Bi-directional machine translation for conversational interpreter

(Neural machine translation)

2023.12.14
Presenter: Seonhui, Kim

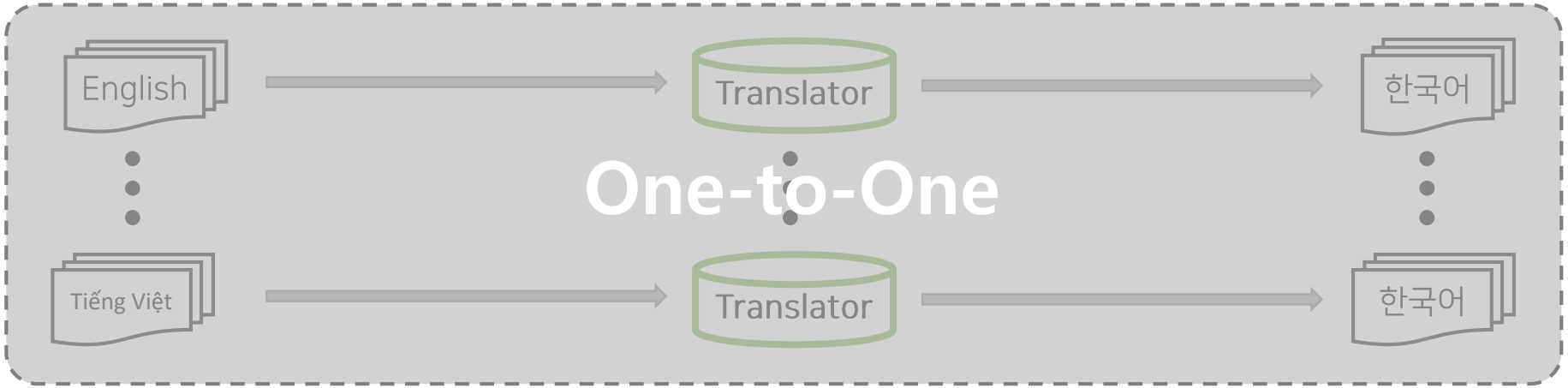
Contents

Previous

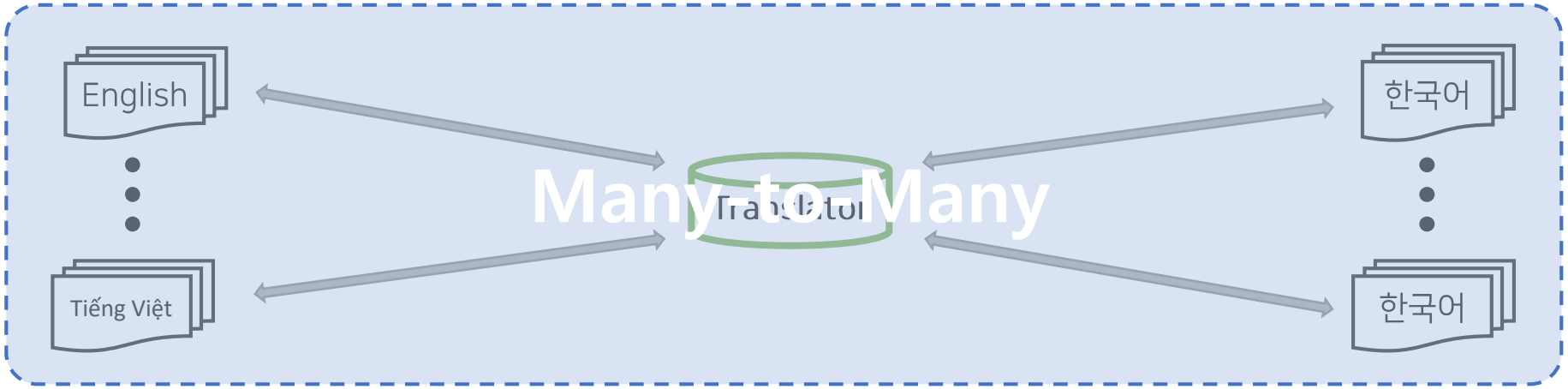
01. Previous

