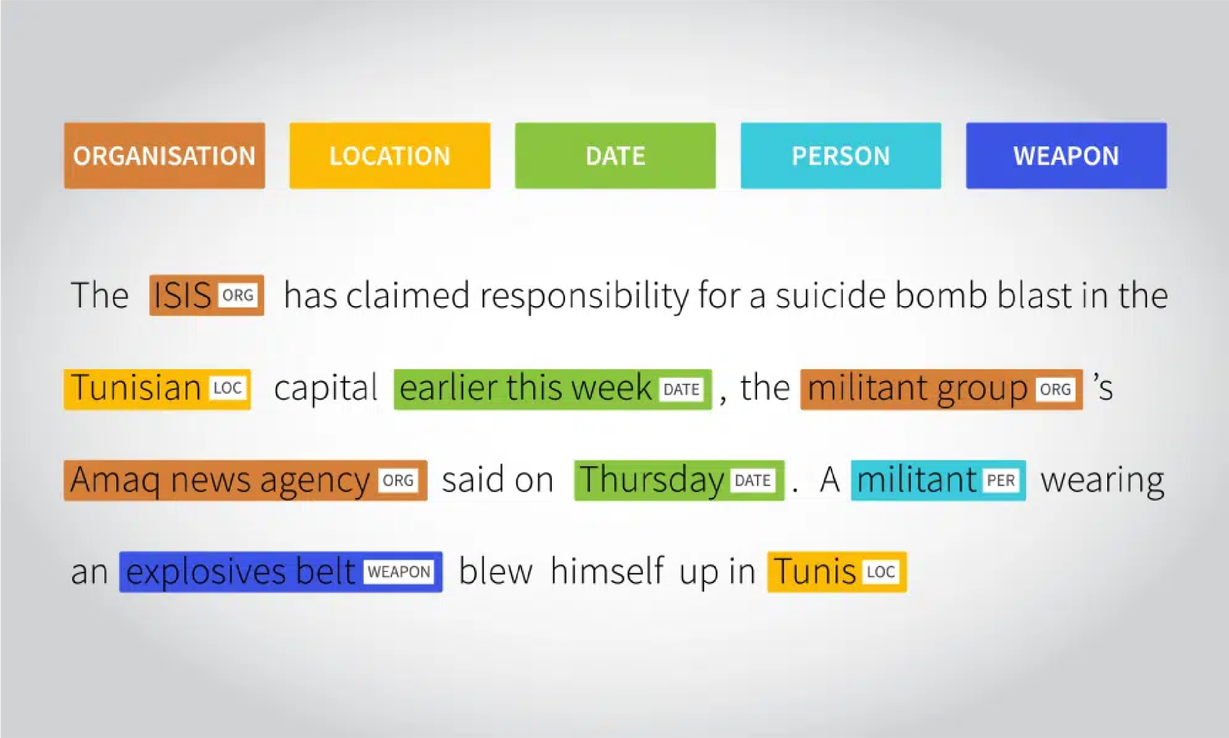
FINE-TUNING BERT MODEL FOR NAMED ENTITY RECOGNITION

NATURAL LANGUAGE PROCESSING

WHAT IS NAMED ENTITY RECOGNITION ?



- Named Entity Recognition (NER) is a task of Natural Language Processing (NLP) that involves **identifying** and **classifying** named entities in a text into predefined categories such as person names, organizations, locations, and others.
- The goal of NER is to **extract structured information from unstructured text data** and **represent it in a machine-readable format**.

OVERVIEW

01

THE PROBLEM

02

THE SOLUTION

03

THE RESULT

04

THE FUTURE WORK

THE PROBLEM

By addressing these challenges, NER can become a more powerful tool in extracting and analyzing information from natural texts, opening up numerous application possibilities across different fields

01

DATA REQUIREMENTS

Requires a large amount of annotated data to achieve high performance.

02

MODEL COMPLEXITY

The large size and high complexity of BERT models demand significant computational resources.

03

DOMAIN SPECIFICITY

Performance can vary significantly across different domains, necessitating domain-specific fine-tuning for optimal results.

THE PROBLEM

By addressing these challenges, NER can become a more powerful tool in extracting and analyzing information from natural texts, opening up numerous application possibilities across different fields

04

AMBIGUITY RESOLUTION

Struggles to distinguish between similar entities.

