Paper Review

Semantics-aware BERT for Language Understanding

Zhang et al., AAAI, Feb 2020

Myeongsup Kim

Integrated M.S./Ph.D. Student
Data Science & Business Analytics Lab.
School of Industrial Management Engineering
Korea University

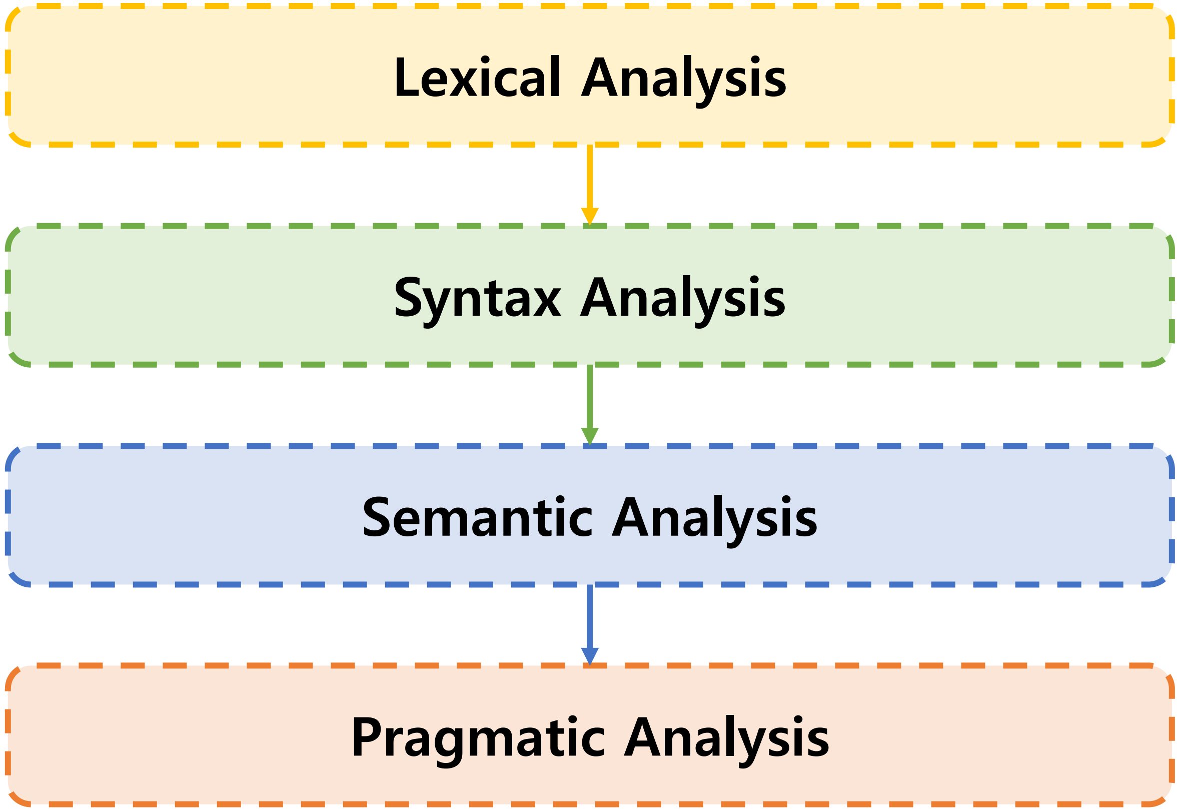
Myeongsup_kim@korea.ac.kr

CONTENTS

1. Prerequisites

- Steps of Natural Language Processing
- BERT: Bidirectional Encoder Representations from Transformers
- Semantic Role Labeling

<Steps of Natural Language Processing>



1 Prerequistes

-BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

<BERT>

양방향적 학습이 가능하도록 Transformer Encoder의 목적함수를 변형한 형태
대형 언어 코퍼스에 대해 비지도 학습으로 Language Model을 구축하고, 지도 학습으로 하위 NLP task
에 적용하는 준지도 학습의 언어 모형

Sen A: I am MS.
Sen B: I study deep learning.

Masked Language Model

Input: I am [Mask1]. I [Mask2] deep learning.
Label: [Mask1] = MS, [Mask2] = study

Next Sentence Prediction

Input: Sen A = I am MS.
Sen B = I study deep learning
Label: IsNext

<Input Representation>

BERT의 Input은 Token Embedding, Positional Embedding, Segment Embedding의 합으로 구성
문장의 시작을 의미하는 [CLS] 토큰과 문장의 구분 및 종료를 의미하는 [SEP] 토큰을 포함하여 Input을 형성

Input: I am MS. I study deep learning.

↓

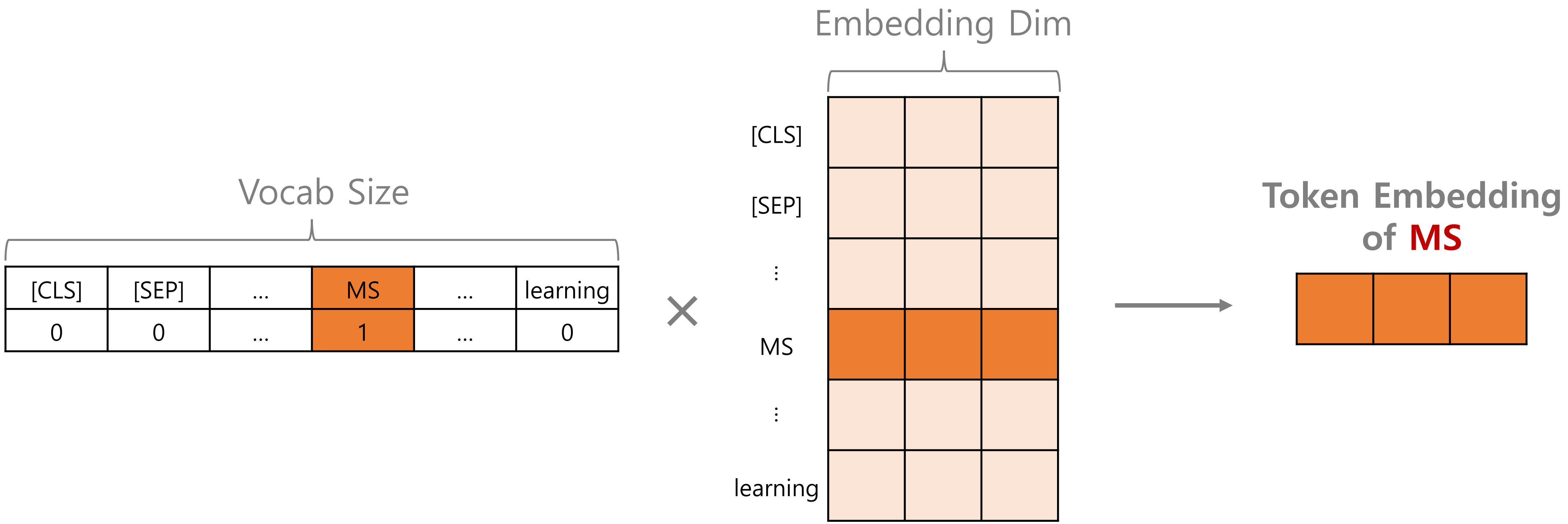
Input: [CLS], I, am, MS, [SEP], I study, deep, learning, [SEP]

↓

Embedding = Token + Segment + Positional Embedding

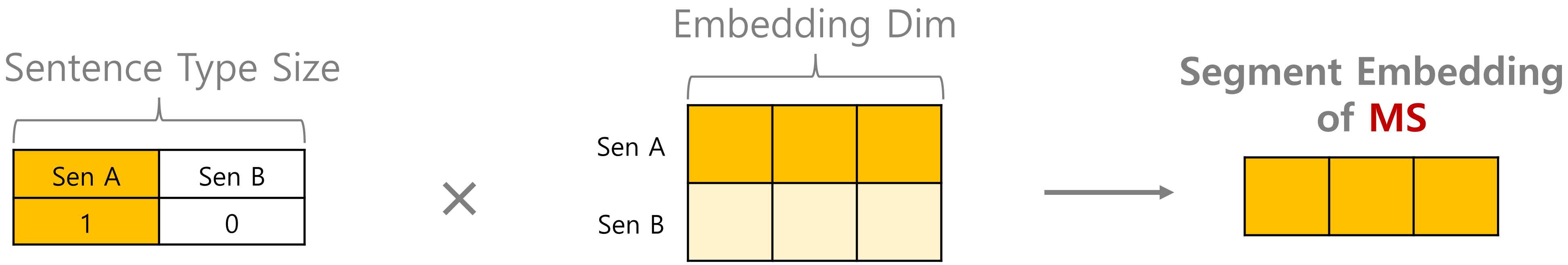
<Token Embedding>

Input: [CLS], I, am, MS, [SEP], I study, deep, learning, [SEP]



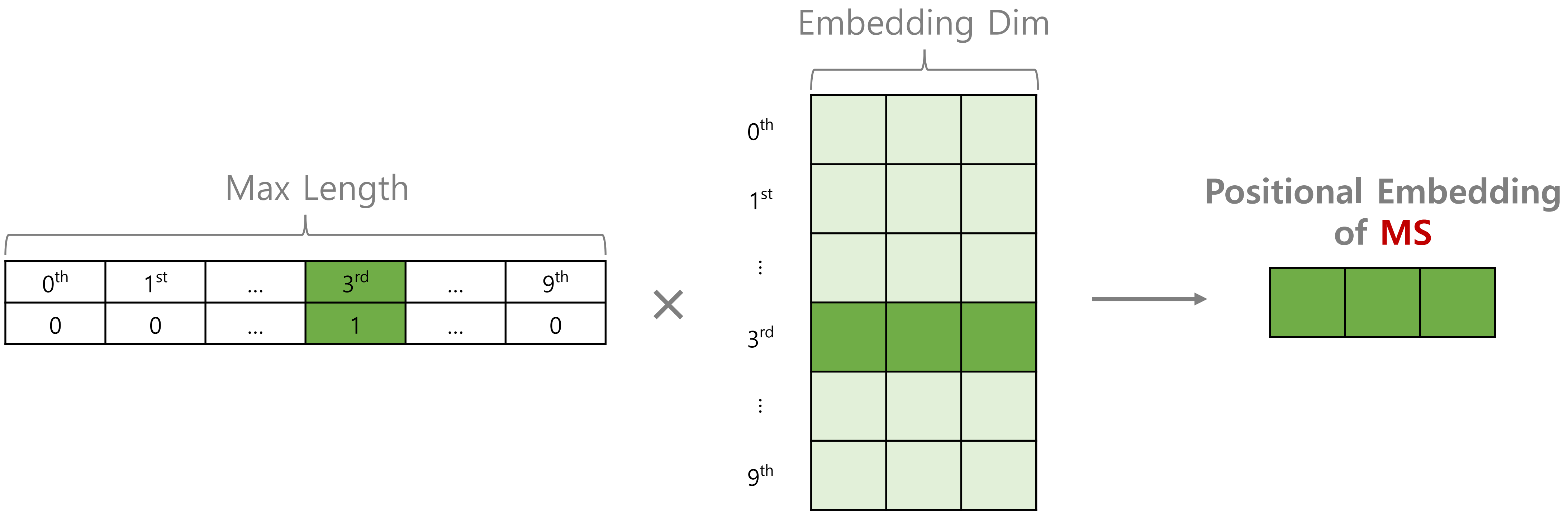
<Segment Embedding>

Input: [CLS], I, am, MS, [SEP], I study, deep, learning, [SEP]



<Positional Embedding>

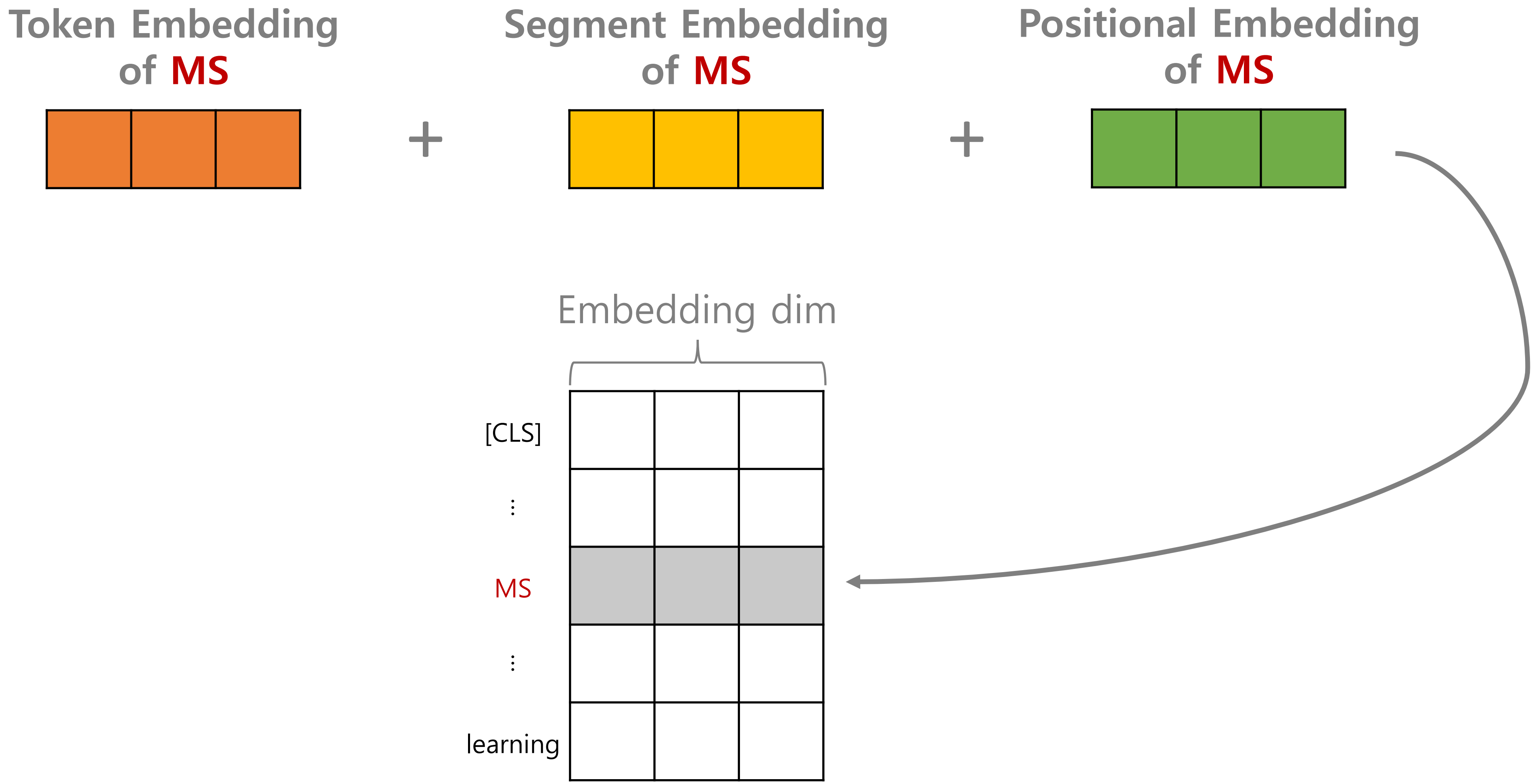
Input: [CLS], I, am, MS, [SEP], I study, deep, learning, [SEP]