Conversational NMT model applicable in real-life scenarios – Multiligual NMT

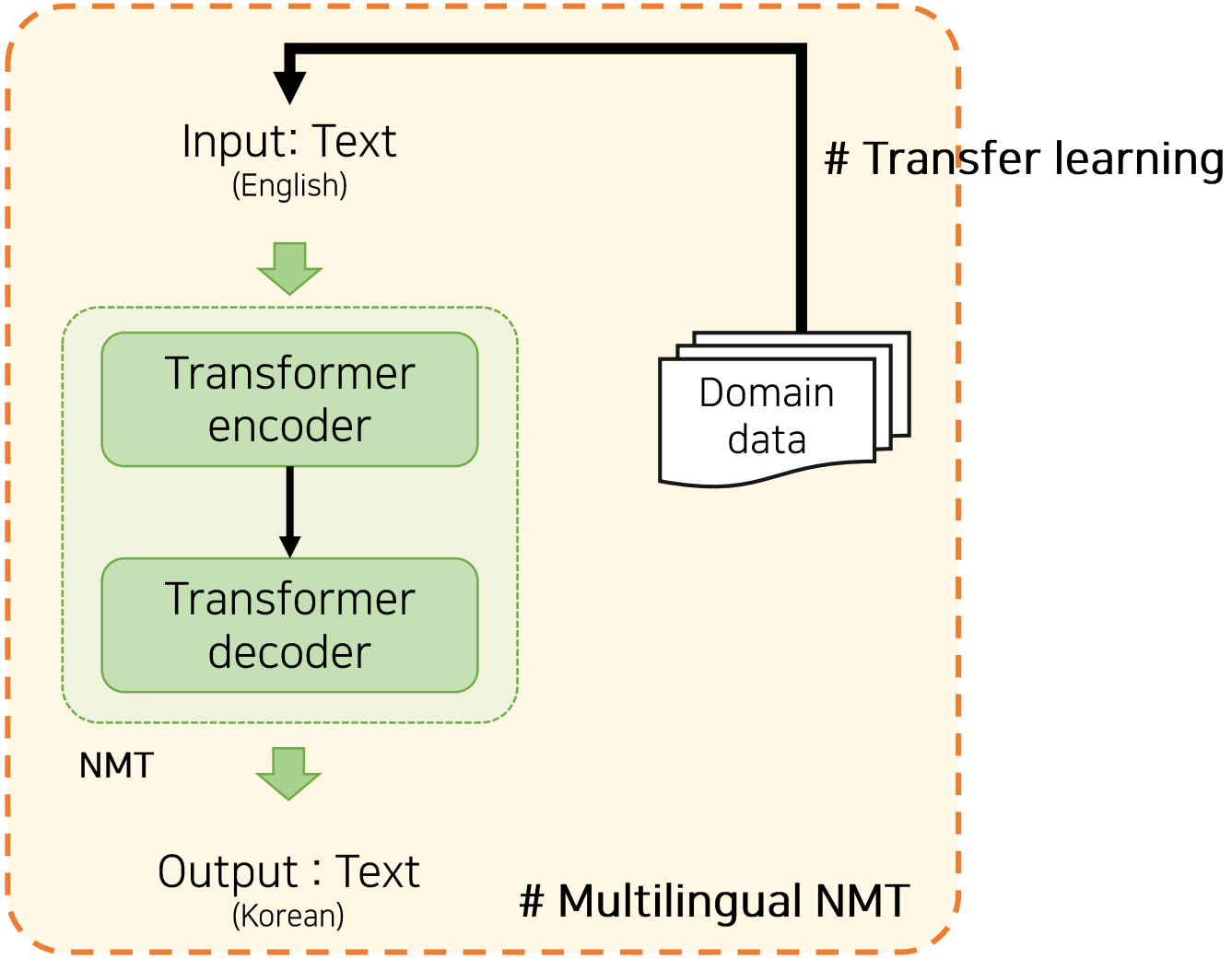
Monolingual



Multilingual



Conversational NMT model applicable in real-life scenarios – Adaptation domain



Utilizing Transparent Display for an Interpreter:

