



國立高雄科技大學

National Kaohsiung University of Science and Technology

Tokenization in Transformer :From Text to Tokens

Author: Shih-Shinh Huang

Date: July 11, 2024





Outline

- Introduction
 - About Transformer
 - About Tokenization
- Tokenization Strategy
 - Character Tokenization
 - Word Tokenization
 - Subword Tokenization





Introduction

- About Transformer
 - The Transformer neural network is a novel and widely used architecture in NLP.
 - Text classification
 - Word labeling
 - Question-Answering

A. Vaswani, “Attention Is All You Need,” *NIPS*, 2017.

NLP: natural language processing

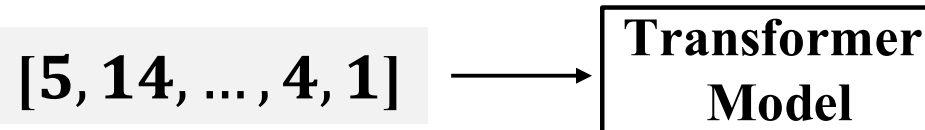


Introduction

- About Transformer
 - Most data in NLP is human-readable text strings

“Tokenization is a core task of NLP!”

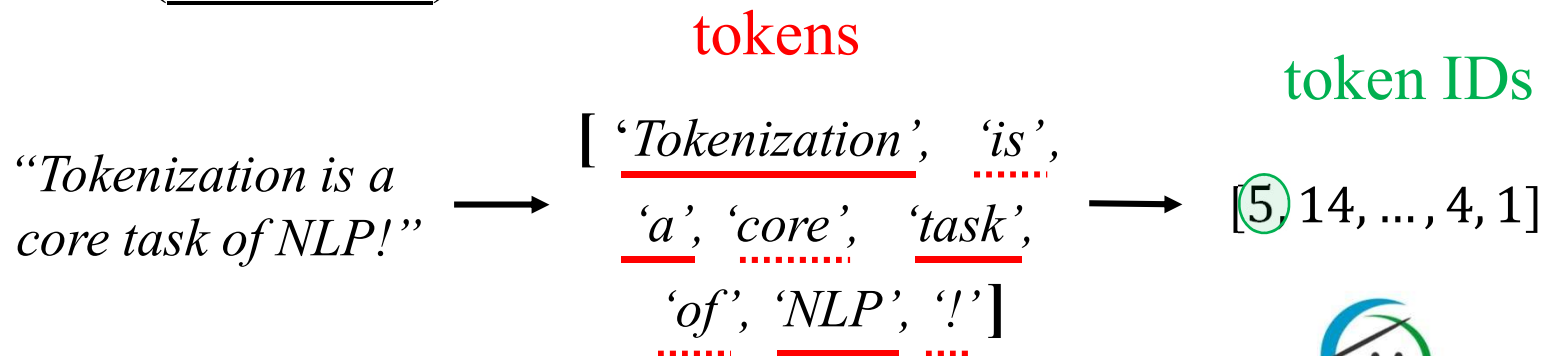
- The transformer models can only receive a list of integer numbers as input.





Introduction

- About Tokenization
 - Tokenization is a step in NLP to convert text strings to a list of integer numbers
 - break down text strings into meaningful units (tokens)
 - encode tokens into transformer-readable integers (token IDs).





Tokenization Strategy

- Character Tokenization
 - split text strings into a list of characters and consider each character as a token.

“Tokenization is a core task of NLP!”



T	o	k	e	n	i	z	a	t	i	o	n		i	s		a		c	o	r	e		t	a	s	k		o	f		N	L	P	!
---	---	---	---	---	---	---	---	---	---	---	---	--	---	---	--	---	--	---	---	---	---	--	---	---	---	---	--	---	---	--	---	---	---	---

token list





Tokenization Strategy

- Character Tokenization
 - transform the tokens into a list of integers

character-based vocabulary

a → 1	A → 27	0 → 53	 → 63
b → 2	B → 28	1 → 54	! → 64
c → 3	C → 29	2 → 55	. → 65
⋮	⋮	⋮	⋮
z → 26	Z → 52	9 → 62	