1 Prerequistes

-BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

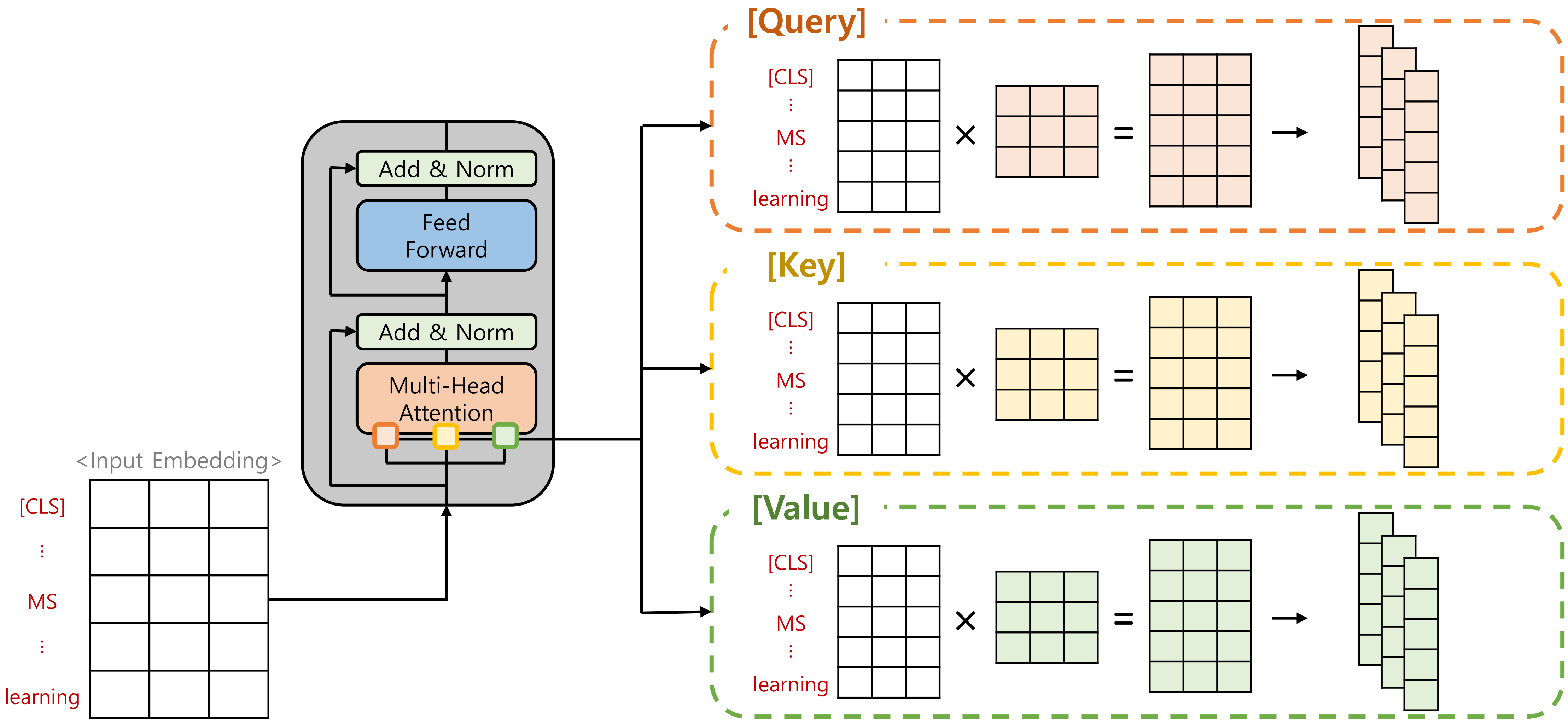
<Input Representation>



1 Prerequistes

-BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

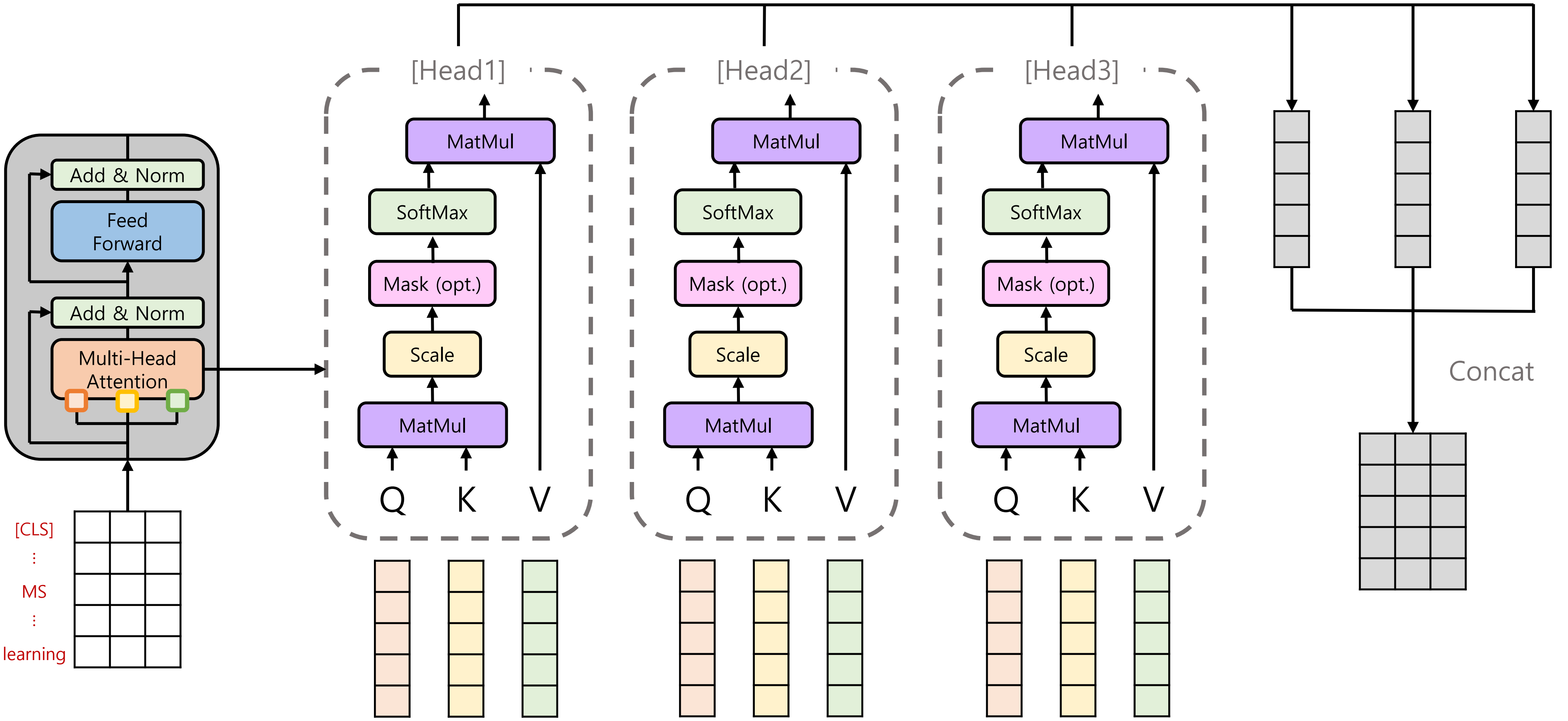
<Encoder>



1 Prerequistes

-BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

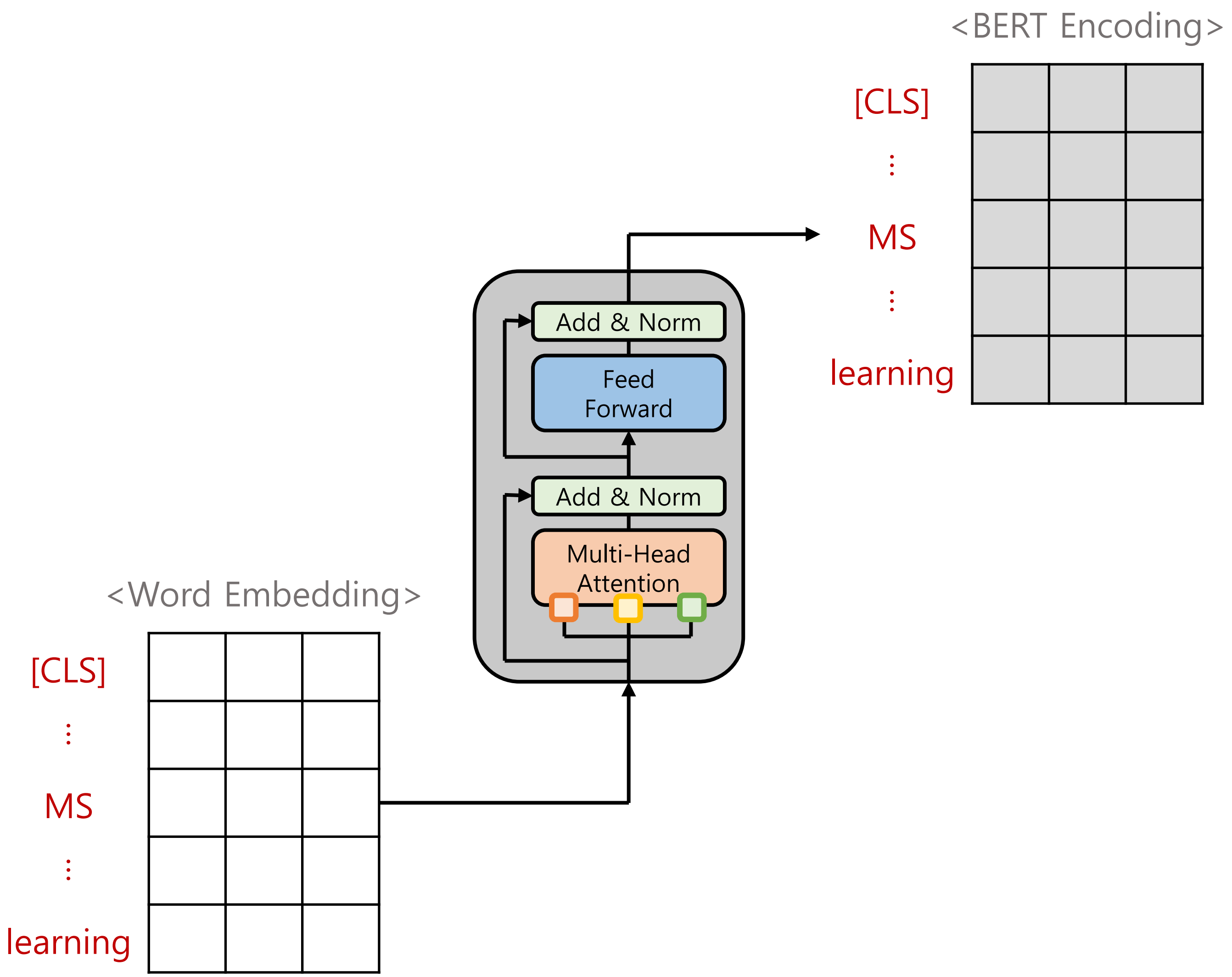
<Encoder>



1 Prerequistes

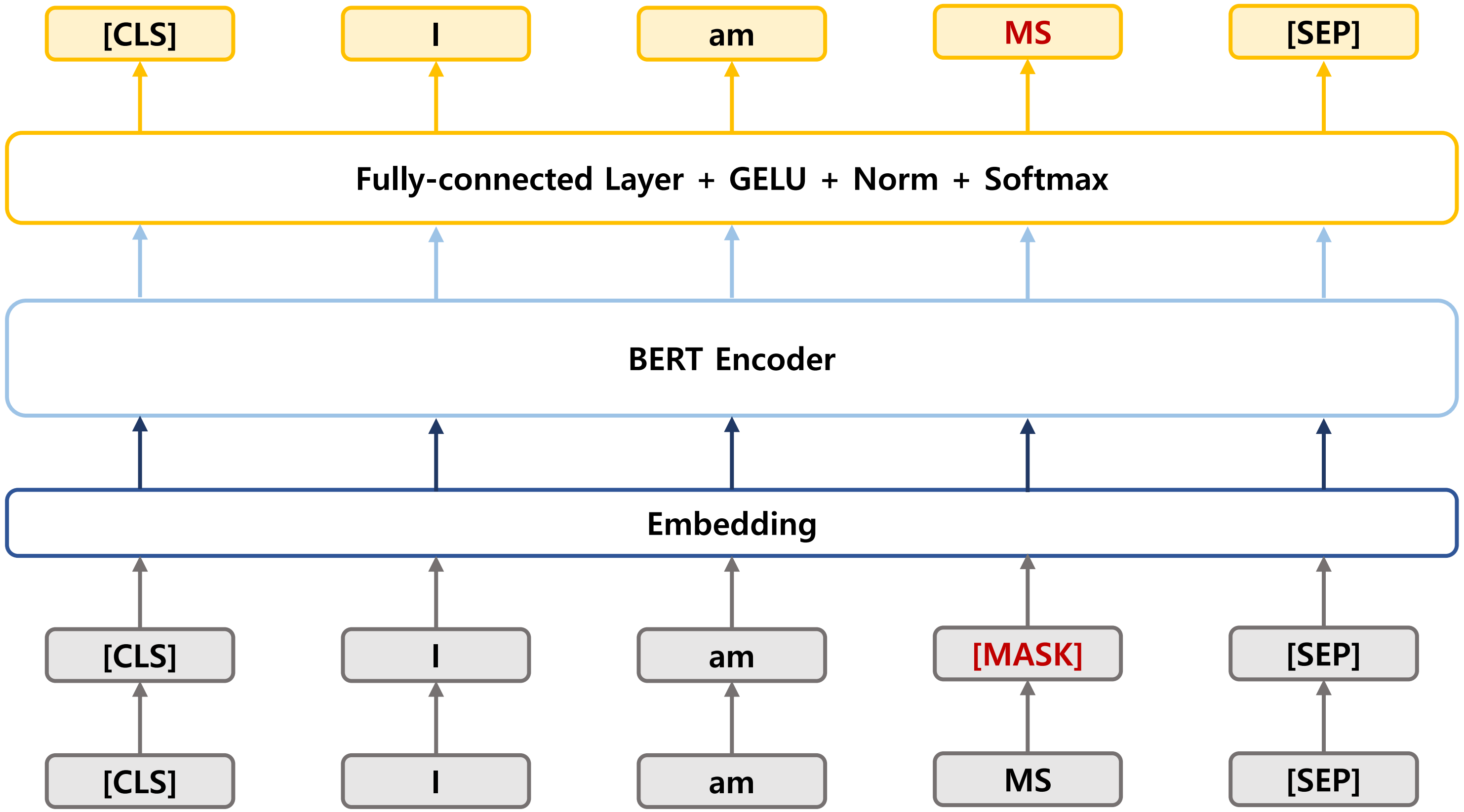
-BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

<Encoder>



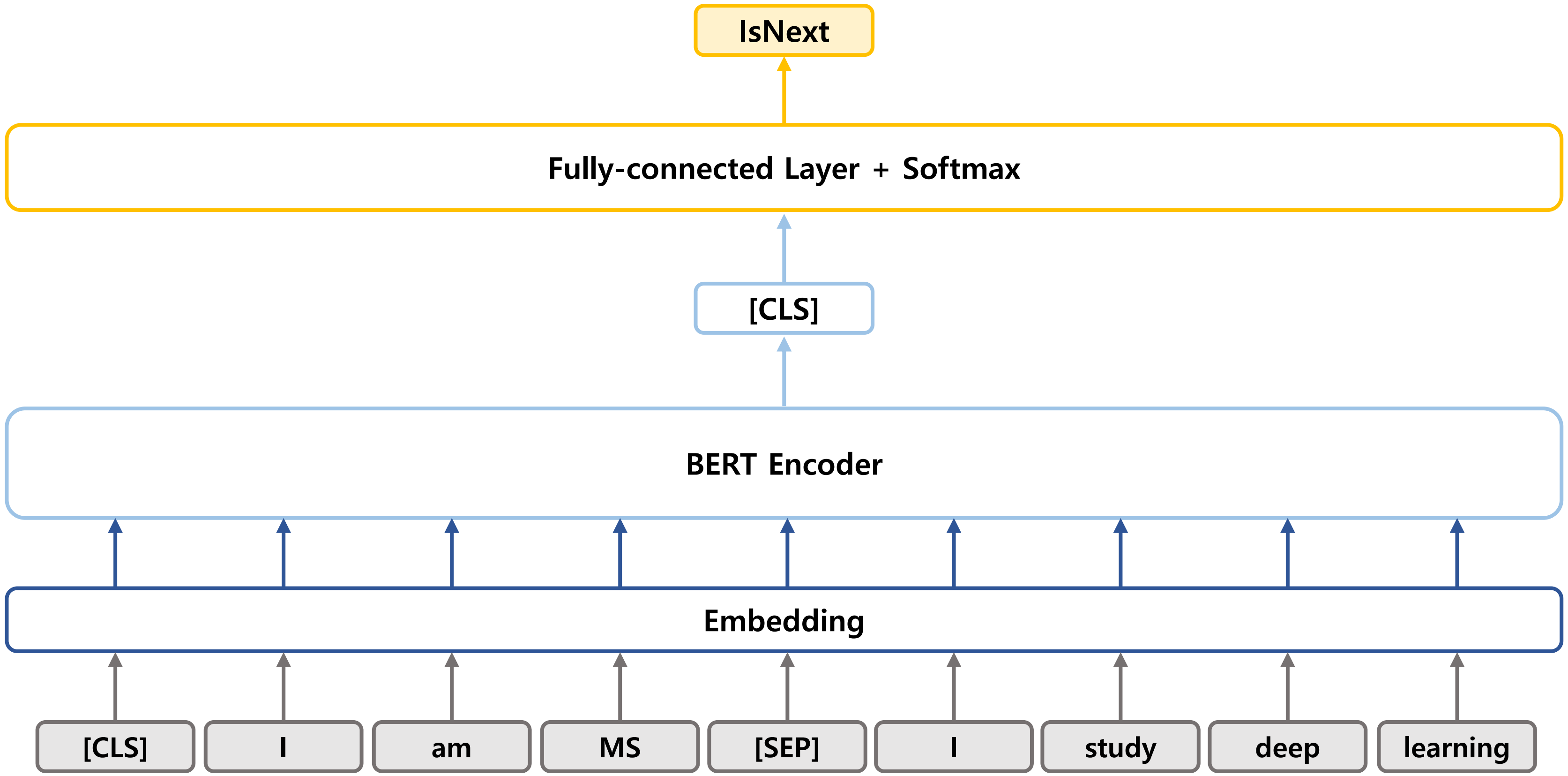
<Masked Language Model>

Random으로 15%의 Token을 [MASK] Token으로 변환한 후, 주변 문맥을 이용하여 [MASK] Token을 예측하는 Pre-training 방법