✓ Requires a Multiligual Language Model(Translator)

Bi-directional NMT Experiences & Results

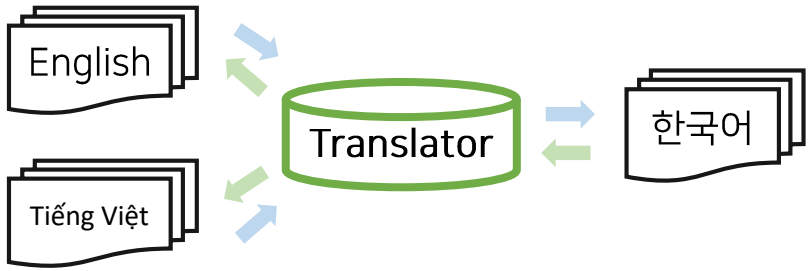
Explain train directions

Uni-Directional

- ✓ Source language: Korean
- ✓ Target language: English

If two parallel corpus exist,
two models are generated.

One-to-Many, Many-to-One



One-to-Many

- ✓ Source language: Korean
- ✓ Target language: English, Vietnam

Many-to-One

- ✓ Source language: English, Vietnam
- ✓ Target language: Korean

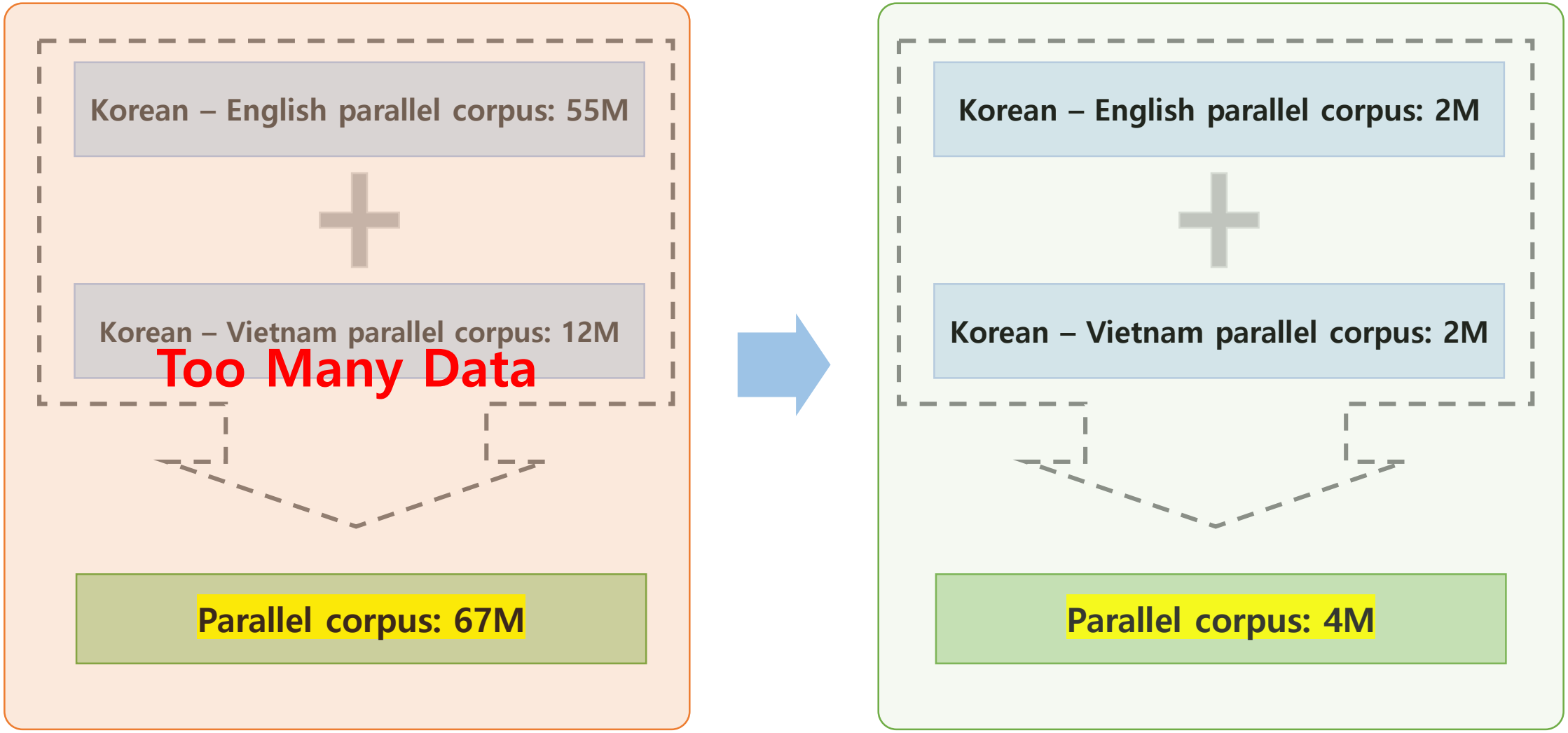
If two parallel corpus exist,
two models are generated.

