# Conversational Korean-Vietnam translator for interpreters

(Neural machine translation)

2023.11.09
Presenter: Seonhui, Kim

# Contents

01. Introduction

02. Related works

03. Experience

04. Results

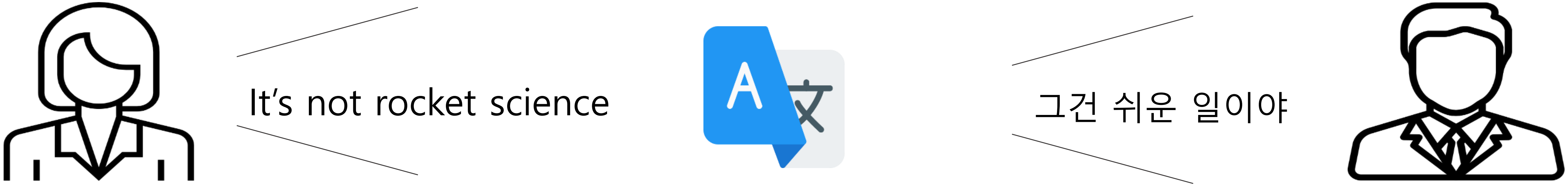
05. Future work

# Introduction

# 01. Introduction

## Conversational

- **Improvement of translation quality**
  - ✓ Translating natural expression of foreigners into appropriate Korean for situation
  - ✓ Improving smooth communication in real-life interpreting situations



It’s not rocket science

그건 쉬운 일이야

# Related works

# 02. Related works

WMT22

DeepL

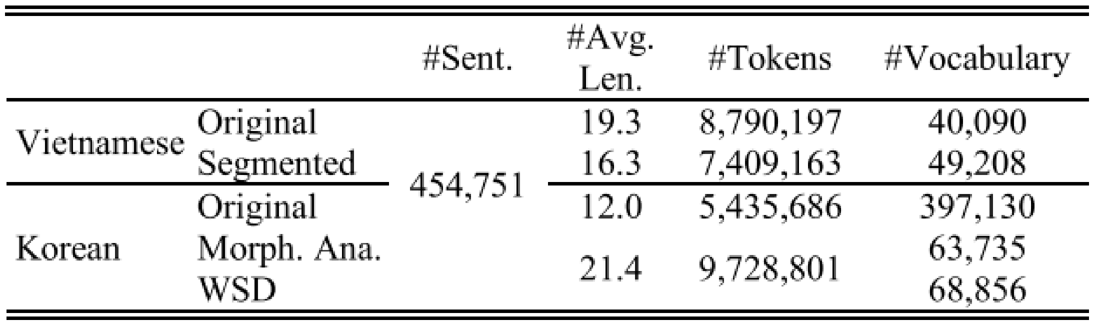
Seventh Conference on Machine Translation

Confirmed language pairs:

| | | |
|---|---|---|
| • English | - | Chinese |
| • English | - | Czech |
| • English | - | German |
| • English | - | Japanese |
| • English | - | Russian |
| • French | - | German |
| • Croatian | - | English |
| • Livonian | - | English |
| • Yakut | - | Russian |