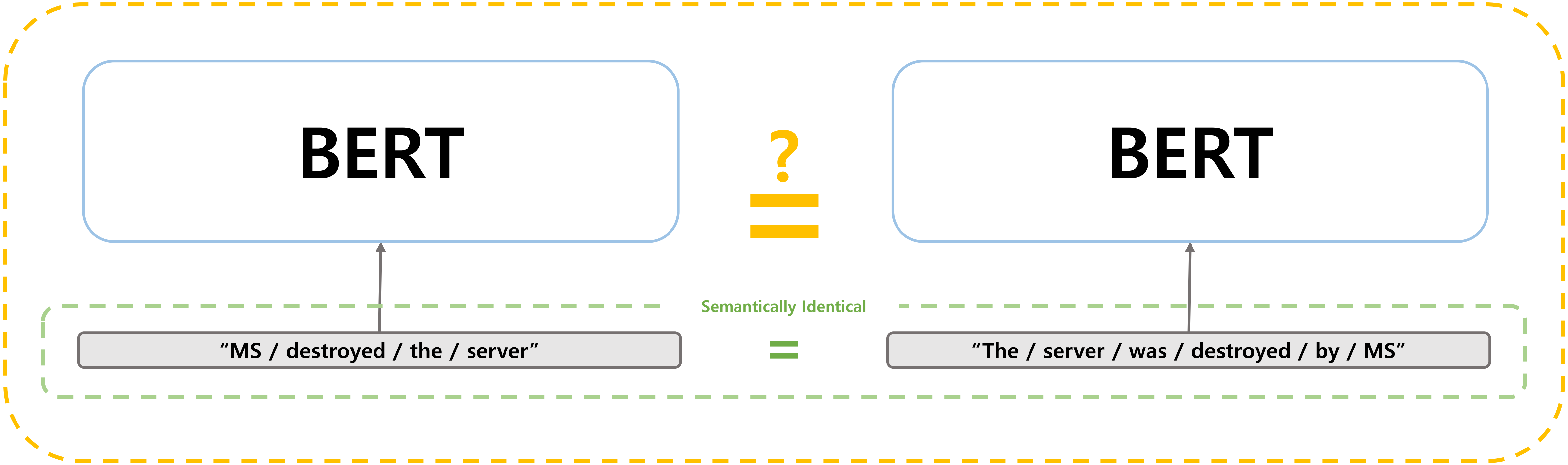


<Next Sentence Prediction>

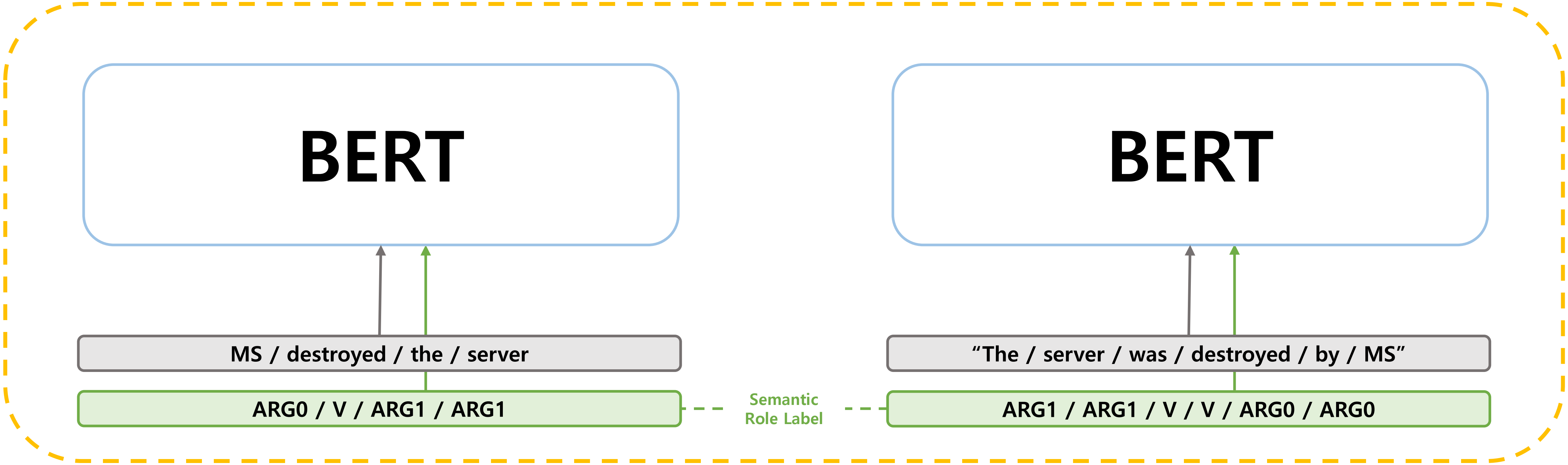
Corpus에서 두 문장을 이어 붙여 해당 문장이 원래 Corpus에서 이어져 있던 문장인지를 예측하는 Pre-training 방법



<Does BERT Understand Semantics?>

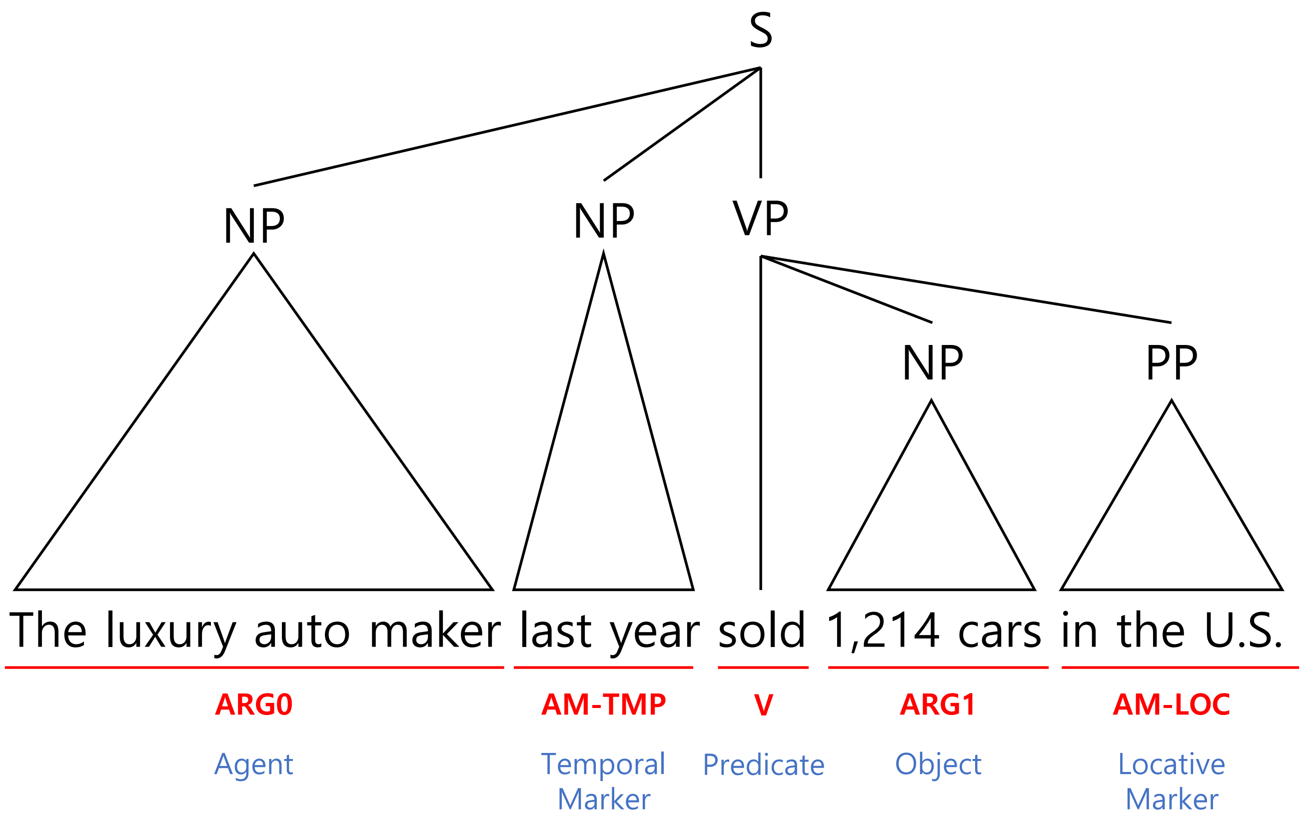


<Does BERT Understand Semantics?>



<Semantic Role Labeling>

문장에서 서술어(Predicate)를 중심으로 “누가(Who), 무엇을(What), 어떻게(How), 왜(Why)” 등을 인식하는 Task, 서술어에 대한 논항(Argument)를 찾고, 각 논항의 의미 역할을 결정



1 Prerequistes
-Semantic Role Labeling

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

↓ tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

↓ tagging

"Steve Jobs born in 1955"

"Steve Jobs co founded Apple in 1976"

<BIO Tagging>

“Beginning, Inside, Outside”의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

“Born in 1955, Steve Jobs co-founded Apple in 1976”

tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

tagging

“Steve Jobs born in 1955”

V-B										

Verb-Beginning

“Steve Jobs co founded Apple in 1976”

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

tagging

"Steve Jobs born in 1955"

V-B	AM-TMP-B									

Adjuncts temporal-Beginning

"Steve Jobs co founded Apple in 1976"

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	**in**	**1955**	**,**	**Steve**	**Jobs**	**co-**	**founded**	**Apple**	**in**	**1976**

tagging

"Steve Jobs born in 1955"

V-B	**AM-TMP-B**	**AM-TMP-I**								

Adjuncts temporal-Inside

"Steve Jobs co founded Apple in 1976"

1 Prerequistes
-Semantic Role Labeling

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

tagging

"Steve Jobs born in 1955"

V-B	AM-TMP-B	AM-TMP-I	O							
			Outside							

"Steve Jobs co founded Apple in 1976"

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	**in**	**1955**	**,**	**Steve**	**Jobs**	**co-**	**founded**	**Apple**	**in**	**1976**

tagging

"Steve Jobs born in 1955"

V-B	AM-TMP-B	AM-TMP-I	O	ARG0-B	ARG0-I					
				Argument-Biginning	*Argument-Inside*					

"Steve Jobs co founded Apple in 1976"

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

tagging

"Steve Jobs born in 1955"

V-B	AM-TMP-B	AM-TMP-I	O	ARG0-B	ARG0-I	O	O	O	O	O
						Outside	*Outside*	*Outside*	*Outside*	*Outside*

"Steve Jobs co founded Apple in 1976"

<BIO Tagging>

"Beginning, Inside, Outside"의 약자로,
Token-level의 Tag를 통해 각 Argument 또는 Predicate의 시작과 끝을 표기하는 기법

"Born in 1955, Steve Jobs co-founded Apple in 1976"

tokenizing

Born	in	1955	,	Steve	Jobs	co-	founded	Apple	in	1976

tagging

"Steve Jobs born in 1955"

V-B	AM-TMP-B	AM-TMP-I	O	ARG0-B	ARG0-I	O	O	O	O	O

"Steve Jobs co founded Apple in 1976"

O	O	O	O	ARG0-B	ARG0-I	V-B	V-I	ARG1-B	AM-TMP-B	AM-TMP-I

<Proposition Bank Style>

Palmer et al., 2004

Arguments (ARG-, AA)		Adjuncts (AM-)	
ARG0	agent	AM-ADV	general-purpose
ARG1	patient, object	AM-CAU	cause
ARG2	…	AM-DIR	direction
ARG3	…	AM-DIS	discourse marker
ARG4	…	AM-EXT	extent
AA	…	AM-LOC	location
References (R-)		AM-MNR	manner
R-ARG0	reference of agent	AM-MOD	modea verb
R-ARG1	reference of patient	AM-NEG	negation marker
R-AM-TMP	reference of temporal	AM-PNC	purpose
…	…	AM-PRD	predication
…	…	AM-REC	reciprocal
Verbs (V)		AM-TMP	temporal
V	predicate	AM-INS	instrument

1 Prerequistes
-Semantic Role Labeling

<Syntactic Variation>

Yesterday,	Kristina	hit	Scott	with a baseball bat
AM-TMP	ARG0	V	ARG1	AM-INS
Temporal	Agent	Predicate	Object	Instrument

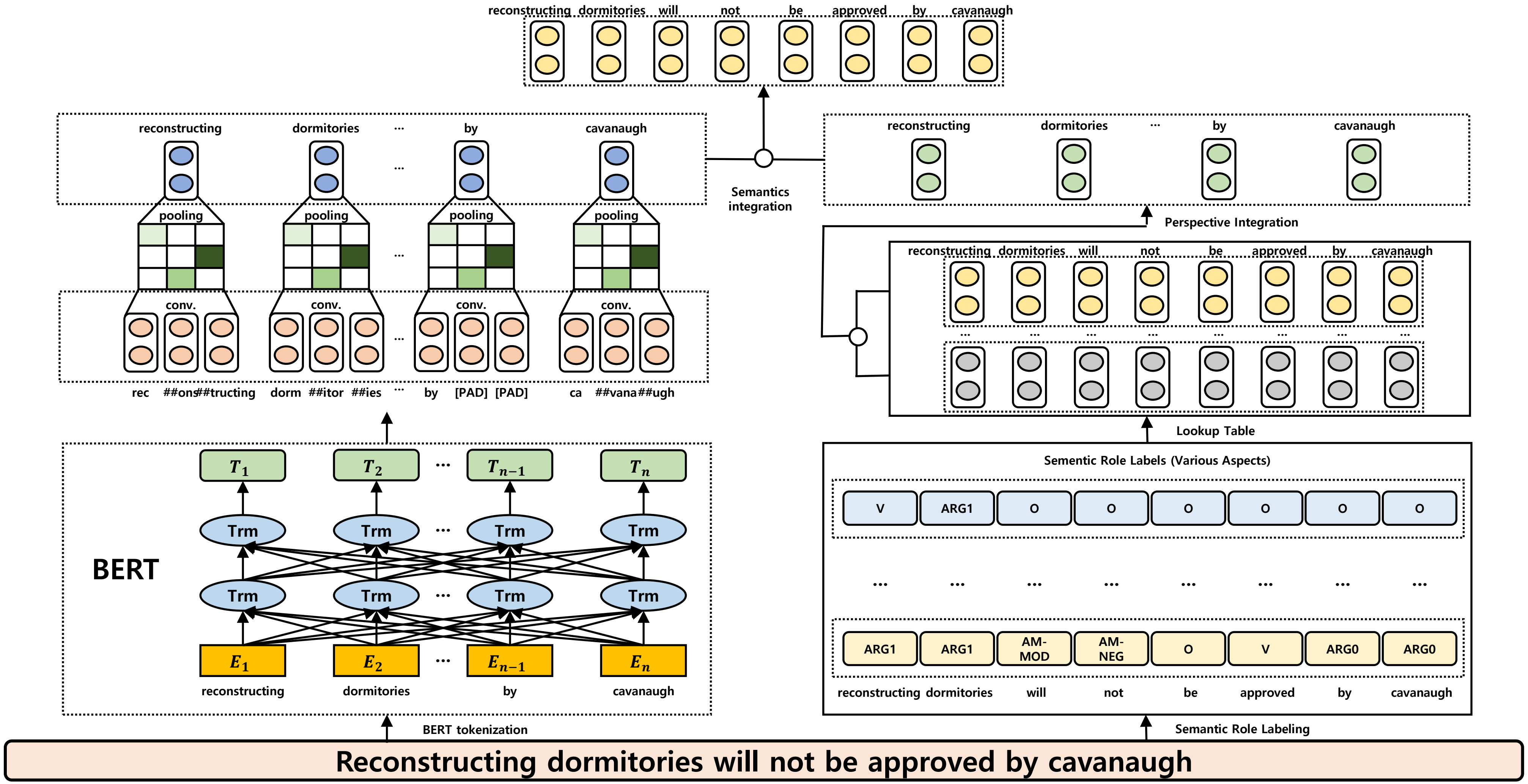
- ✓ Scott was hit by Kristina yesterday with a baseball bat
- ✓ Yesterday, Scott was hit with a baseball bat by Kristina
- ✓ With a baseball bat, Kristina hit Scott yesterday
- ✓ Yesterday Scott was hit by Kristina with a baseball bat
- ✓ Kristina hit Scott with a baseball bat yesterday