Bi-Directional

- ✓ Source language: Korean, English
- ✓ Target language: English, Korean

If two parallel corpus exist,
One models are generated.

Parallel corpus for Machine Translation



Korean-English, Korean-Vietnam Machine Translation

Data collection: Conversational parallel corpus configuration

- Korean - English

Language	Number of Sentences	Number of words
Korean	2M	1.6M
English	2M	16M

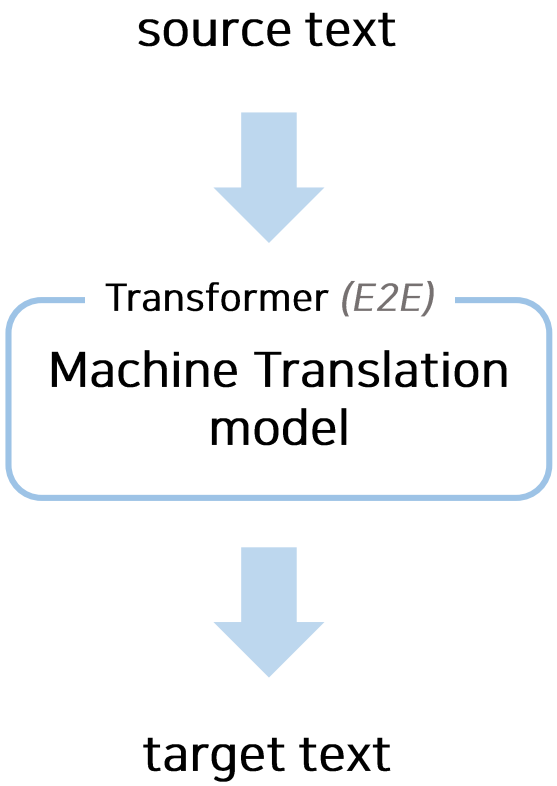
- Korean - Vietnam

Language	Number of Sentences	Number of words
Korean	2M	1.5M
Vietnam	2M	17M

Korean-English, Korean-Vietnam Machine Translation

Experience configuration & Results

Configuration	Uni-Directional	Results
Attention Block: 6 Attention Dim: 512 Attention Head: 8 FF layers: 2048 Token size: 10,000 Token joint: False	Korean → English	11.2
	English → Korean	18.4
	Korean → Vietnam	17.9
	Vietnam → Korean	23.1