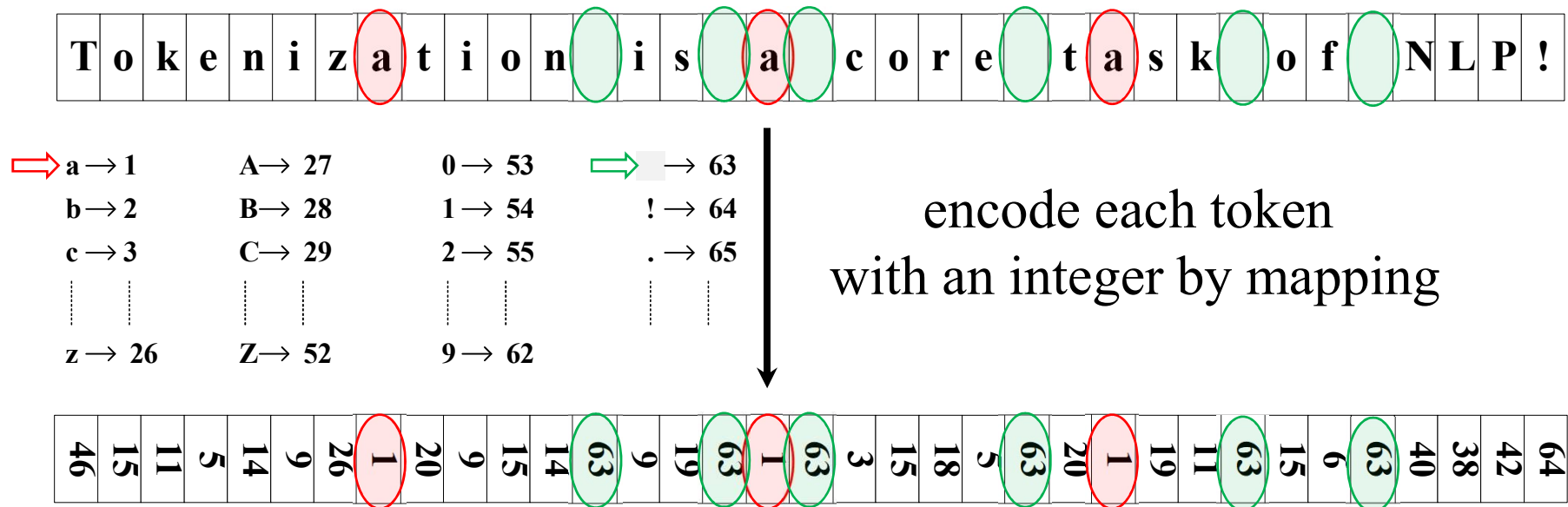
map the character in our vocabulary to a unique integer

vocabulary size of English language is small (≈ 256)



Tokenization Strategy

- Character Tokenization
 - transform the tokens into a list of integers





Tokenization Strategy

- Character Tokenization: Discussion
 - **Advantages:**
 - The vocabulary is easily identified and its size is small.
 - There is no out of vocabulary (OOV) tokens
 - **Drawbacks:** The token sequences is very long and with less meaning and linguistic structures.

T	o	k	e	n	i	z	a	t	i	o	n		i	s		a		c	o	r	e		t	a	s	k		o	f		N	L	P	!
---	---	---	---	---	---	---	---	---	---	---	---	--	---	---	--	---	--	---	---	---	---	--	---	---	---	---	--	---	---	--	---	---	---	---

sequence length = 35





Tokenization Strategy

- Word Tokenization
 - split text strings into a list of words by using whitespace and consider each word as a token.

“Tokenization is a core task of NLP!”



tokenization	is	a	core	task	of	NLP!
--------------	----	---	------	------	----	------

token list





Tokenization Strategy

- Word Tokenization
 - transform the tokens into a list of integers

word-based vocabulary

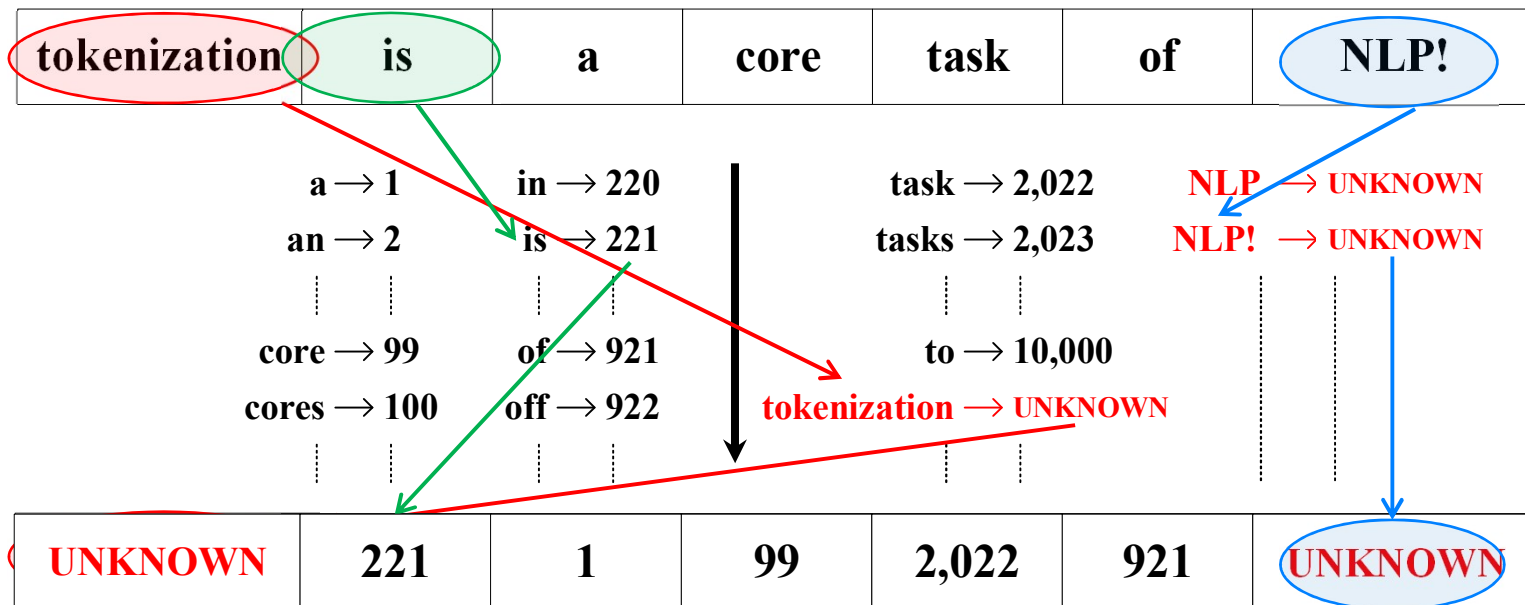
a → 1	in → 220	task → 2,022	NLP → UNKNOWN
an → 2	is → 221	tasks → 2,023	NLP! → UNKNOWN
⋮	⋮	⋮	⋮
core → 99	of → 921	to → 10,000	limit vocabulary size
cores → 100	off → 922	tokenization → UNKNOWN	to 10,000
⋮	⋮	⋮	⋮