2. SemBERT

- Semantic Embedding
- Contextual Embedding
- Semantic Integration

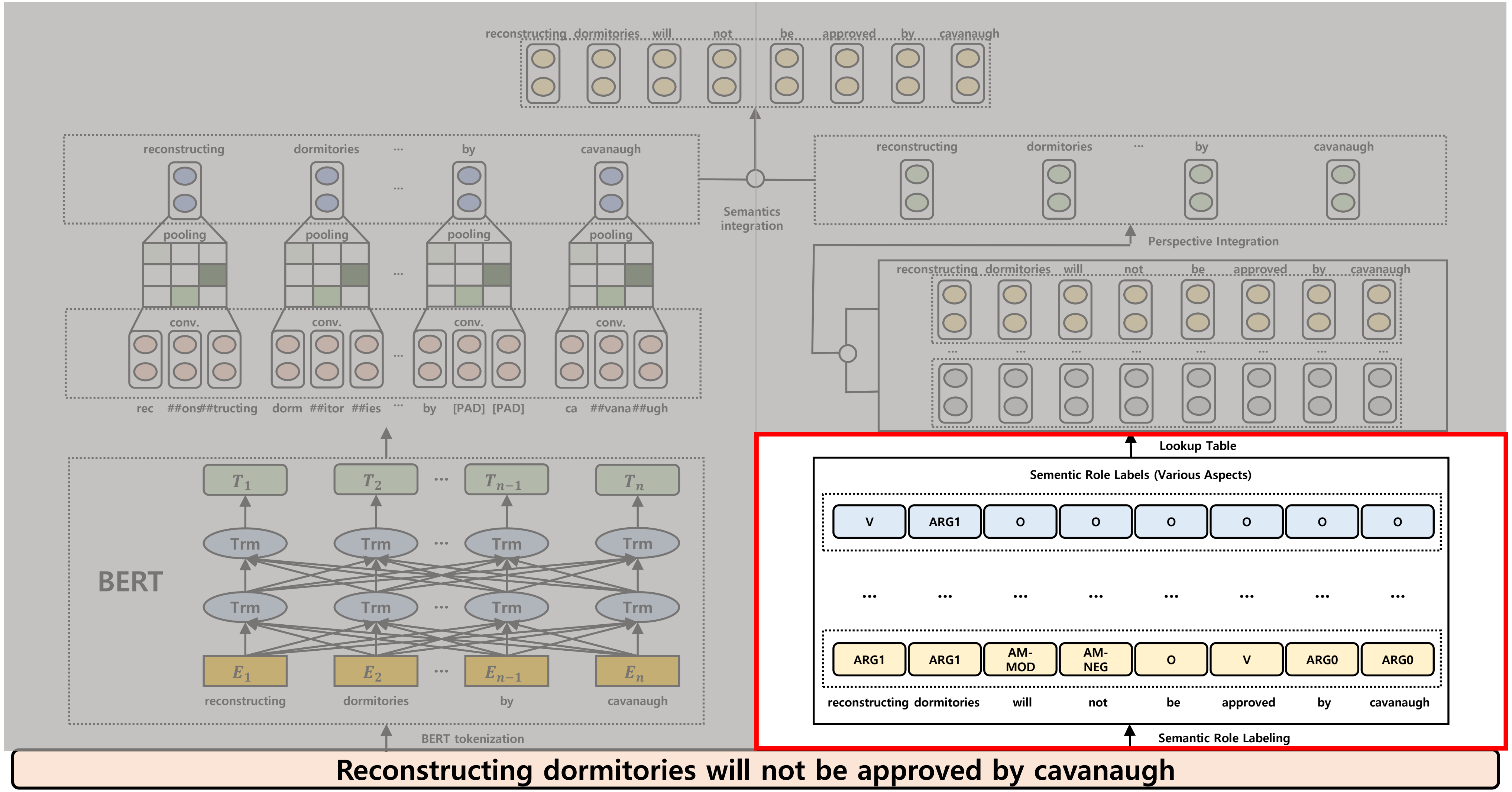
2 SemBERT
-Overall Architecture

<Overall Architecture>



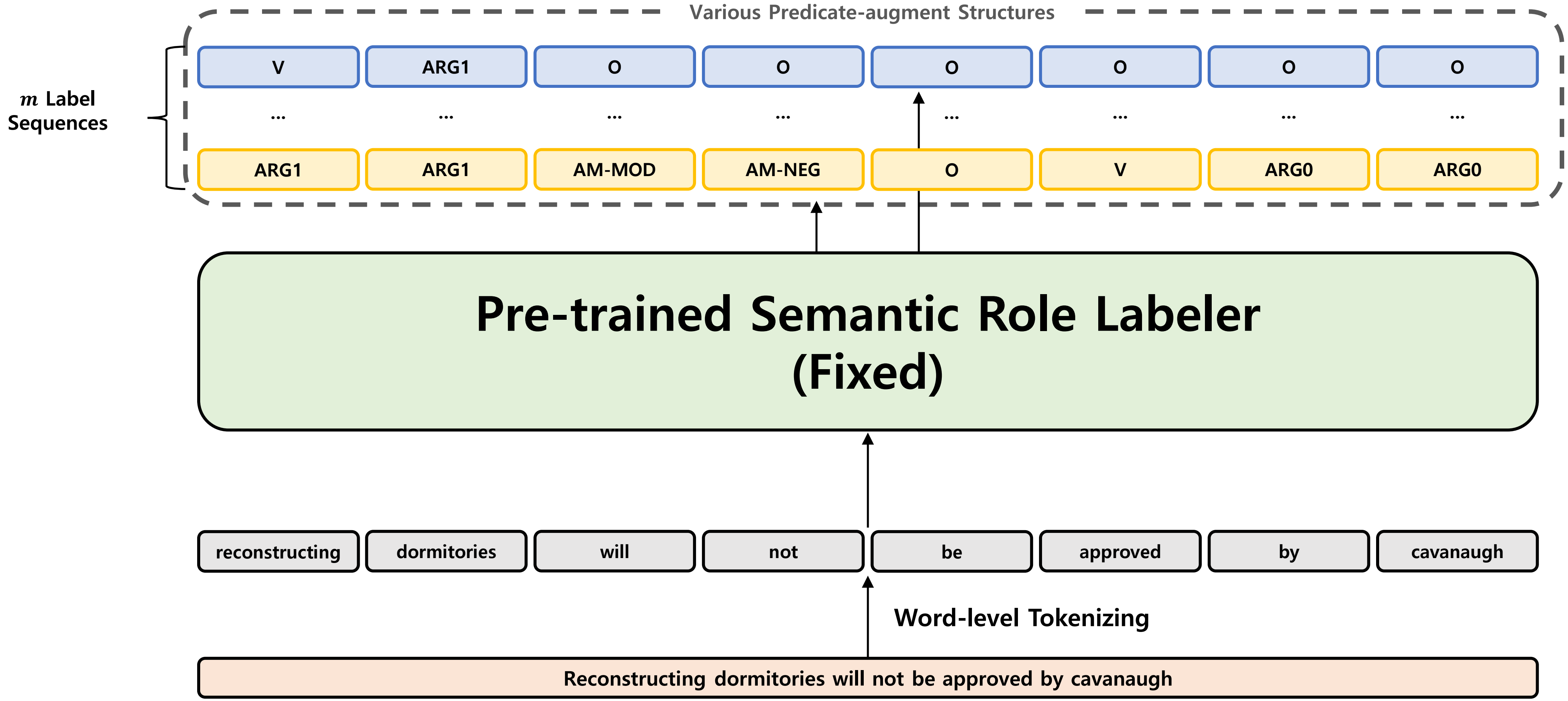
2 SemBERT
-Semantic Embedding

<Overall Architecture>



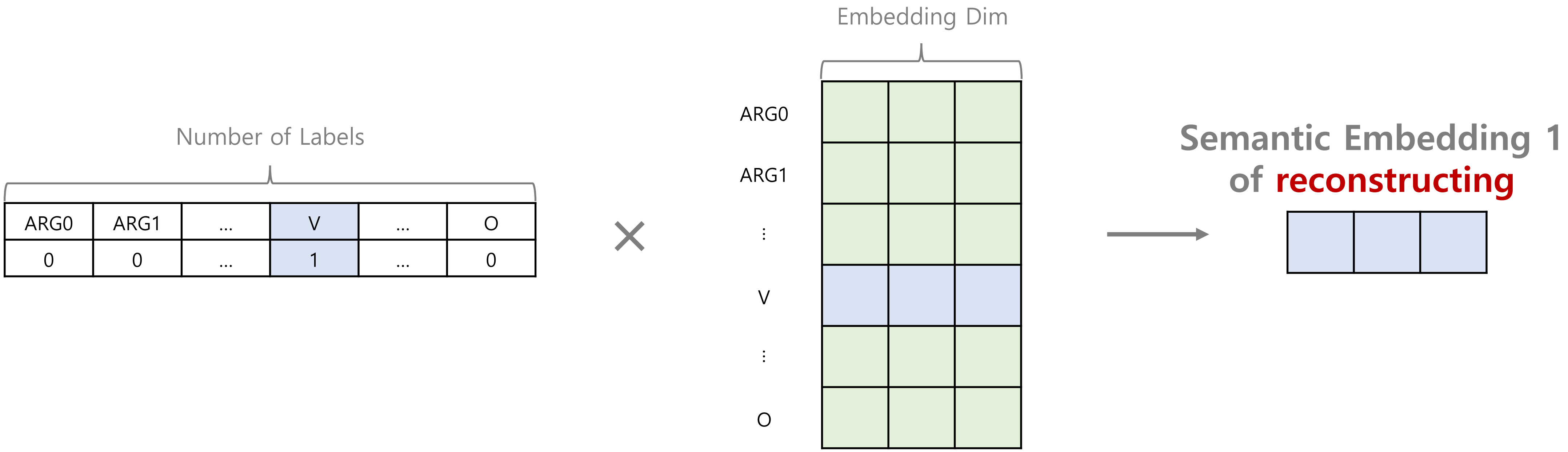
2 SemBERT
-Semantic Embedding

<Semantic Role Labeling>



<Semantic Embedding>

Input Sequence	reconstructing	dormitories	will	not	be	approved	by	cavanaugh
Semantic Role Label 1	V	ARG1	O	O	O	O	O	O
Semantic Role Label 2	ARG1	ARG1	AM-MOD	AM-NEG	O	V	ARG0	ARG0



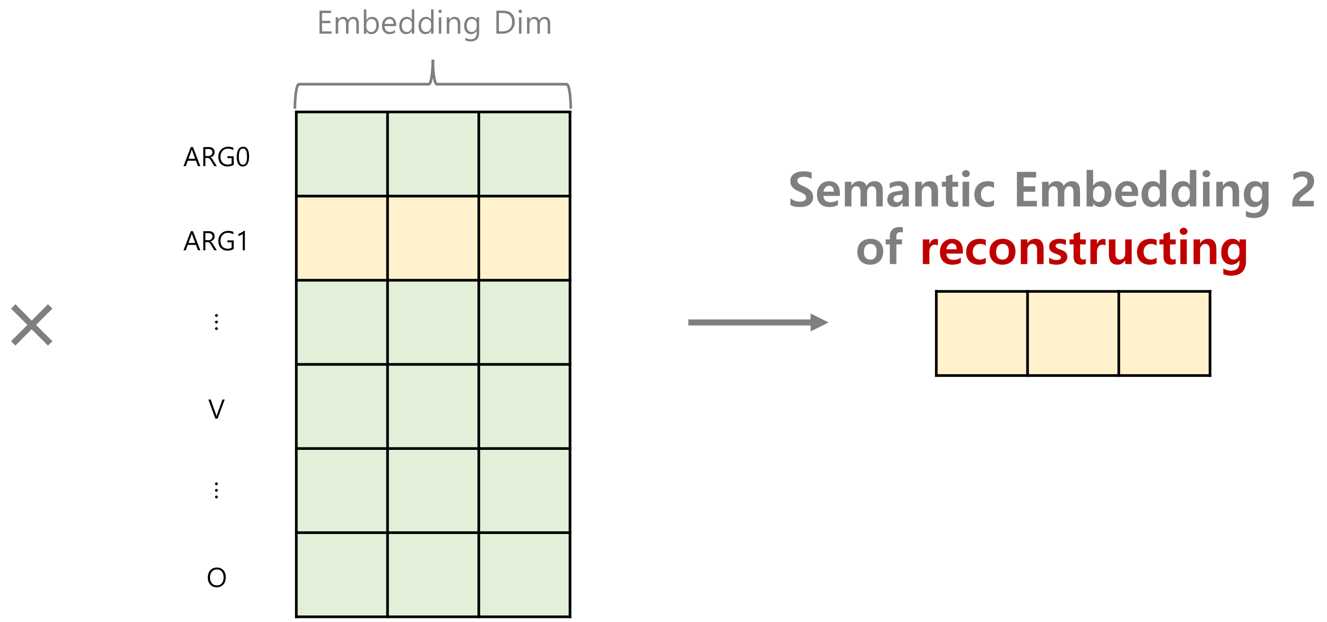
2 SemBERT
-Semantic Embedding

<Semantic Embedding>

Input Sequence	reconstructing	dormitories	will	not	be	approved	by	cavanaugh
Semantic Role Label 1	V	ARG1	O	O	O	O	O	O
Semantic Role Label 2	ARG1	ARG1	AM-MOD	AM-NEG	O	V	ARG0	ARG0