Evaluation matrix

- BLEU score

One-to-Many, Many-to-One Machine Translation

Data collection: Conversational parallel corpus configuration

- Korean - English

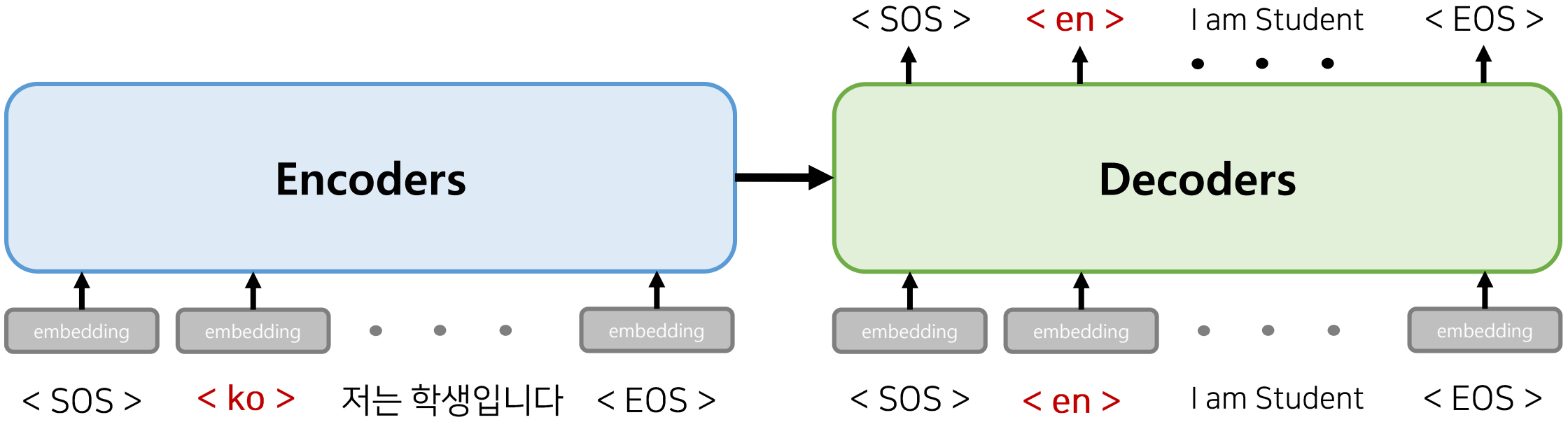
Language	Number of Sentences	Number of words
Korean	2M	1.6M
English	2M	16M