05

DOMAIN ADAPTATION

Requires significant resources across various fields.

06

LANGUAGE DEPENDENCY

Effectiveness varies by language.

THE PROBLEM

By addressing these challenges, NER can become a more powerful tool in extracting and analyzing information from natural texts, opening up numerous application possibilities across different fields

07

UNSTRUCTURED DATA PROCESSING

Demands advanced techniques

08

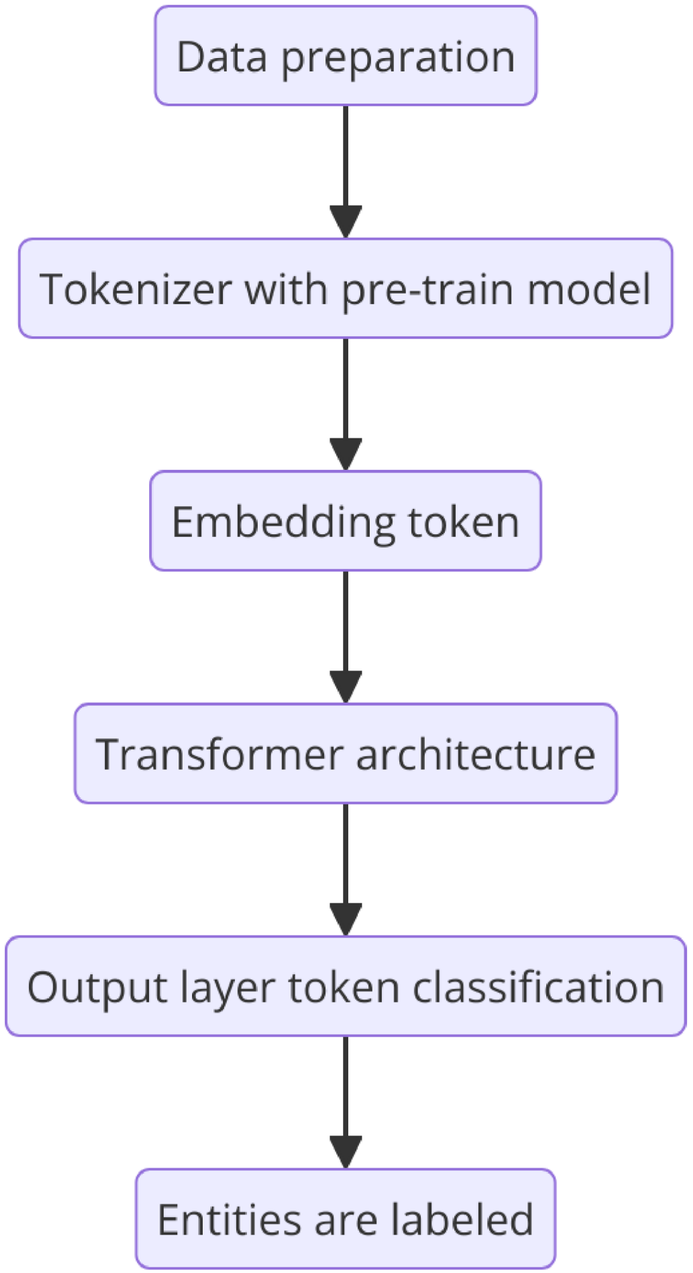
PERFORMANCE MEASUREMENT

Accurate evaluation is complex

09

REAL-TIME PROCESSING

Balancing speed with accuracy is challenging.


Data preparation
Tokenizer with pre-train model
Embedding token
Transformer architecture
Output layer token classification
Entities are labeled

THE PROCESS

1. DATASET
THE CONLL-2003

- This task targets four types of named entities: **persons, locations, organizations**, and **miscellaneous entities** that do not fit into the other three categories.

DATA EXAMPLE

Attribute	Value
chunk_tags	[11, 12, 12, 21, 13, 11, 11, 21, 13, 11, 12, 13, 11, 21, 22, 11, 12, 17, 11, 21, 17, 11, 12, 12, 21, 22, 22, 13, 11, 0]
id	"0"
ner_tags	[0, 3, 4, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
pos_tags	[12, 22, 22, 38, 15, 22, 28, 38, 15, 16, 21, 35, 24, 35, 37, 16, 21, 15, 24, 41, 15, 16, 21, 21, 20, 37, 40, 35, 21, 7]
tokens	["The", "European", "Commission", "said", "on", "Thursday", "it", "disagreed", "with", "German", "advice", "to", "consumers", "to", "shun", "British", "lamb", "until", "scientists", "determine", "whether", "mad", "cow", "disease", "can", "be", "transmitted", "to", "sheep", "."]