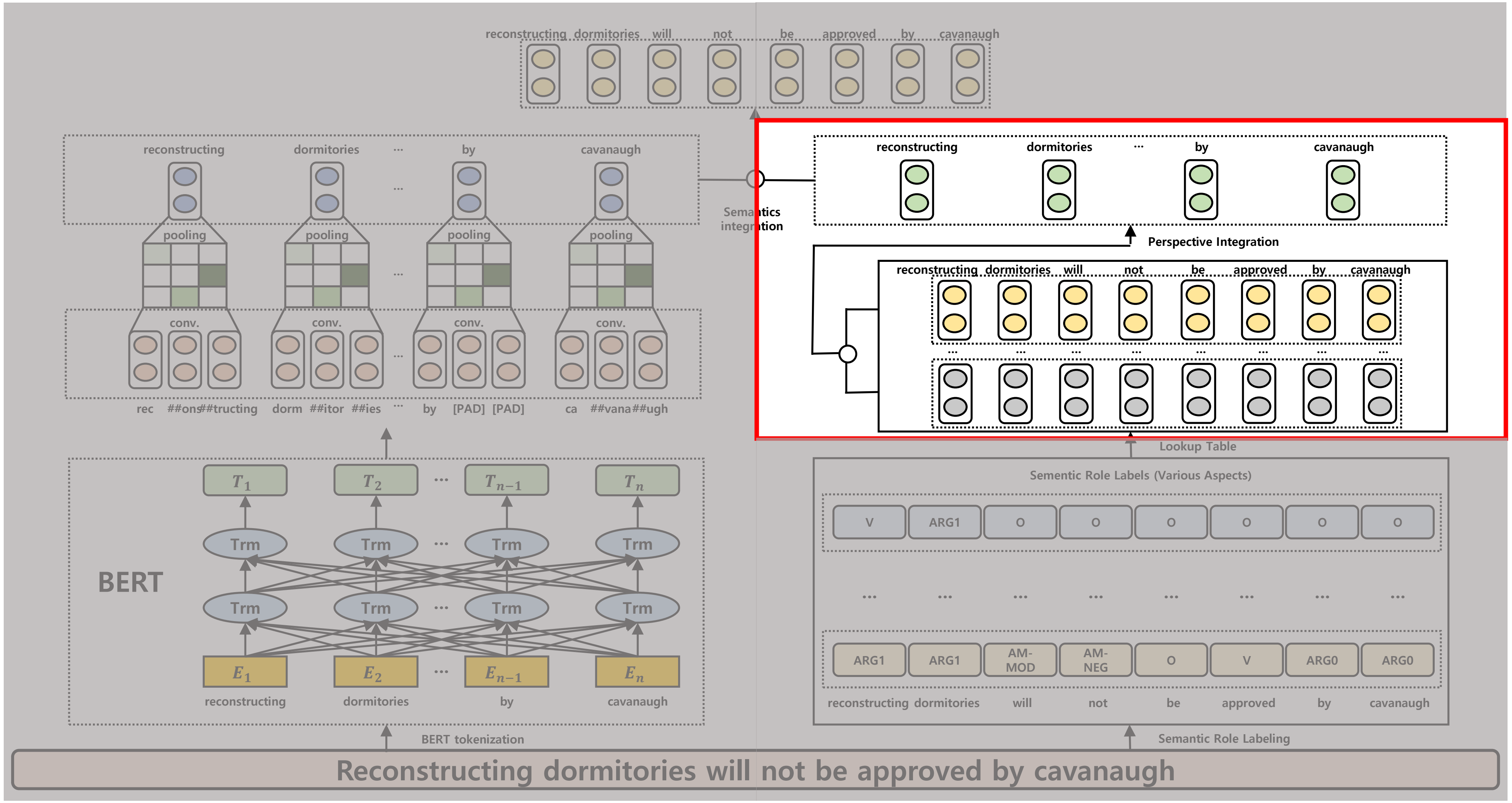
Number of Labels

ARG0	ARG1	...	V	...	O
0	1	...	0	...	0

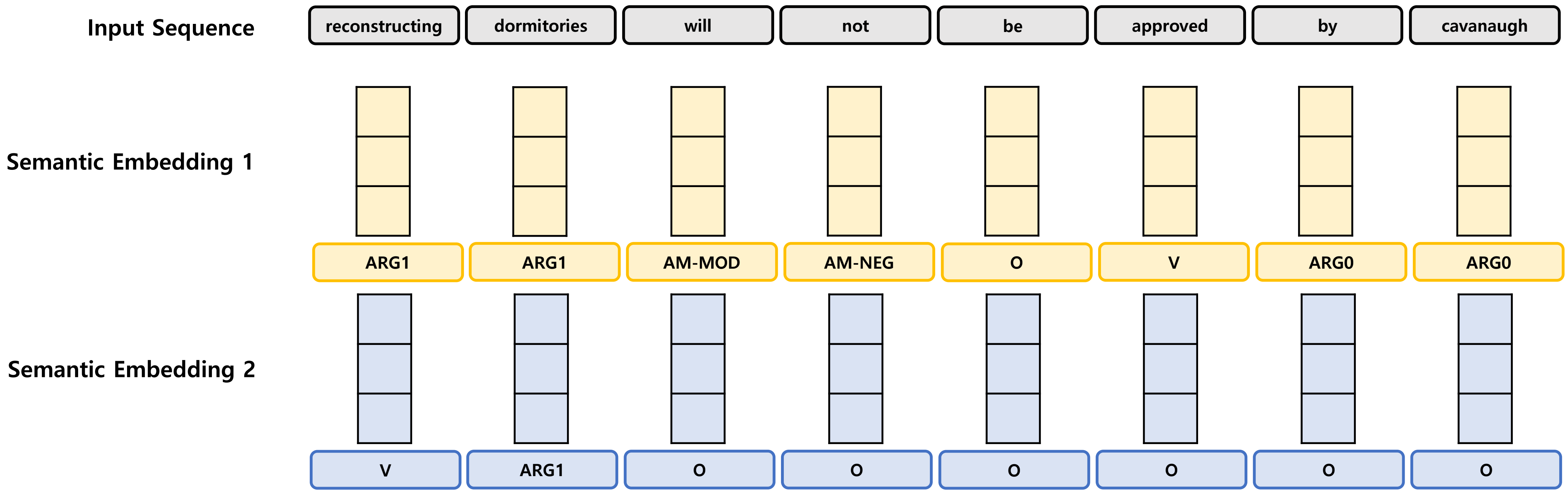


2 SemBERT
-Semantic Embedding

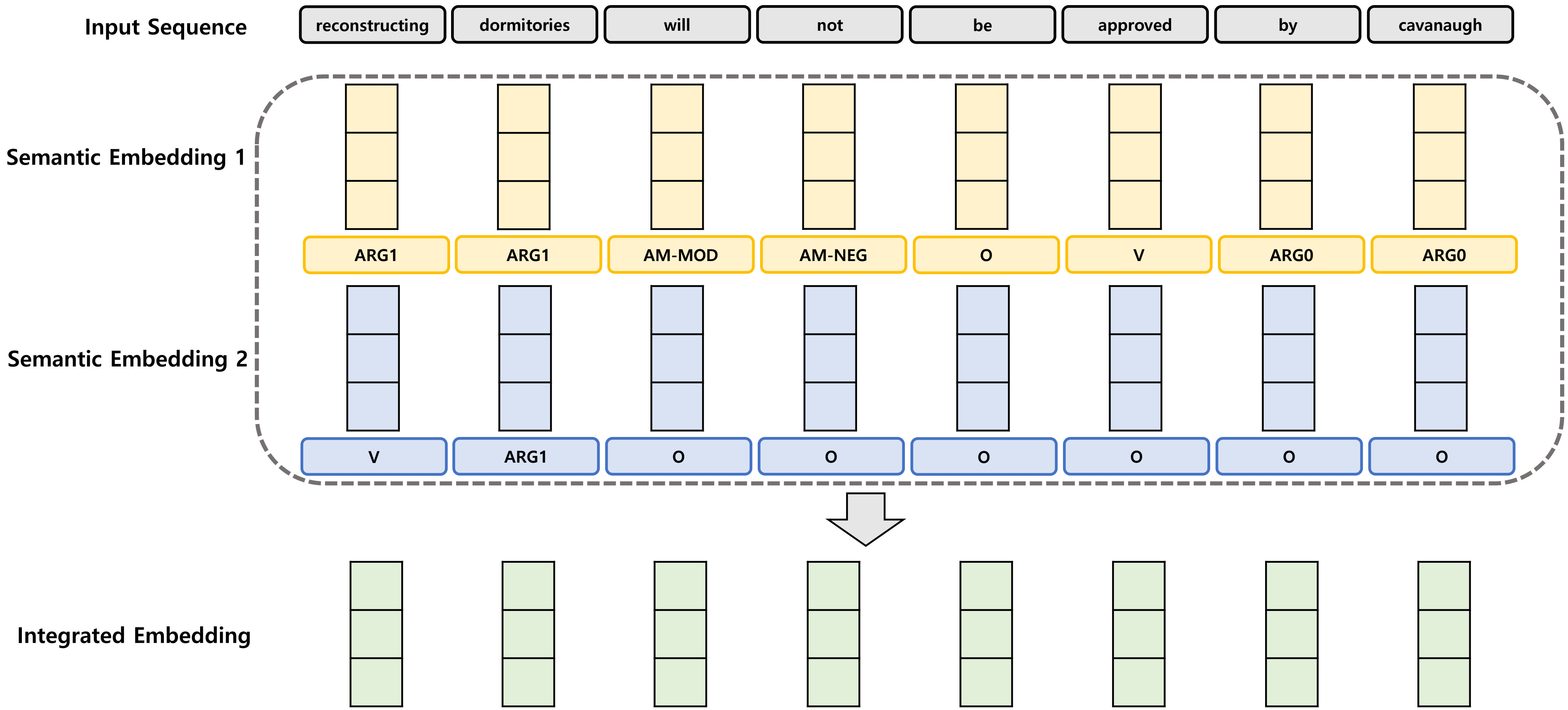
<Overall Architecture>



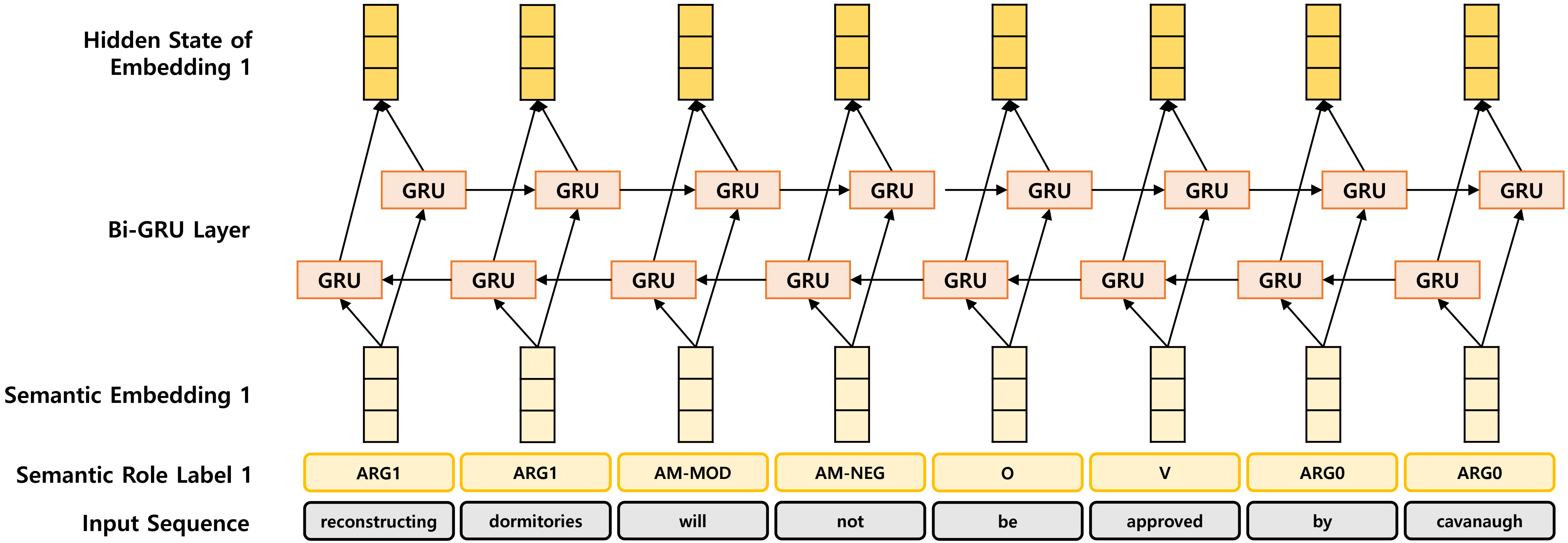
<Perspective Integration>



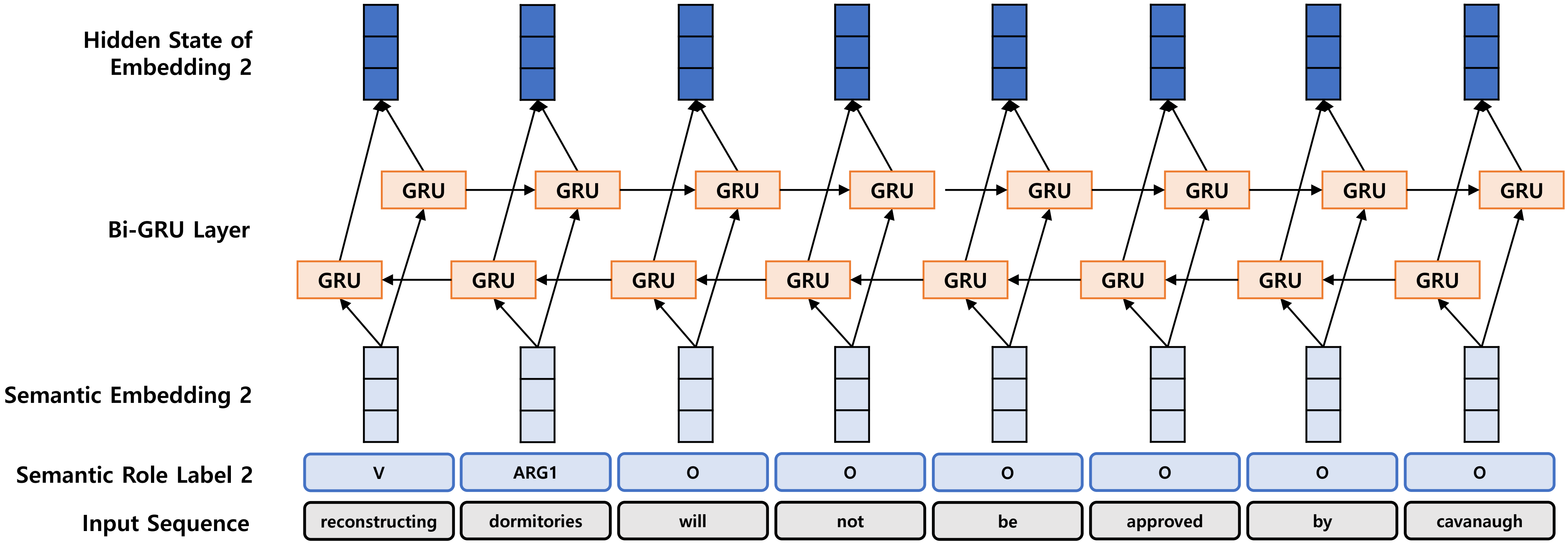
<Perspective Integration>



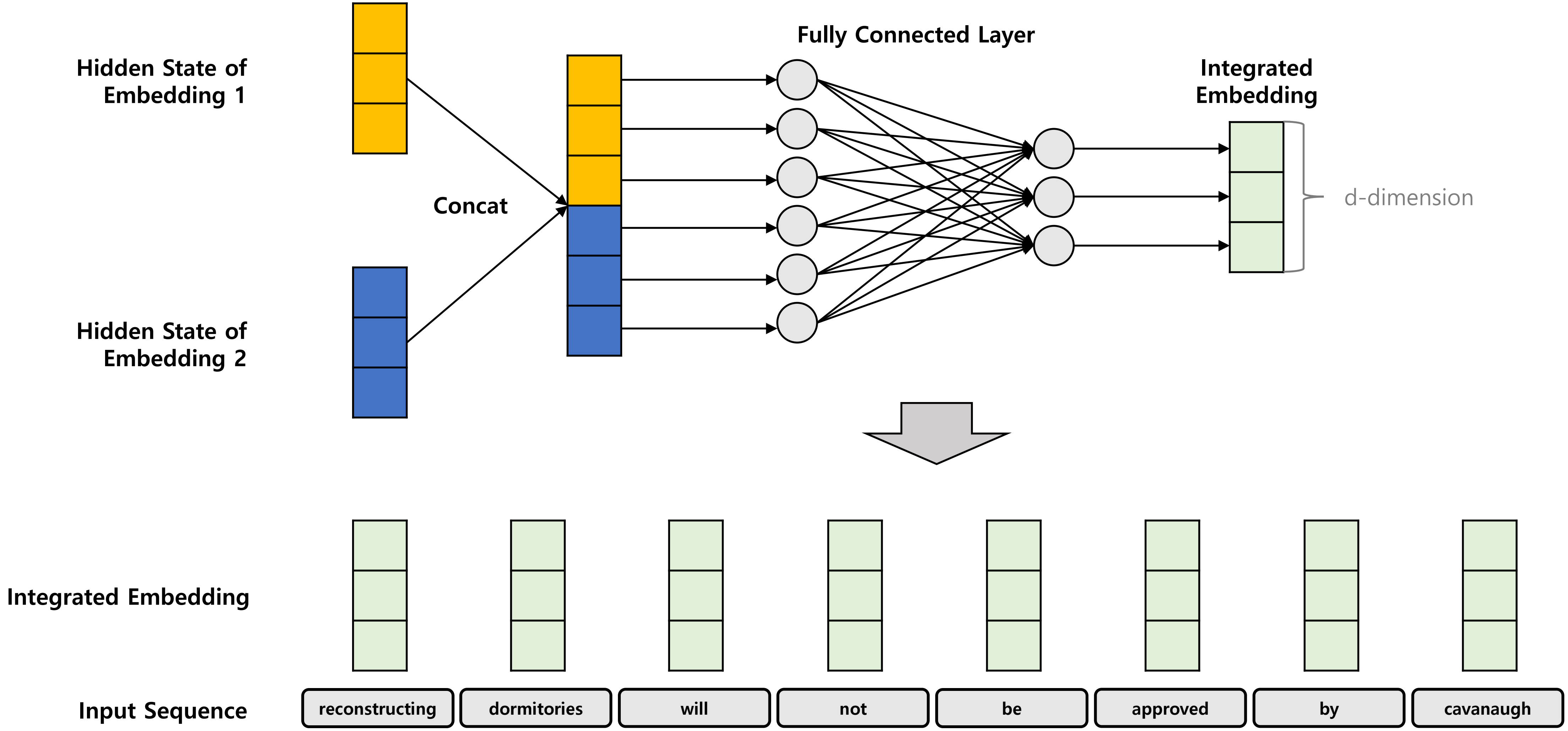
<Perspective Integration>



<Perspective Integration>

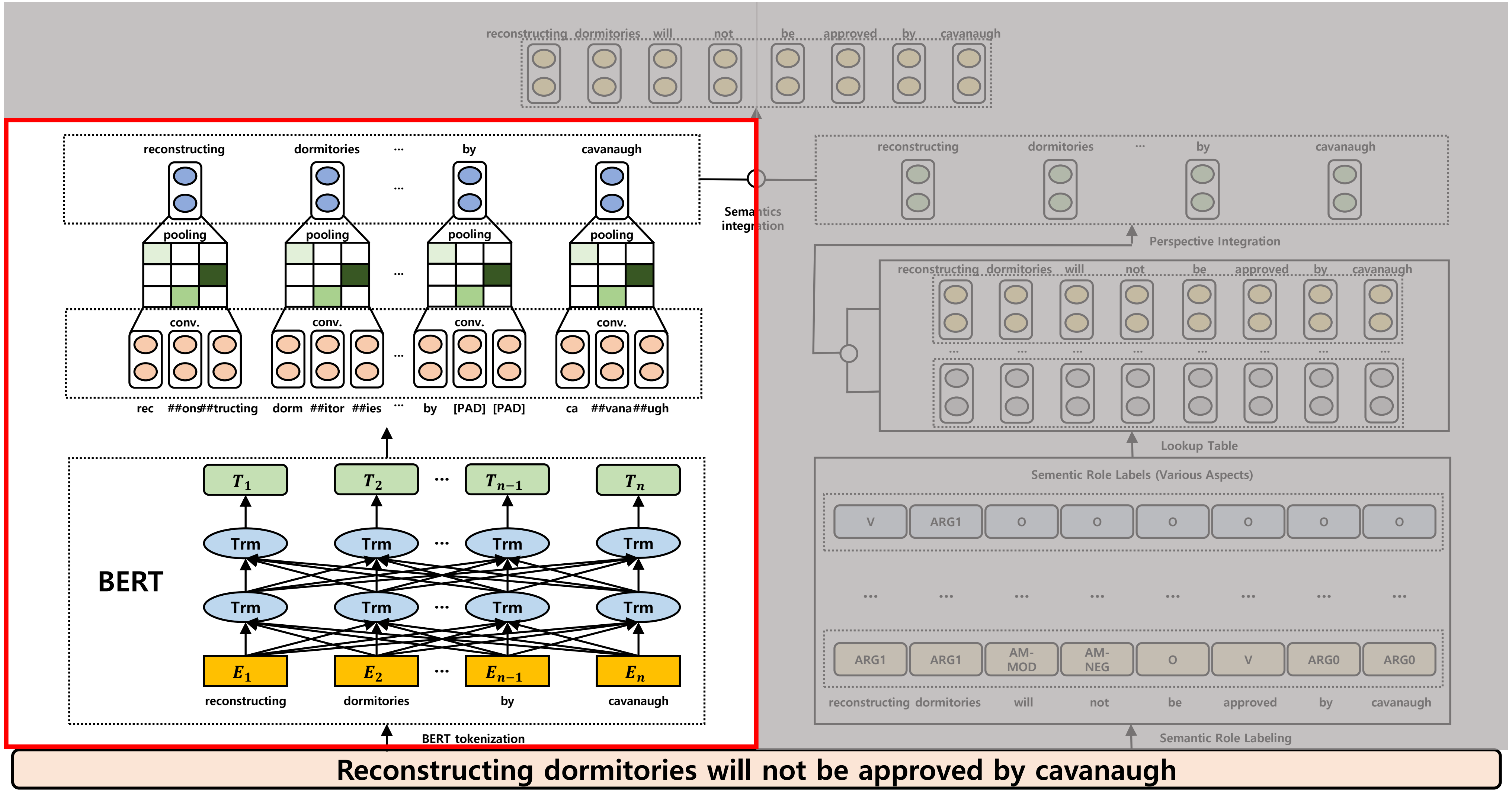


<Perspective Integration>



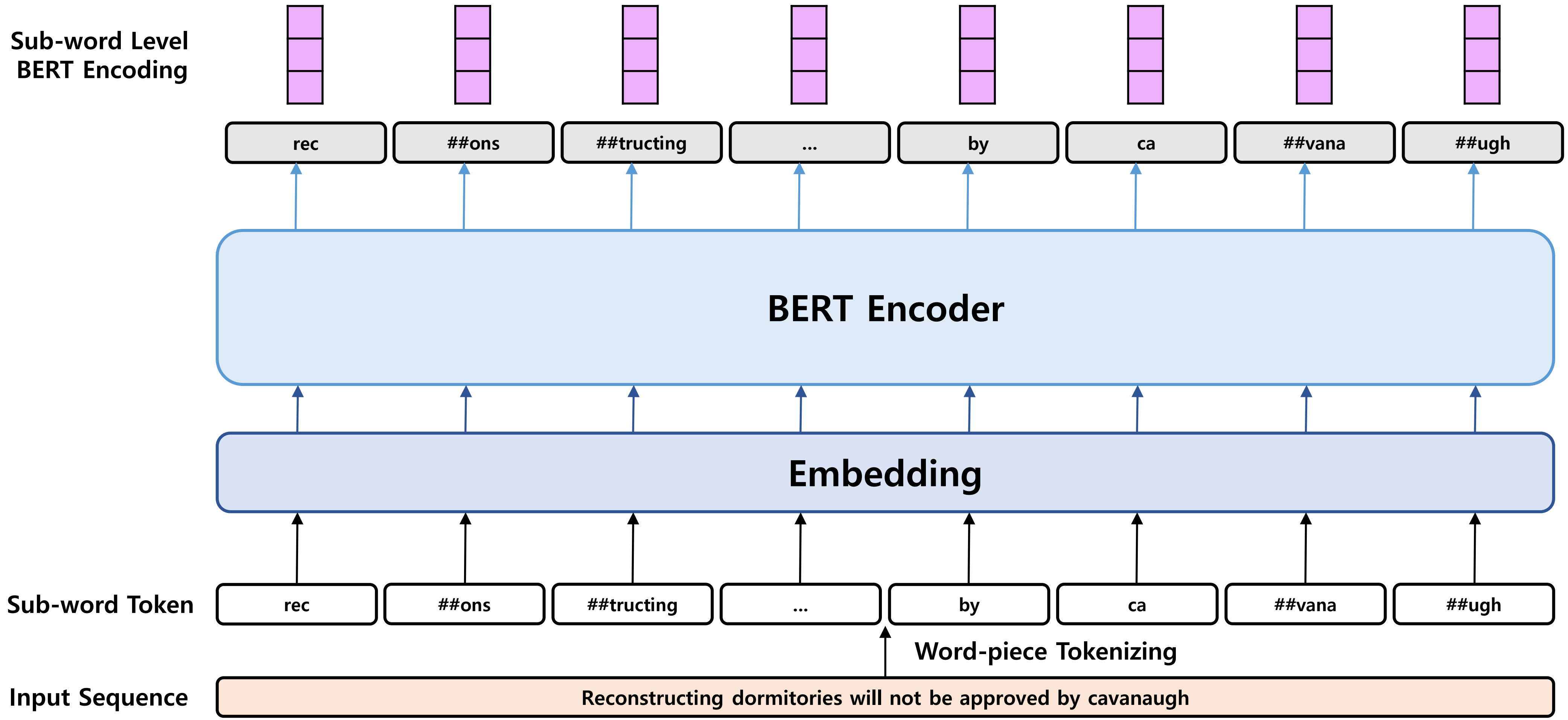
2 SemBERT
-Contextual Embedding

<Overall Architecture>

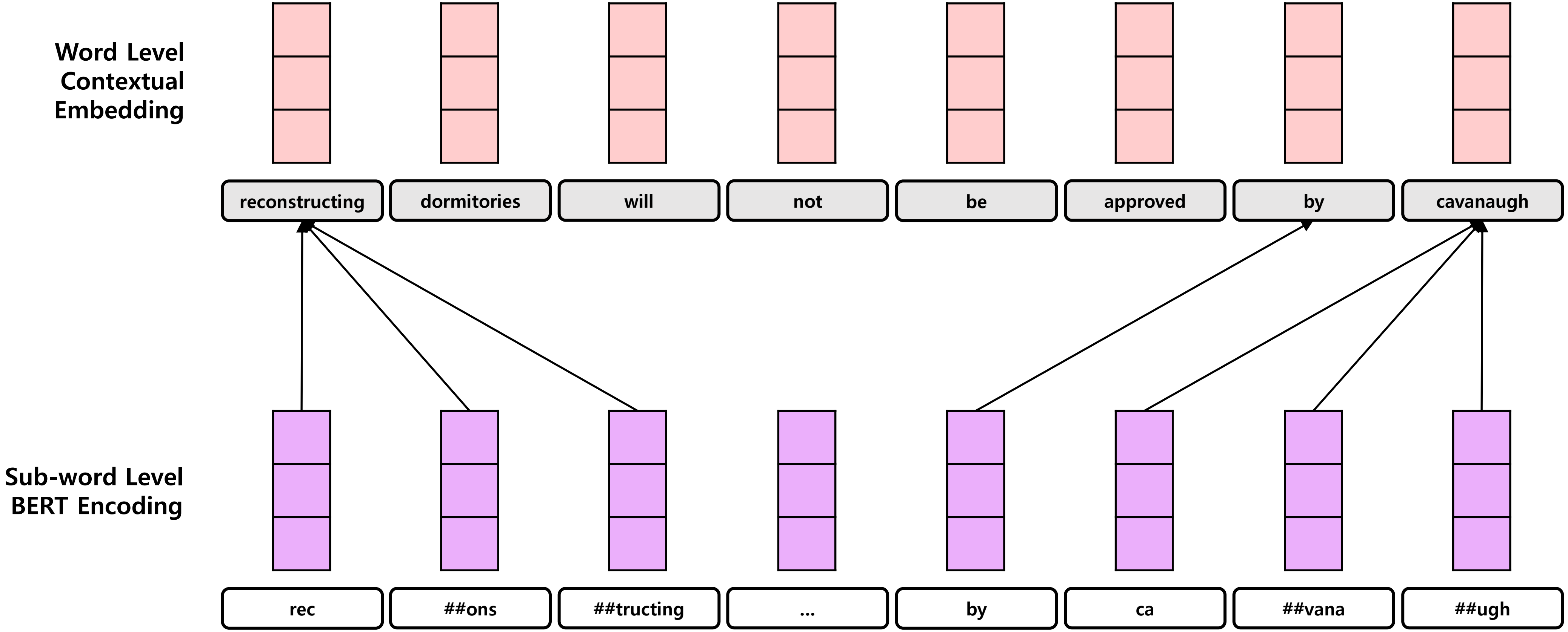


2 SemBERT
-Contextual Embedding

<Pre-trained BERT Encoding>

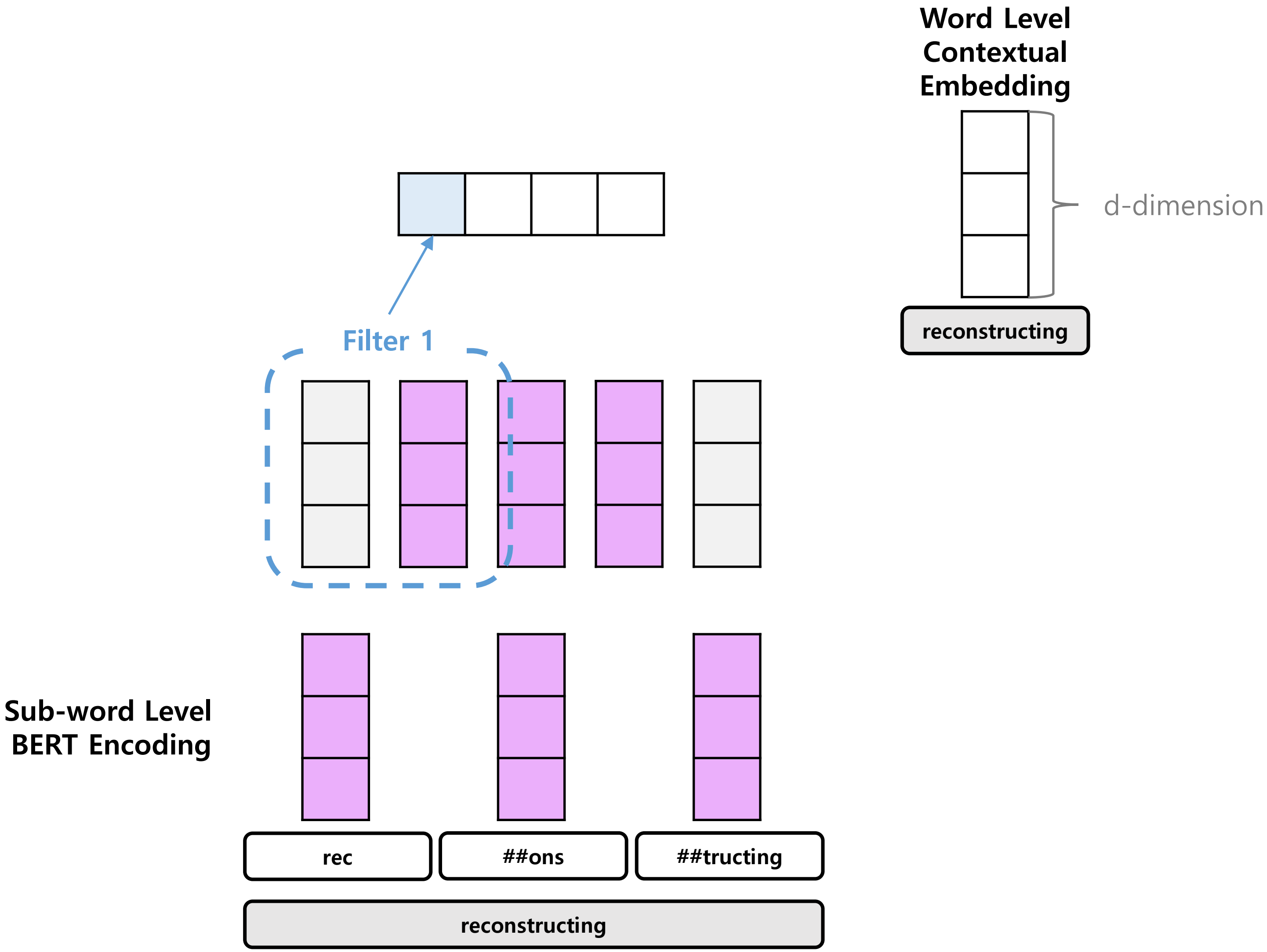


<Word Level Embedding>

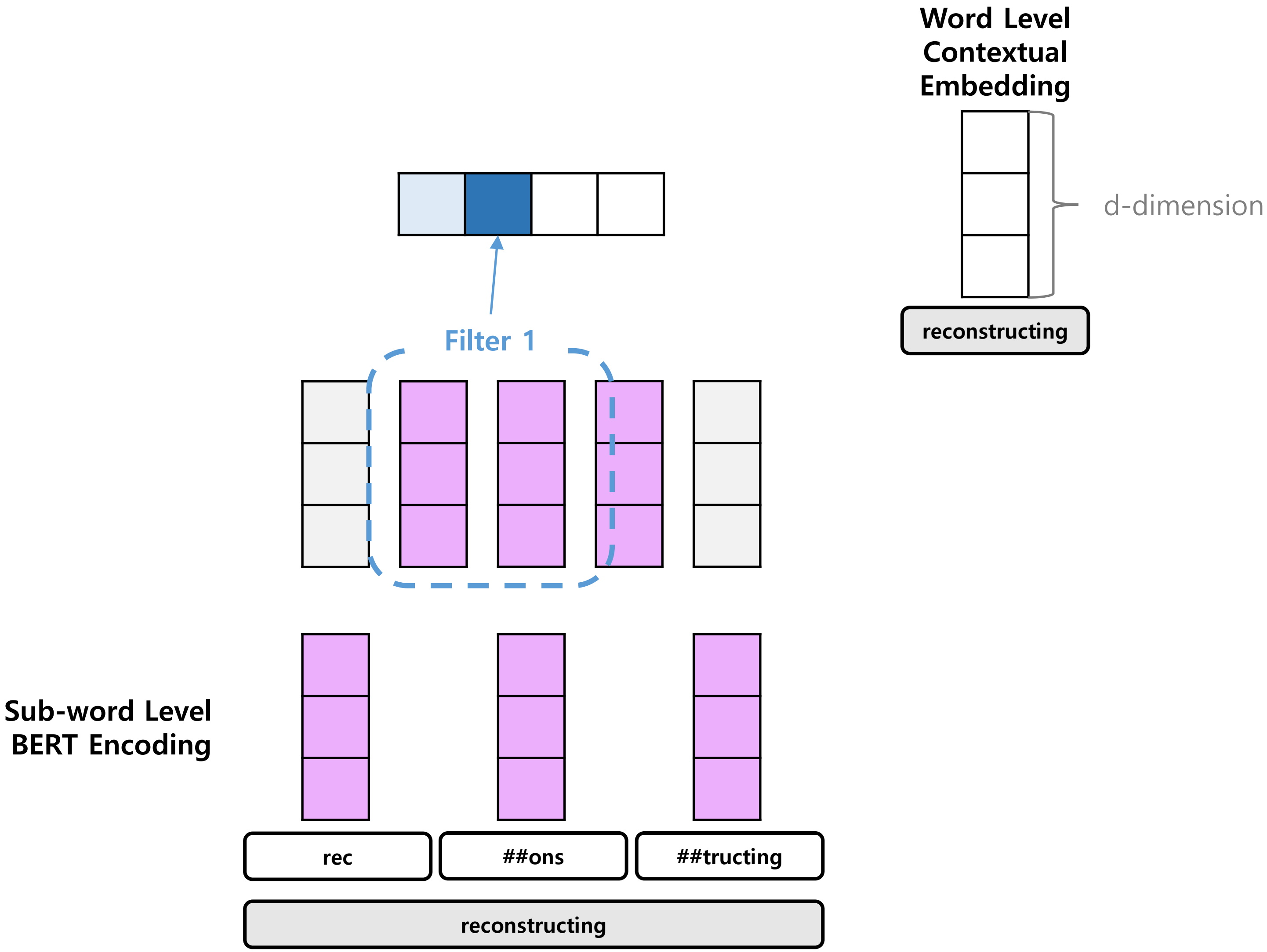


2 SemBERT
-Contextual Embedding

<1D Convolution>

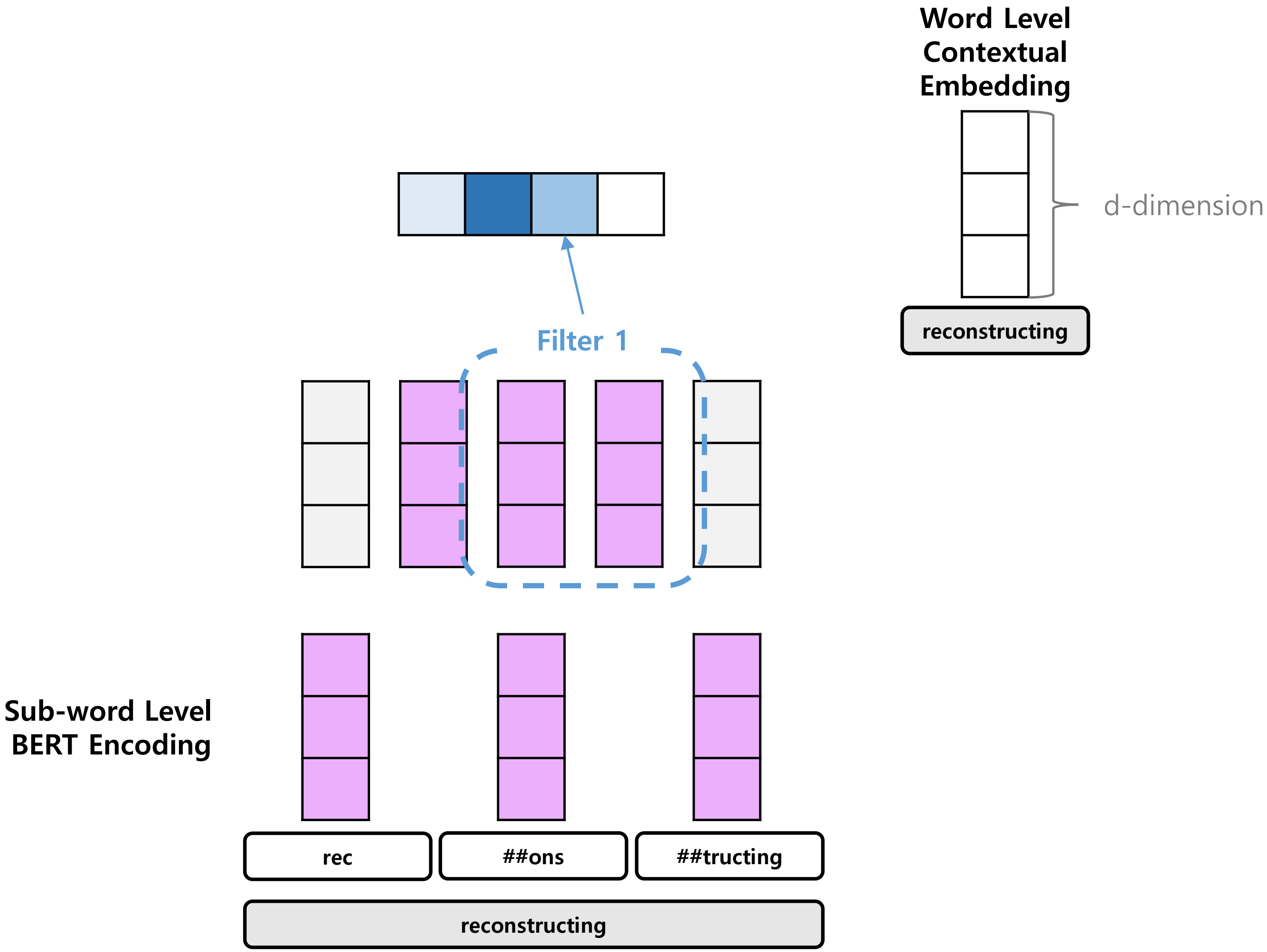


<1D Convolution>

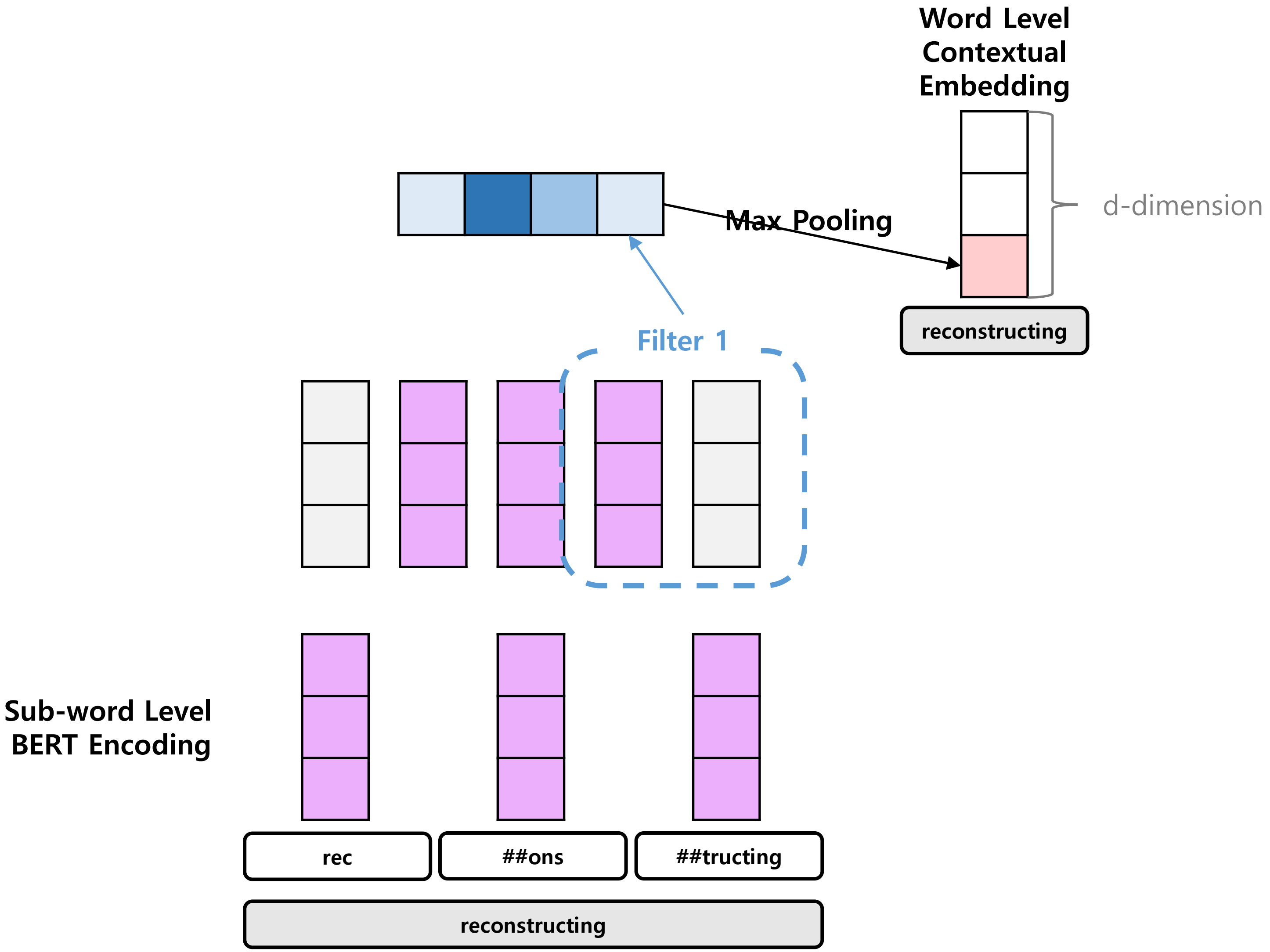


2 SemBERT
-Contextual Embedding

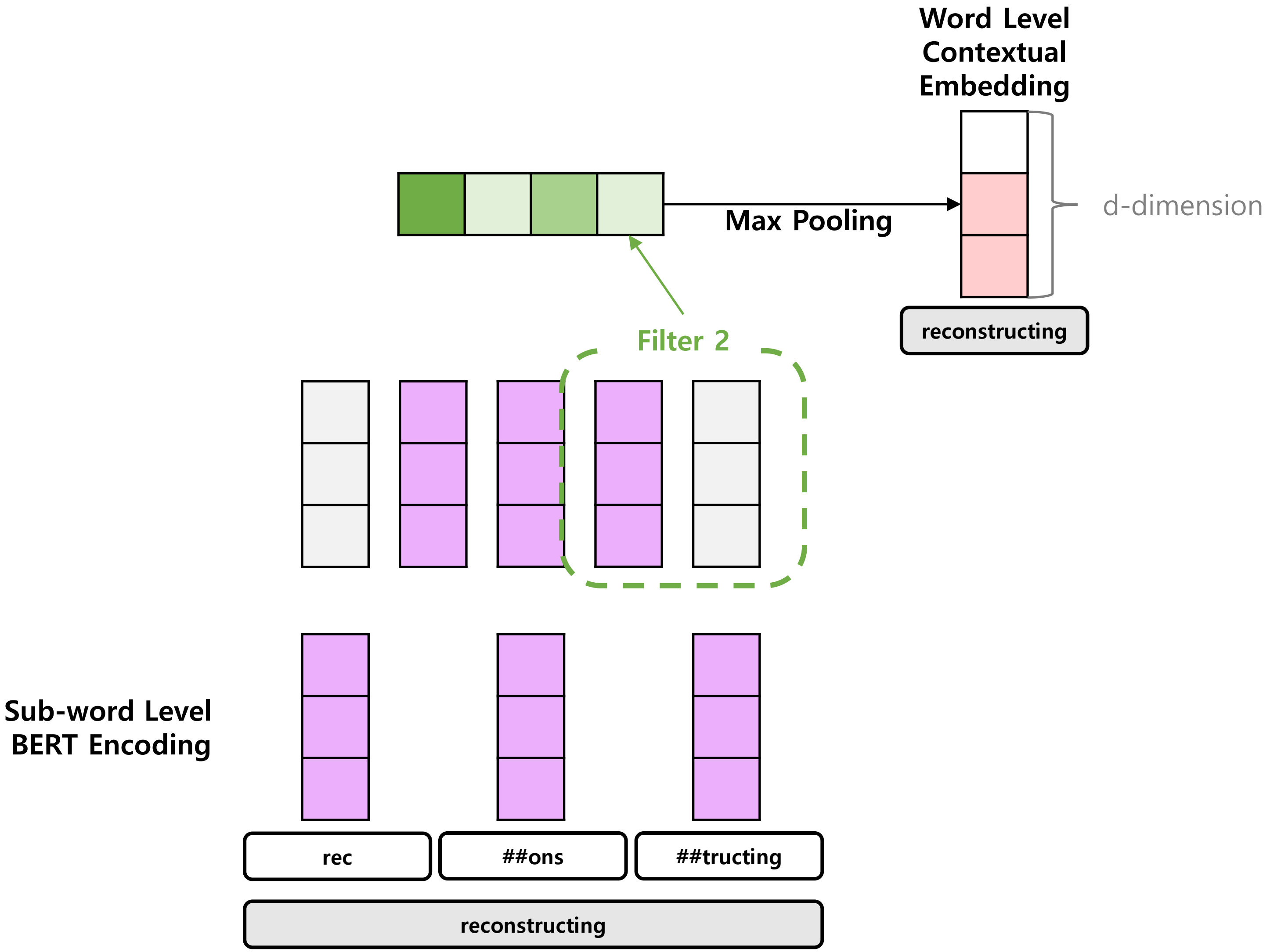
<1D Convolution>



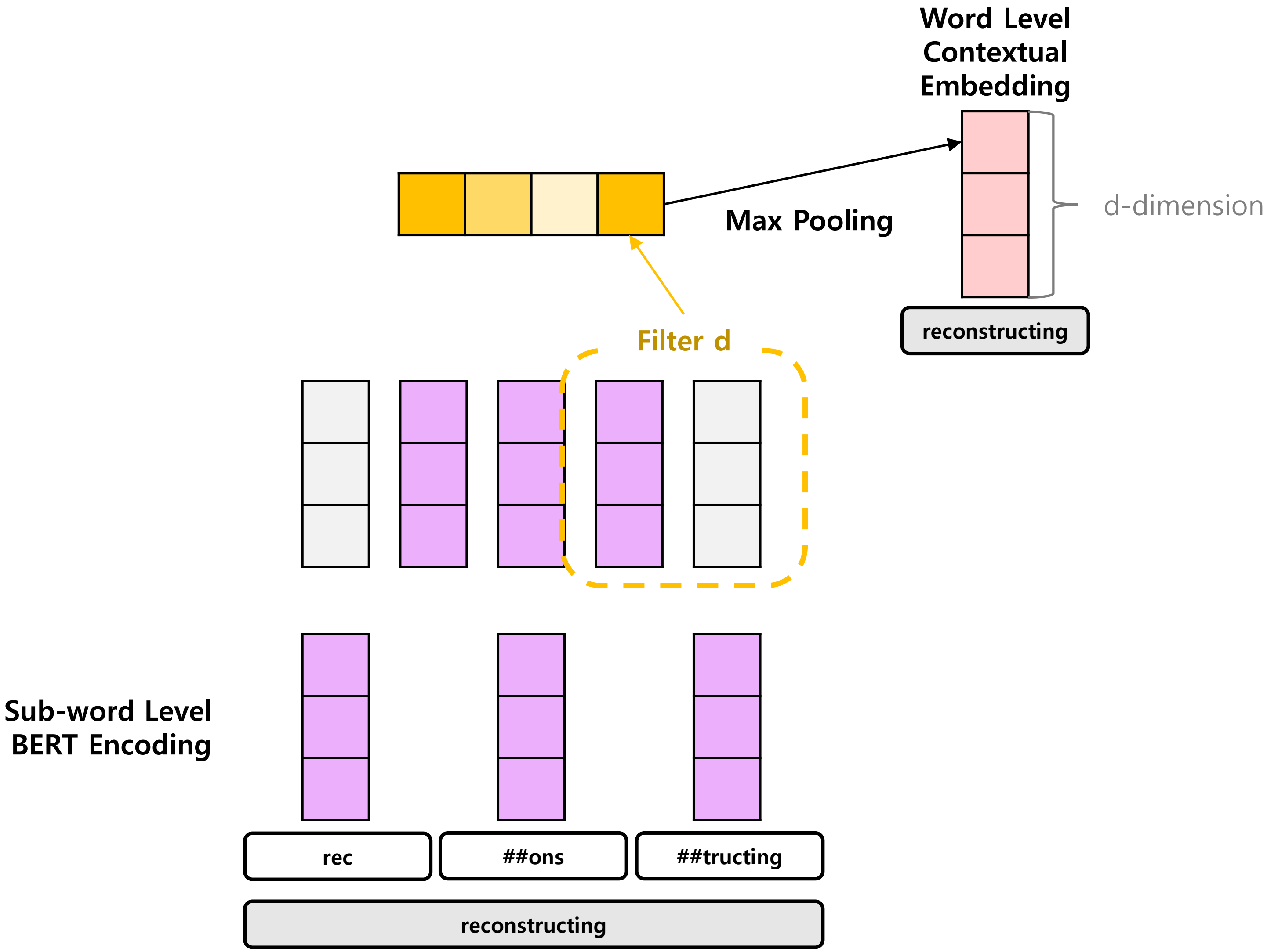
<1D Convolution>



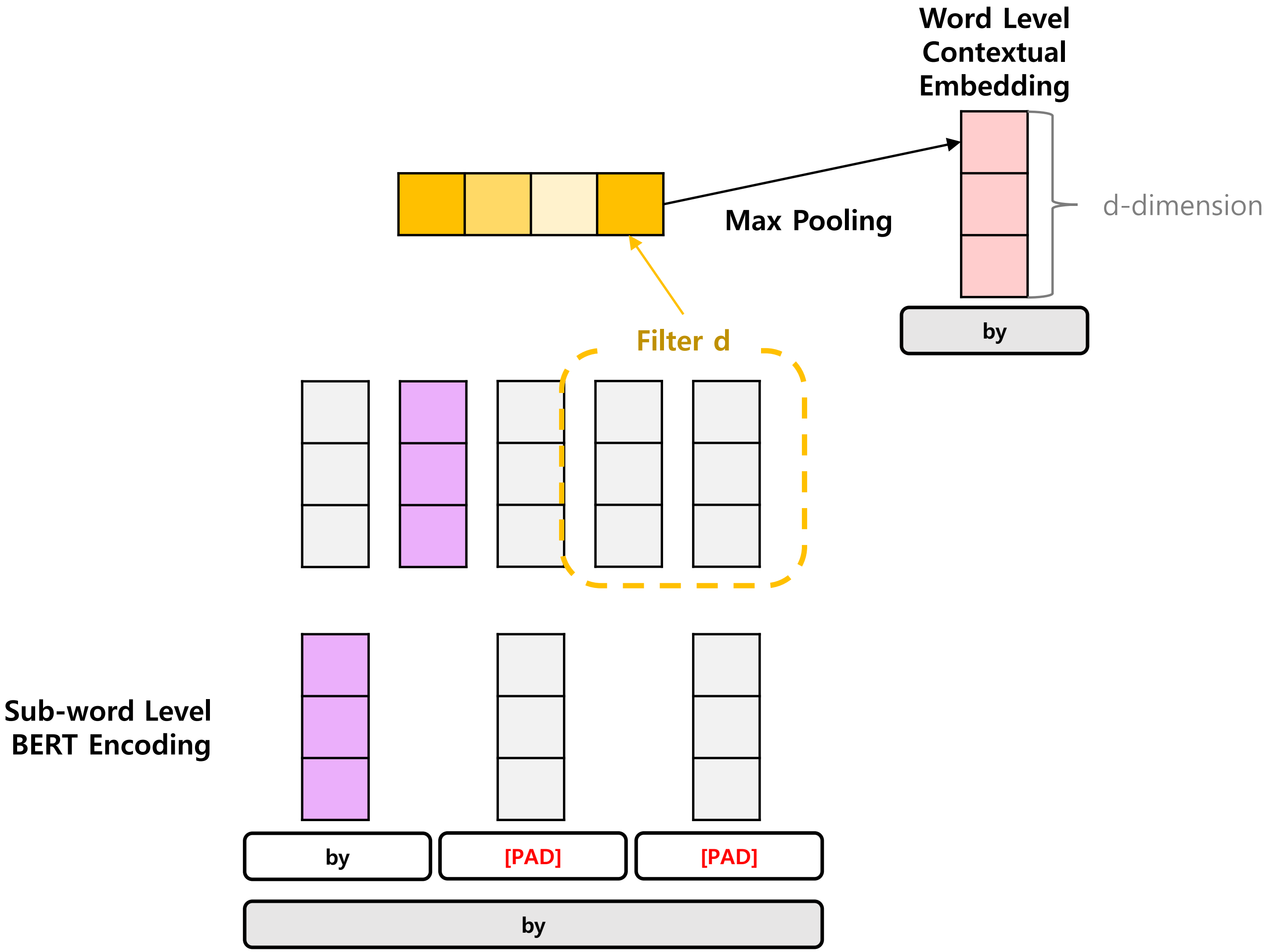
<1D Convolution>



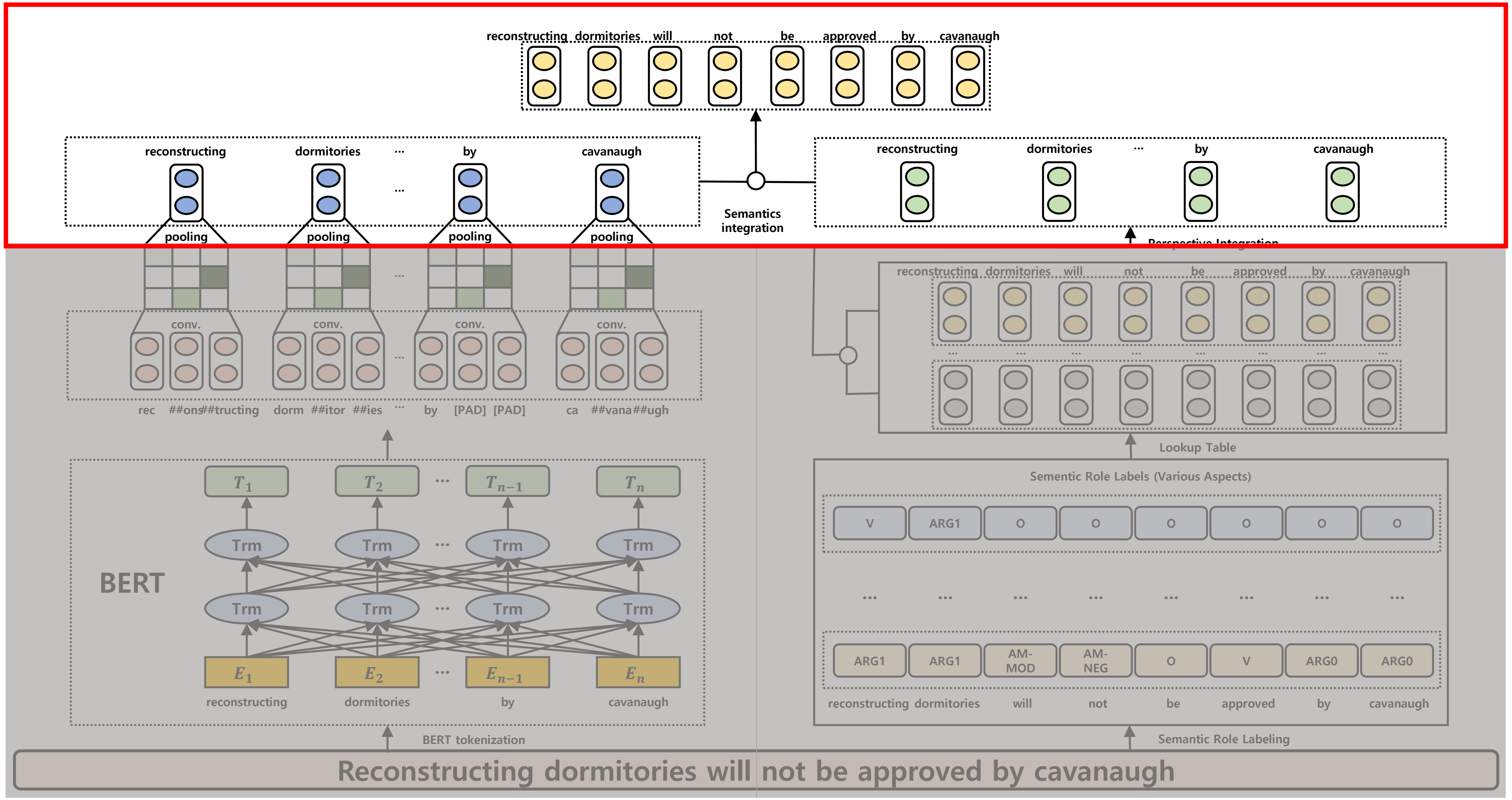
<1D Convolution>



<1D Convolution>

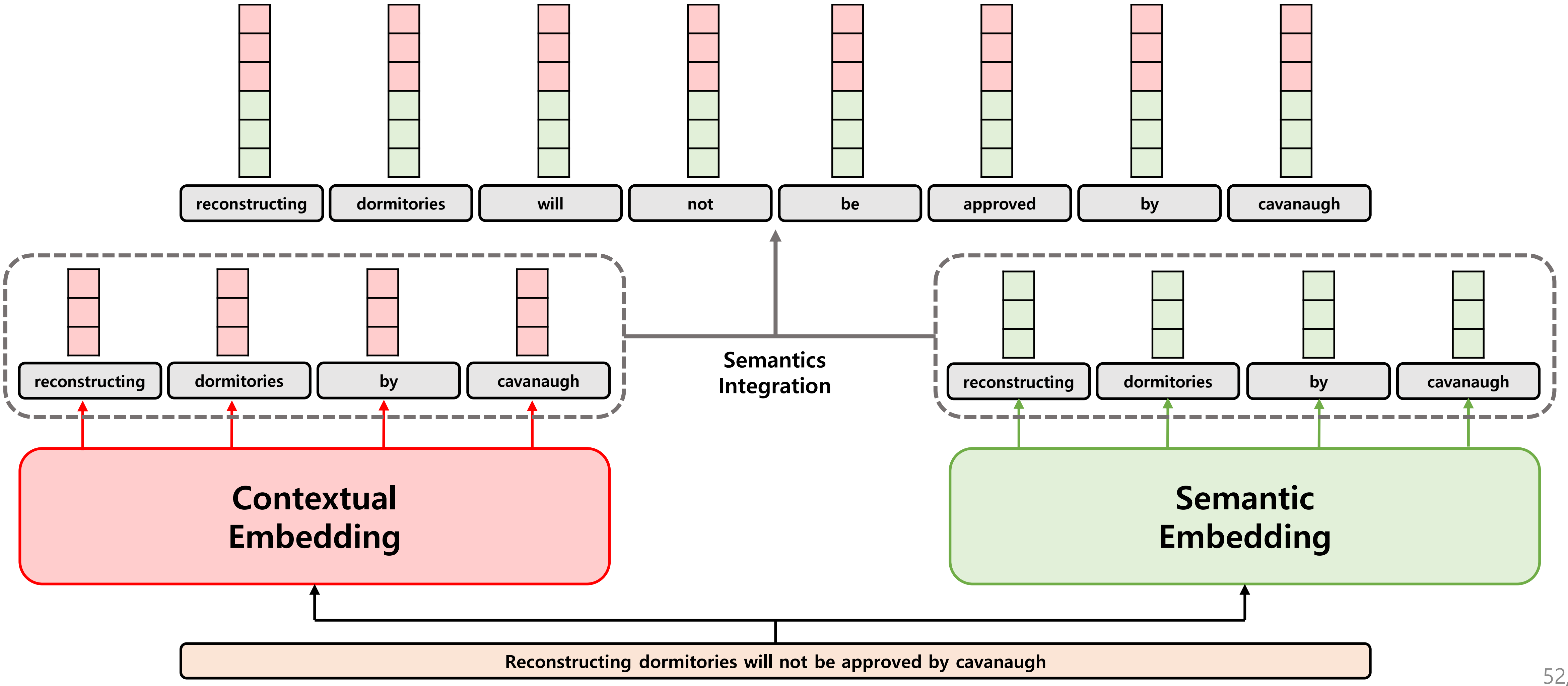


<Overall Architecture>