- Korean - Vietnam

Language	Number of Sentences	Number of words
Korean	2M	1.5M
Vietnam	2M	17M

One-to-Many, Many-to-One Machine Translation

Data collection: Input language tag



One-to-Many

- ✓ Source language: Korean
- ✓ Target language: English, Vietnam

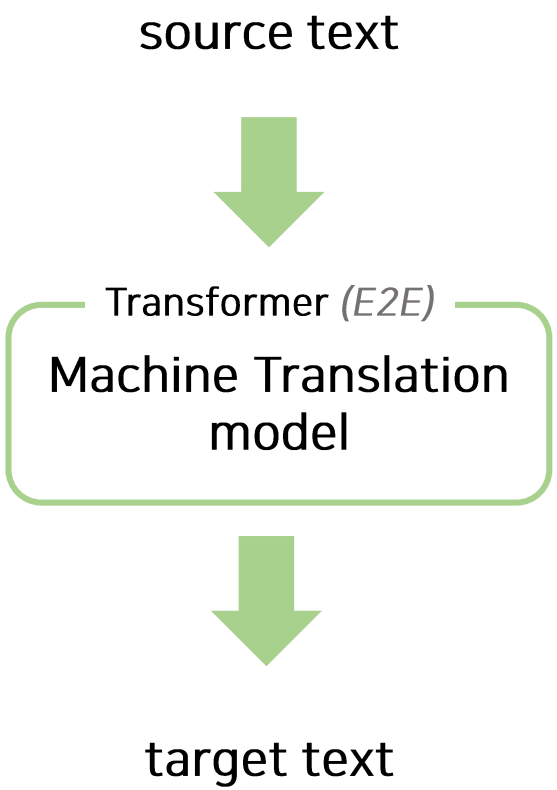
Many-to-One

- ✓ Source language: English, Vietnam
- ✓ Target language: Korean

One-to-Many, Many-to-One Machine Translation

Experience configuration & Results

Configuration	Direction	Results
Attention Block: 6 Attention Dim: 512 Attention Head: 8 FF layers: 2048 Token size: 10,000 Token joint: False	Many-to-One (English→Korean)	12.3 / 11.2
	Many-to-One (Vietnam→Korean)	19.1 / 18.4
	One-to-Many (Korean → English)	17.95 / 17.90
	One-to-Many (Korean→Vietnam)	23.50 / 23.1



Evaluation matrix

- BLEU score

Bi-Directional Machine Translation

Data collection: Conversational parallel corpus configuration

- Korean - English

Language	Number of Sentences	Number of words
Korean	2M	1.6M
English	2M	16M

Common experience configuration

Hyperparameter
Attention Block: 6
Attention Dim: 512
Attention Head: 8
FF layers: 2048

Bi-Directional

- ✓ Source language: Korean, English
- ✓ Target language: English, Korean

Bi-Directional Machine Translation

Detail experience configuration & Results

Model	Token Joint	Token size	Enc-Dec	Epoch	Result
Bi-Directional	False	5,000	False	50	16.3
				100	17.4
			True	50	16.3
				100	17.5
Korean→English	False	5,000	False	100	18.4
			True	100	18.2

Model	Token Joint	Token size	Enc-Dec	Epoch	Result
Bi-Directional	True	10,000	False	50	16.3
				100	18
			True	50	16.9
				100	18
Korean→English	True	10,000	False	100	18.2
			True	100	17.5

Bi-Directional Machine Translation

Compare Results

Model	Token joint	Token size	Enc-Dec	BLEU
Korean→English	False	5,000	False	18.4 Best!!
Bi-Directional				17.4
Korean→English	False	5,000	True	18.2
Bi-Directional				17.5

Model	Token joint	Token size	Enc-Dec	BLEU
Korean→English	True	10,000	False	18.2
Bi-Directional				18
Korean→English	True	10,000	True	17.5
Bi-Directional				18

Future work

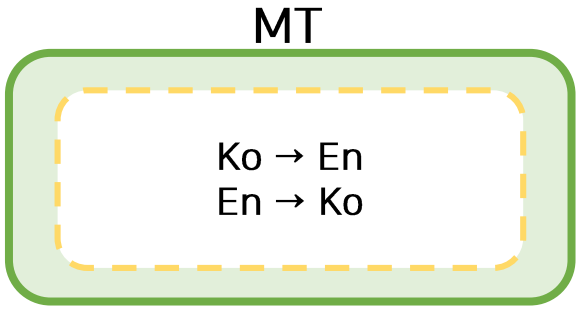
04. Bi-directional

Bi-Directional Machine Translation

Future Training

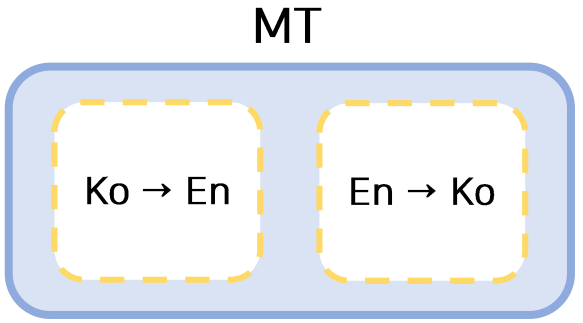
1 Epoch

<Ko> 저는 학생입니다. → <En> I am student.
<En> It's a bit chilly this evening → <Ko> 오늘 저녁은 꽤 쌀쌀합니다.
⋮
Mix data
<Ko> 밥을 많이 먹었습니다. → <En> I ate a lot of food.