NER TAGS

Label	Value	Explanation
O	0	O stands for "Outside" and indicates no entity.
B-PER	1	B-PER indicates the beginning of a person's name.
I-PER	2	I-PER indicates the continuation of a person's name.
B-ORG	3	B-ORG indicates the beginning of an organization's name.
I-ORG	4	I-ORG indicates the continuation of an organization's name.
B-LOC	5	B-LOC indicates the beginning of a location name.
I-LOC	6	I-LOC indicates the continuation of a location name.
B-MISC	7	B-MISC indicates the beginning of a miscellaneous entity.
I-MISC	8	I-MISC indicates the continuation of a miscellaneous entity.

1. DATASET THE CONLL-2003

- This task targets four types of named entities: **persons, locations, organizations**, and **miscellaneous entities** that do not fit into the other three categories.

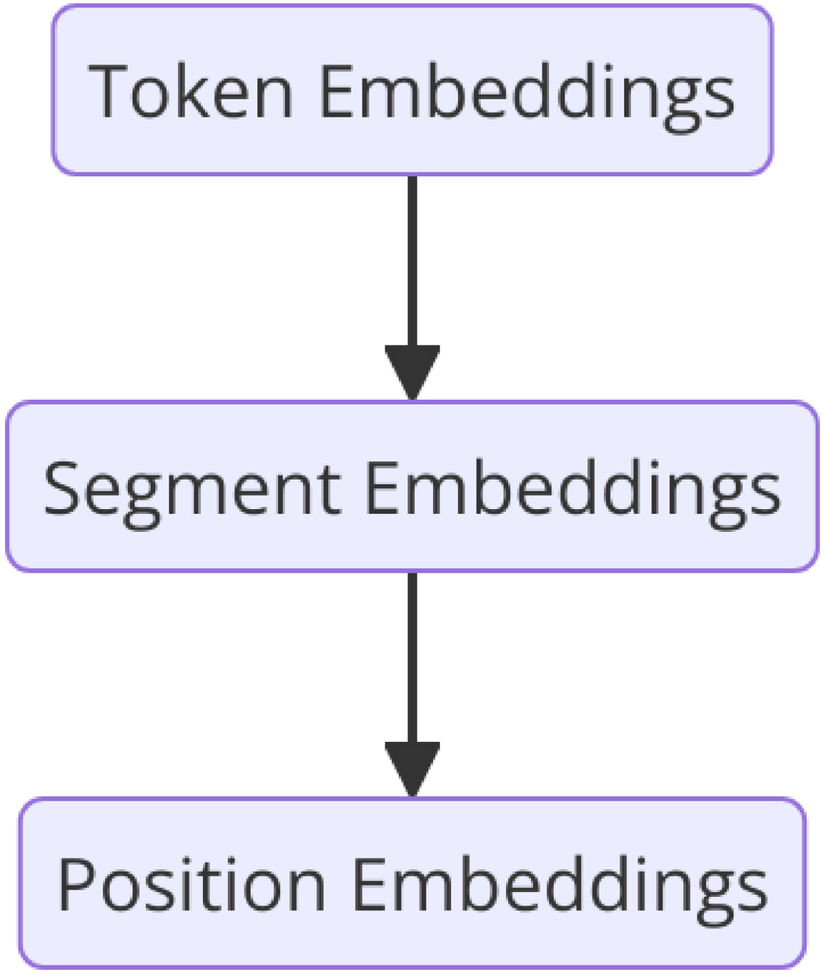
2. TOKENIZER

THE SUBWORD TOKENIZATION OUTPUT

will be convert to id

Input sentence	Barack Obama was born in Honolulu and worked for the United Nations.							
Tokenization	**[CLS]**	**Barack**	**Obama**	**was**	**born**	**in**	**Honolulu**	
Subword Tokenization	**[CLS]**	**Barack**	**Obama**	**was**	**born**	**in**	**Honolu**	**##lu**
Tokenization	**and**	**worked**	**for**	**the**	**United**	**Nations**	**[SEP]**	
Subword Tokenization	**and**	**worked**	**for**	**the**	**United**	**Nations**	**[SEP]**	

3.1 EMBEDDING



TOKEN EMBEDDINGS

Each token in the sentence is converted into a numeric vector through a pre-trained embeddings lookup table.