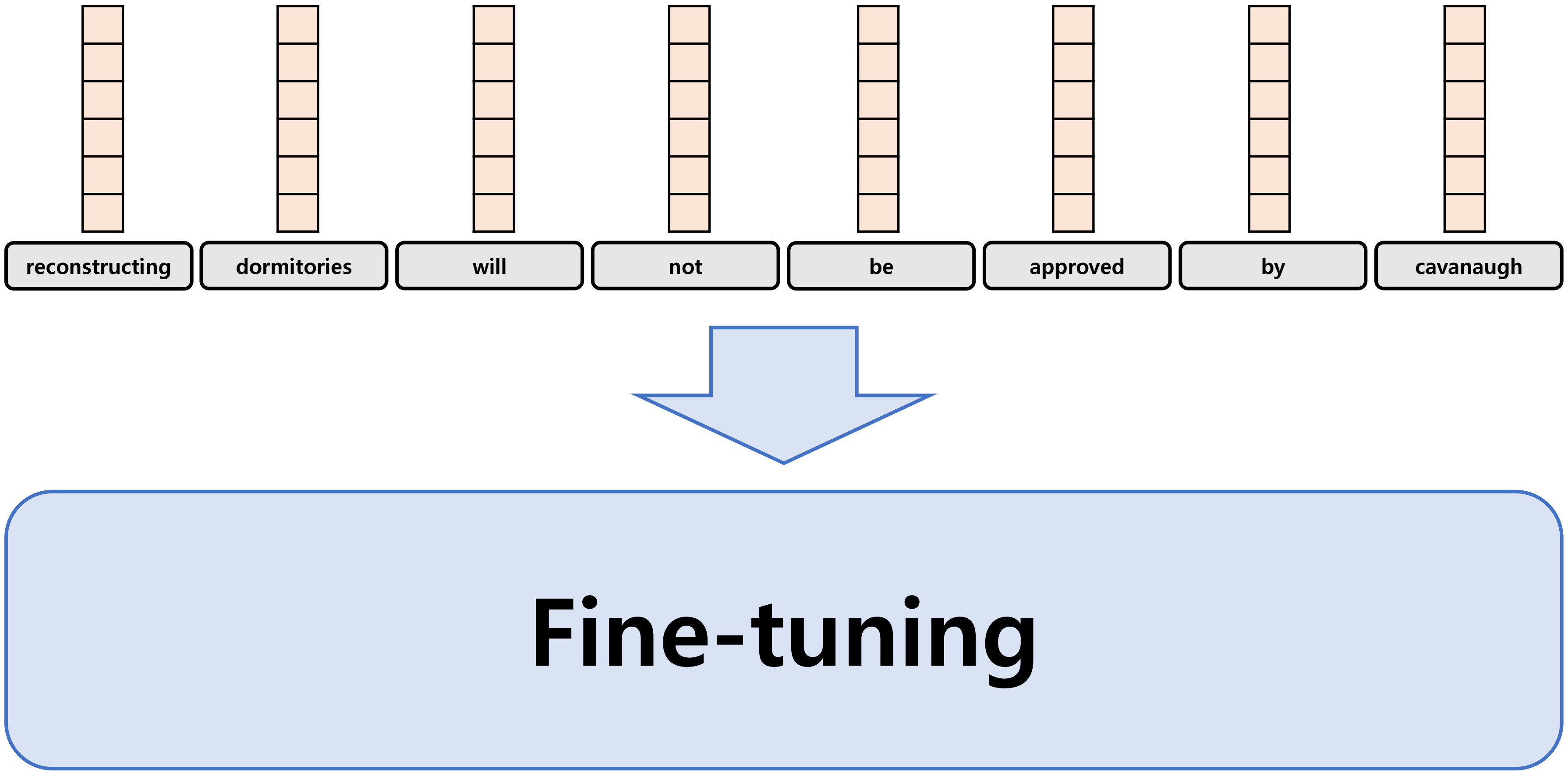


2 SemBERT
-Semantics Integration

<Semantics Integration>



<Semantics Integration>



3. Experiments

- GLUE
- SQuAD 2.0
- SNLI
- Parameters
- Ablation Study

3 Experiments
-GLUE Dataset

<GLUE>

Method	Classification		Natural Language Inference			Semantic Similarity			Score
	CoLA	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STS-B	-
	(mc)	(acc)	m/mm(acc)	(acc)	(acc)	(F1)	(F1)	(pc)	-
				Leaderborad (September, 2019)					
ALBERT	69.1	97.1	91.3/91.0	99.2	89.2	93.4	74.2	92.5	89.4
RoBERTa	67.8	96.7	90.8/90.2	98.9	88.2	92.1	90.2	92.2	88.5
XLNET	67.8	96.8	90.2/89.8	98.6	86.3	93.0	90.3	91.6	88.4
				In literature (April, 2019)					
BiLSTM+ELMo+Attn	36.0	90.4	76.4/76.1	79.9	56.8	84.9	64.8	75.1	70.5
GPT	45.4	91.3	82.1/81.4	88.1	56.0	82.3	70.3	82.0	72.8
GPT on STILTs	47.2	93.1	80.8/80.6	87.2	69.1	87.7	70.1	85.3	76.9
MT-DNN	61.5	95.6	86.7/86.0	-	75.5	90.0	72.4	88.3	82.2
BERTBASE	52.1	93.5	84.6/83.4	-	66.4	88.9	71.2	87.1	78.3
BERTLARGE	60.5	94.9	86.7/85.9	92.7	70.1	89.3	72.1	87.6	80.5
				Our implementation					
SemBERTBASE	57.8	93.5	84.4/84.0	90.9	69.3	88.2	71.8	87.3	80.9
SemBERTLARGE	62.3	94.9	87.6/86.3	94.6	84.5	91.2	72.8	87.8	82.9

<GLUE>

Method	Classification		Natural Language Inference			Semantic Similarity			Score
	CoLA	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STS-B	-
	(mc)	(acc)	m/mm(acc)	(acc)	(acc)	(F1)	(F1)	(pc)	-