1 Epoch

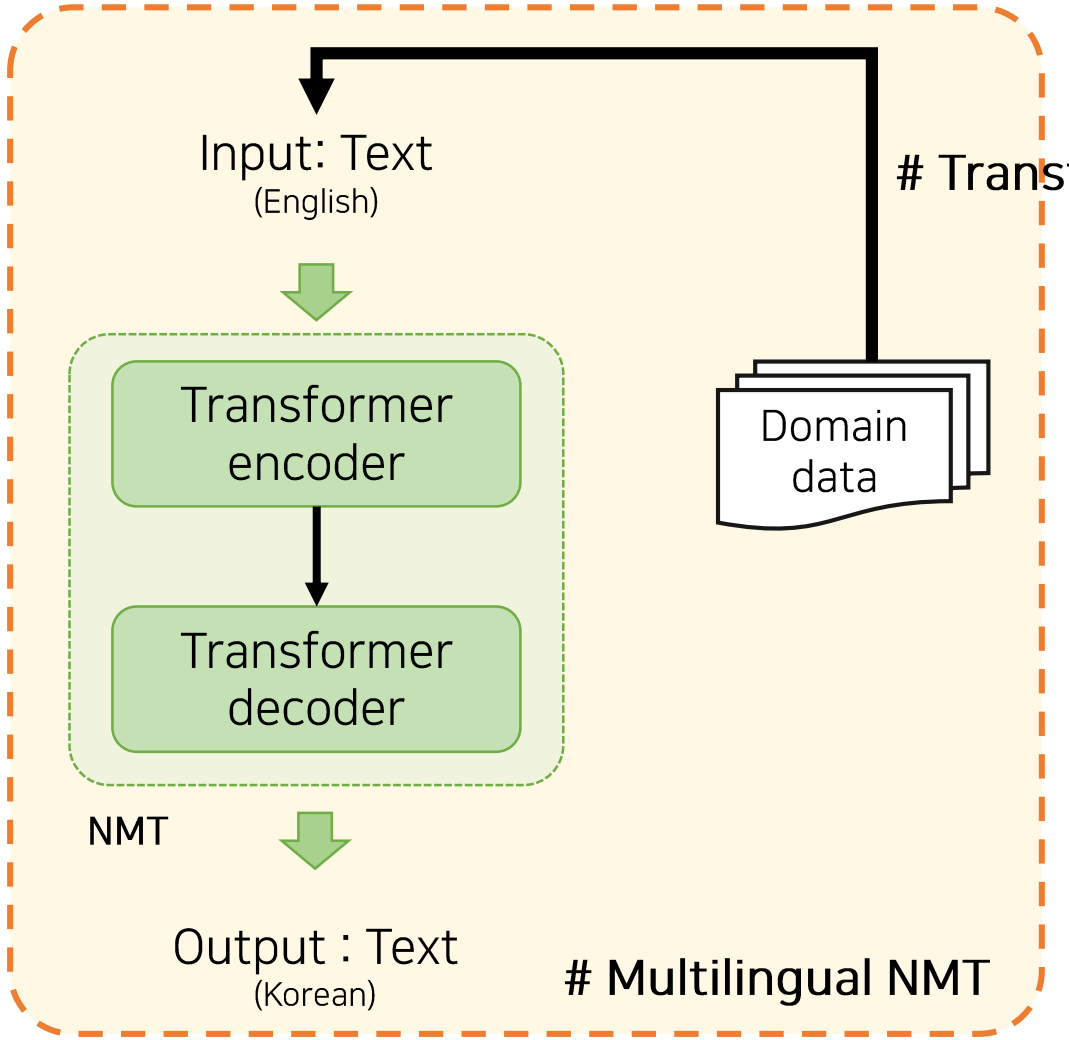
<Ko> 저는 학생입니다. → <En> I am student.
<Ko> 밥을 많이 먹었습니다. → <En> I ate a lot of food.
⋮
Separate data
<En> It's a bit chilly this evening → <Ko> 오늘 저녁은 꽤 쌀쌀합니다.