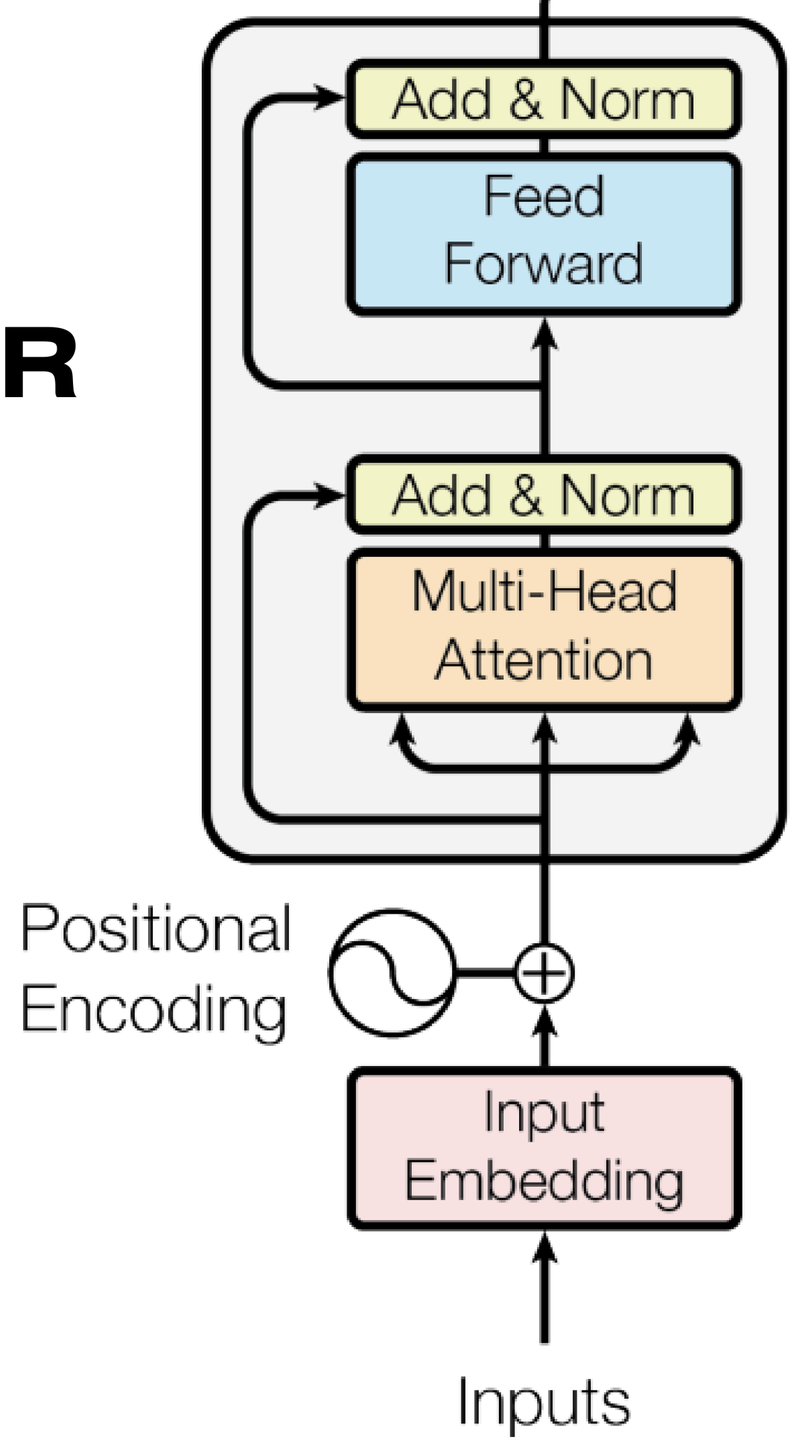
SEGMENT EMBEDDINGS

BERT is designed to process a whole sentence or a pair of sentences.