			Leaderborad (September, 2019)						
ALBERT	69.1	97.1	91.3/91.0	99.2	89.2	93.4	74.2	92.5	89.4
RoBERTa	67.8	96.7	90.8/90.2	98.9	88.2	92.1	90.2	92.2	88.5
XLNET	67.8	96.8	90.2/89.8	98.6	86.3	93.0	90.3	91.6	88.4
			In literature (April, 2019)						
BiLSTM+ELMo+Attn	36.0	90.4	76.4/76.1	79.9	56.8	84.9	64.8	75.1	70.5
GPT	45.4	91.3	82.1/81.4	88.1	56.0	82.3	70.3	82.0	72.8
GPT on STILTs	47.2	93.1	80.8/80.6	87.2	69.1	87.7	70.1	85.3	76.9
MT-DNN	61.5	95.6	86.7/86.0	-	75.5	90.0	72.4	88.3	82.2
BERTBASE	52.1	93.5	84.6/83.4	-	66.4	88.9	71.2	87.1	78.3
BERTLARGE	60.5	94.9	86.7/85.9	92.7	70.1	89.3	72.1	87.6	80.5
			Our implementation						
SemBERTBASE	57.8	93.5	84.4/84.0	90.9	69.3	88.2	71.8	87.3	80.9
SemBERTLARGE	62.3	94.9	87.6/86.3	94.6	84.5	91.2	72.8	87.8	82.9

<SQuAD 2.0>

Model	EM	F1
#1 BERT + DAE + AoA	85.9	88.6
#2 SG-Net	85.2	87.9
#3 BERT + NGM + SST	85.2	87.7
U-Net (Sun et al. 2018)	69.2	72.6
RMR + ELMo + Verifier (Hu et al. 2018)	71.7	74.2
Our implementation		
$\textbf{BERT}_{\textbf{LARGE}}$	80.5	83.6
$\textbf{SemBERT}_{\textbf{LARGE}}$	82.4	85.2
$\textbf{SemBERT*}_{\textbf{LARGE}}$	84.8	87.9

<SNLI>

Model	Dev	Test
In literature		
DRCN (Kim et al. 2018)	-	90.1
SJRC (Zhang et al. 2019)	-	91.3
MT-DNN (Liu et al. 2019)	92.2	91.6
Our implementation		
BERT$_{\text{BASE}}$	90.8	90.7
SemBERT$_{\text{BASE}}$	91.2	91.0
BERT$_{\text{LARGE}}$	91.3	91.1
SemBERT$_{\text{LARGE}}$	92	91.6