Domain Tuning Experiences & Results

05. Domain Tuning

Tuning Model



Pre-trained model
- Korean -> English 55M
- English -> Korean 55M

Domain: Metro Service
- Confrontation with a subway conductor
  - ➢ My transit card isn't working, can you help?
- Metro map information
  - ➢ Seoul station, Gwanghwamun station
  - ➢ Ticket Vending Machines, T-money

Korean – English Machine Translation Domain Tuning

Data collection: Metro information parallel corpus

Data	Name	Amount	Detail
Station list	Station	777	역/Station 'O'
Station list	Station_org	777	역/Station 'X'
Metro Q&A list	QA_list	215	

- Example

Station
✓ 광화문역 - Gwanghwamun Station
✓ 홍대역 - Hongdae Station
✓ 명동역 - Myeongdong Station

Station_org
✓ 광화문 - Gwanghwamun
✓ 홍대 - Hongdae
✓ 명동 - Myeongdong

QA_list
✓ 게이트에 오류가 있었던 것 같습니다. - I think there's a problem with the gate.
✓ 이 카드를 다시 사용할 수 있습니까? - Can I use this card again?

Korean – English Machine Translation Domain Tuning

Data collection: Eval set

- Station eval set
  - Station Information
  - Use tuning data
  - 777 parallel corpus
- Station_org eval set
  - Station Information
  - Use tuning data
  - 777 parallel corpus
- QA_list eval set
  - Metro service Q&A list
  - Use tuning data
  - 216 parallel corpus

- Toeic eval set
  - Toeic test
  - Not use tuning data
  - 1,607 parallel corpus
- Drama eval set
  - Drama caption list
  - Not use tuning data
  - 5,630 parallel corpus

Korean – English Machine Translation Domain Tuning

Metro data tuning results

Station			
	Before tuning	Tuning 50epoch	Tuning 100epoch
Korean -> English	59.4	66.9	73.6
English -> Korean	71.1	97.9	99.7

Station_org			
	Before tuning	Tuning 50epoch	Tuning 100epoch
Korean -> English	34	44.1	55.7
English -> Korean	60	98.1	99.4

QA_list			
	Before tuning	Tuning 50epoch	Tuning 100epoch
Korean -> English	35.6	36.1	40.2
English -> Korean	24.9	95	98

Korean – English Machine Translation Domain Tuning

Metro data tuning results

Toeic			
	Before tuning	Tuning 50epoch	Tuning 100epoch
Korean -> English	21.5	21.6	21.4
English -> Korean	17.1	14.7	14.6

Drama			
	Before tuning	Tuning 50epoch	Tuning 100epoch
Korean -> English	24.8	24.7	24.3
English -> Korean	10.8	9.4	9.4

Korean – English Machine Translation Domain Tuning

Metro data tuning results

Station eval set

Korean_reference	English_reference	Before tuning	After tuning
당정역	Dangjeong Station	Tangjeong Station	Dangjeong Station
천안역	Cheonan Station	Cheonan Station	Cheonan Station

Station_org eval set

Korean_reference	English_reference	Before tuning	After tuning
신사	Gentlemen	Gentlemen.	Gentlemen
신논현	Sinnonhyeon	Sinnonhyeon	Sinnonhyeon

Side effects still exist.

Q & A

2023.11.09
Presenter: Seonhui, Kim