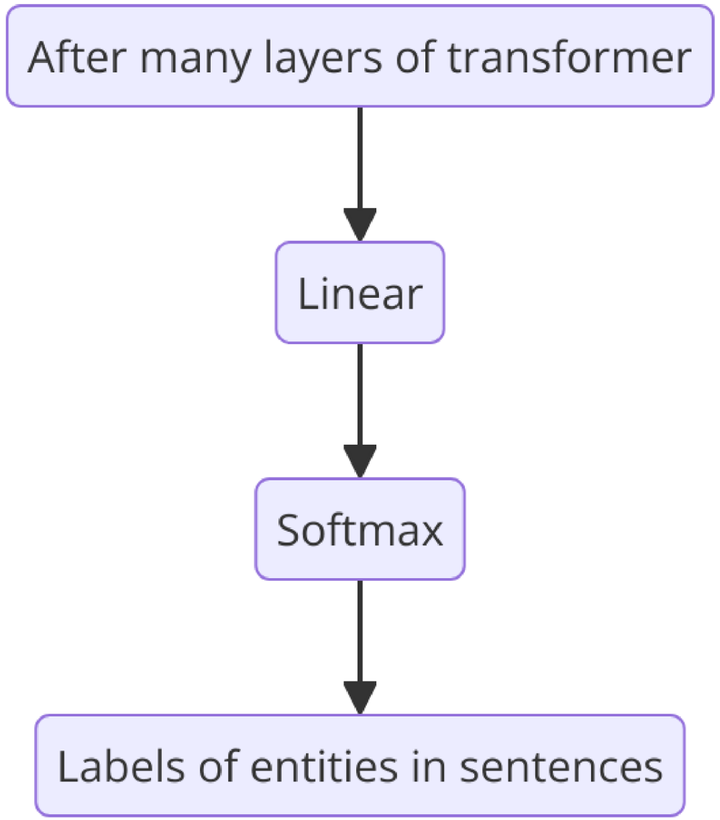
POSITION EMBEDDINGS

BERT needs to understand the order of words in a sentence.

3.2 TRANSFOMER LAYER



BERT



3.3 OUTPUT LAYER

This is the result of example input “ Barack Obama was born in Honolulu and worked for the United Nations.” :

entity	score	index	word	start	end
B-PER	0.9993924	1	barack	0	6
I-PER	0.99744177	2	obama	7	12
B-LOC	0.9985078	6	honolulu	25	33
B-ORG	0.99950993	11	united	53	59
I-ORG	0.9989102	12	nations	60	67

3.4 PREDICTION

EXAMPLE A

Input: In Venezula Vietnam, Canadaa, Apple, apple.
Output:

entity	score	index	word	start	end
B-LOC	0.9979493	2	ve	3	5
B-LOC	0.9983057	3	##ne	5	7
B-LOC	0.99828255	4	##zu	7	9
B-LOC	0.9985959	5	##la	9	11
B-LOC	0.96063536	6	vietnam	12	19
B-LOC	0.9941064	8	canada	21	27
B-LOC	0.77804106	9	##a	27	28
B-LOC	0.9133823	11	apple	30	35
B-LOC	0.55110127	13	apple	37	42