BERT$_{\text{WWM}}$	92.1	91.6
SemBERT$_{\text{WWM}}$	92.2	91.9

<Parameters>

Model	Params (M)	Shared (M)	Rate (M)
MT-DNN	3,060	340	9.1
BERT on STILT	335	-	1.0
BERT	335	-	1.0
SemBERT	340	-	1.0

<Ablation Study>

Model	SNLI	SQuAD 2.0	
	Dev	EM	F1
BERT$_{LARGE}$	91.3	79.6	82.4
BERT$_{LARGE}$ + SRL	91.5	80.3	83.1
SemBERT$_{LARGE}$	92.3	80.9	83.6

<Ablation Study>

Number	1	2	3	4	5
Accuracy	91.49	91.36	91.57	91.29	91.42

<Max Number of Predicate-argument Structure m>

Proportion	0%	20%	40%
SQuAD 2.0 F1	87.93	87.31	87.24

<Turning Labels Proportion>

4. Discussion

- **Does SemBERT Understand Semantics?**

<Now, SemBERT Understands>

Yesterday,	Kristina	hit	Scott	with a baseball bat
AM-TMP	ARG0	V	ARG1	AM-INS
Temporal	Agent	Predicate	Object	Instrument

- ✓ Scott was hit by Kristina yesterday with a baseball bat
- ✓ Yesterday, Scott was hit with a baseball bat by Kristina
- ✓ With a baseball bat, Kristina hit Scott yesterday
- ✓ Yesterday Scott was hit by Kristina with a baseball bat
- ✓ Kristina hit Scott with a baseball bat yesterday

<Now, SemBERT Understands>

✓ How are you?

✓ How old are you?

✓ What is your age?

<Now, SemBERT Understands>

✓ How are you?

R-ARG0 V ARG0

✓ How old are you?

R-ARG0 V ARG0

✓ What is your age?

R-ARG0 V ARG0

< Steps of Natural Language Processing >

Lexical Analysis

↓

Syntax Analysis

↓

Semantic Analysis

↓

Pragmatic Analysis

Q & A

Thank You