map the word in our vocabulary to a unique integer



Tokenization Strategy

- Word Tokenization
 - transform the tokens into a list of integers



encode each token with an integer by mapping



Tokenization Strategy

- Word Tokenization: Discussion
 - **Advantages:** a token as a word in a sentence has a lot of contextual and semantic information
 - **Drawbacks:**
 - the vocabulary size is extremely large
 - some tokens will be UNKNOWN (OOV tokens) that lose information

UNKNOWN	is	a	core	task	of	UNKNOWN
---------	----	---	------	------	----	---------





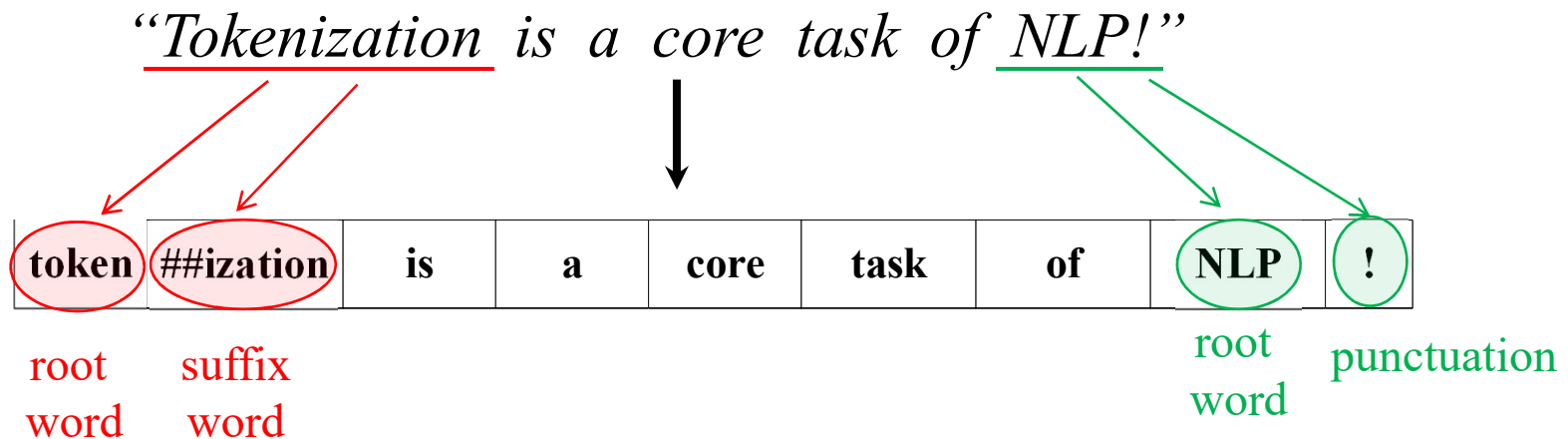
Tokenization Strategy

- Subword Tokenization
 - The main idea is to address the issues faced by character and word tokenizations.
 - split the rarely used words into smaller subwords
 - ⇒ reduce vocabulary size and alleviate OOV problem
 - keep the frequently used words as unique entities
 - ⇒ preserve contextual and semantic information



Tokenization Strategy

- Subword Tokenization



token tokenization real realization organ organization
 ⇒ *token token real real organ organ*
 ##ization ##ization ##ization



Tokenization Strategy

- Subword Tokenization:
 - **WordPiece**: used by BERT and DistilBERT
M. Schuster, *et. al.* “Japanese and Korean Voice Search,”
ICASSP, 2012.
 - **Unigram**: used by XLNet and ALBERT
T. Kudo, “Subword Regularization: Improving Neural Network
Translation Models with Multiple Subword Candidates,” *ACL*, 2018.
 - **Byte-Pair Encoding (BPE)**: used by GPT-2
R. Sennrich, *et. al.* “Neural Machine Translation of Rare Words with
Subword Units,” *ACL*, 2016.



thank you