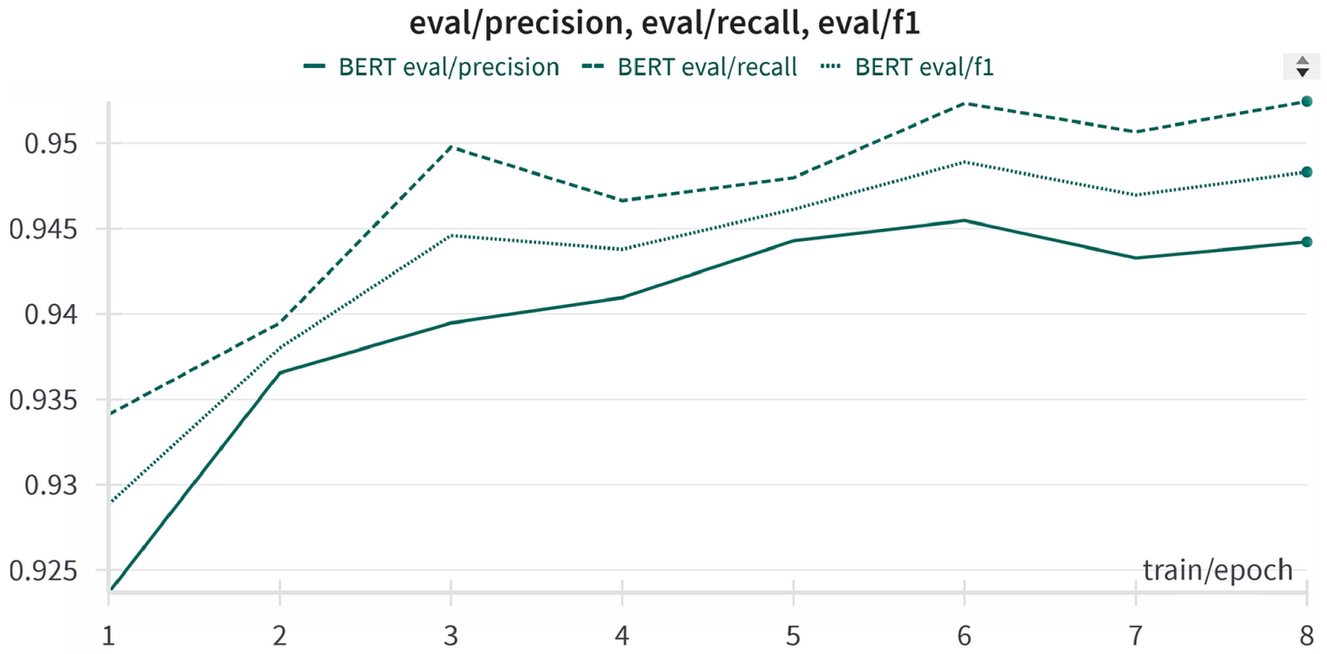
EXAMPLE B

Input: In Venezula Vietnam, Canadaa, Apple, apple. Apple
Output:

entity	score	index	word	start	end
B-LOC	0.9972964	2	ve	3	5
B-LOC	0.9979578	3	##ne	5	7
B-LOC	0.9980026	4	##zu	7	9
B-LOC	0.998351	5	##la	9	11
B-LOC	0.98108023	6	vietnam	12	19
B-LOC	0.9775557	8	canada	21	27
B-ORG	0.89364946	9	##a	27	28
B-ORG	0.90986544	11	apple	30	35
B-ORG	0.98606783	13	apple	37	42
B-ORG	0.9923226	15	apple	44	49

RESULT

- The model's performance **improves** with each epoch, particularly in the early stages.
- Precision, recall, and F1 score all exhibit a **positive trend**, reaching a plateau after the initial epochs.
-> This suggests that the BERT model becomes more **accurate** and **balanced** in recognizing named entities with continued training.



PERFORMANCE COMPARISON

Table 1. Experimental results on the CoNLL-2003 data set

Model	**Precision**	**Recall**	**F1-score**
XLNeT (2020)	90.28	91.20	90.73
BERT-MRC (2019)	92.33	94.61	93.04
BERT (Our)	94.54	95.23	94.88

ACHIEVEMENTS

01

AMBIGUITY AND POLYSEMY

Use BERT to understand context, reducing ambiguity and polysemy.

02

BOUNDARY DETECTION FOR ENTITIES

The tokenize_and_align_labels function aligns labels with tokenized input, handling special tokens like [CLS] and [SEP].

03

ENTITY VARIATION AND SYNONYMS

Tokenizer and model generalize through pre-training on diverse data.

ACHIEVEMENTS

04

LOW-RESOURCE LANGUAGES AND DOMAINS

Use pre-trained BERT to leverage knowledge from large datasets.

05

CONTEXT UNDERSTANDING

BERT captures dependencies and long contexts.

06

OUT-OF-VOCABULARY (OOV) WORDS

Use subword tokenization to handle OOV words.

07

QUALITY AND CONSISTENCY OF ANNOTATION

Use the well-annotated CoNLL-2003 dataset to train the model.

FUTURE WORK

CONTEXT UNDERSTANDING

NESTED ENTITIES

MULTILINGUAL AND CROSS-LINGUAL NER

THANK YOU