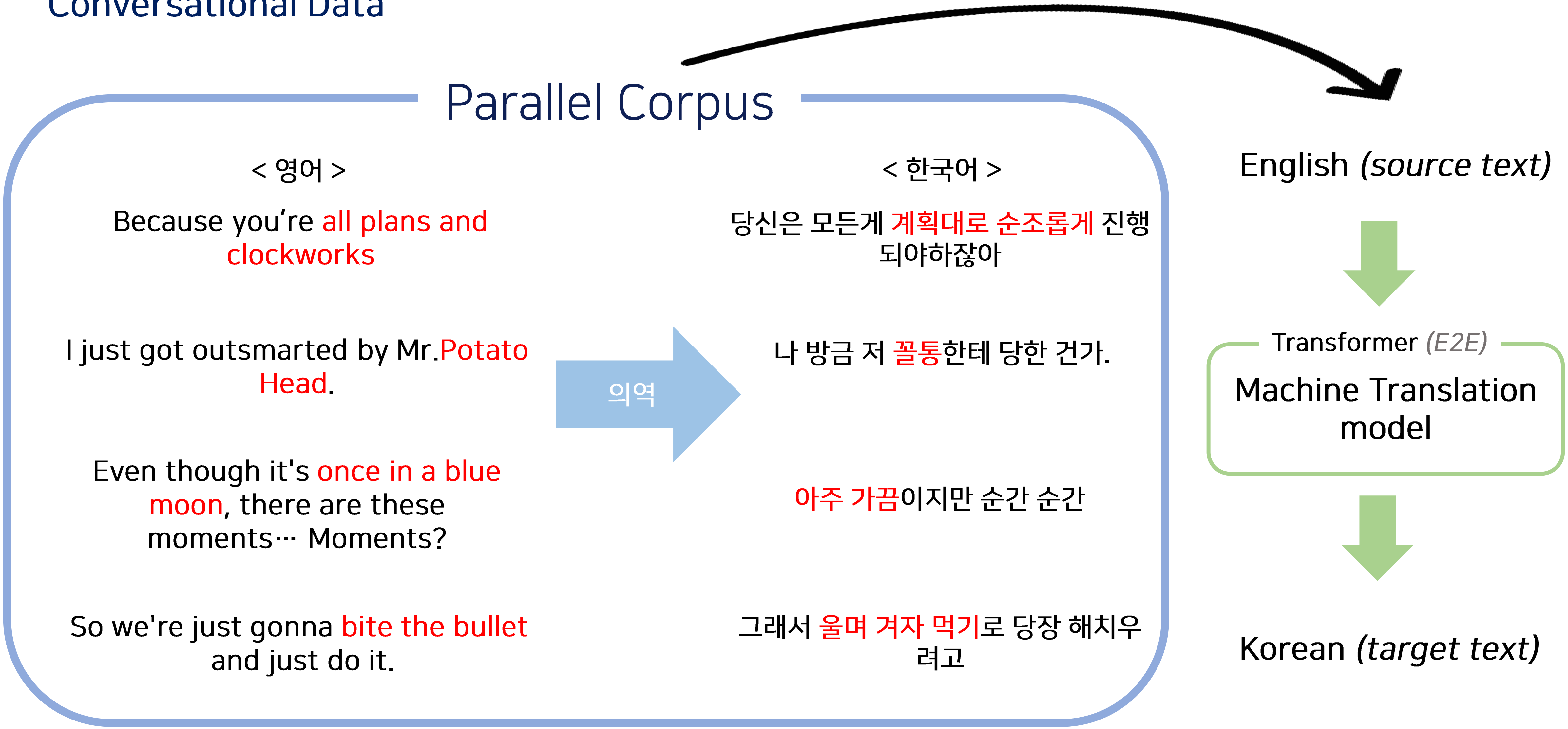
## Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation

**TABLE 11.** **Statistics of the Korean-Vietnamese parallel corpus.**



| | | #Sent. | #Avg. Len. | #Tokens | #Vocabulary |
|---|---|---|---|---|---|
| Vietnamese | Original | 454,751 | 19.3 | 8,790,197 | 40,090 |
| | Segmented | | 16.3 | 7,409,163 | 49,208 |
| Korean | Original | | 12.0 | 5,435,686 | 397,130 |
| | Morph. Ana. | | 21.4 | 9,728,801 | 63,735 |
| | WSD | | | | 68,856 |

## Conversational Data



### Parallel Corpus

| < 영어 > | < 한국어 > |
|---|---|
| Because you're all plans and clockworks | 당신은 모든게 계획대로 순조롭게 진행되야하잖아 |
| I just got outsmarted by Mr.Potato Head. | 나 방금 저 꼴통한테 당한 건가. |
| Even though it's once in a blue moon, there are these moments··· Moments? | 아주 가끔이지만 순간 순간 |
| So we're just gonna bite the bullet and just do it. | 그래서 울며 겨자 먹기로 당장 해치우려고 |

English *(source text)*

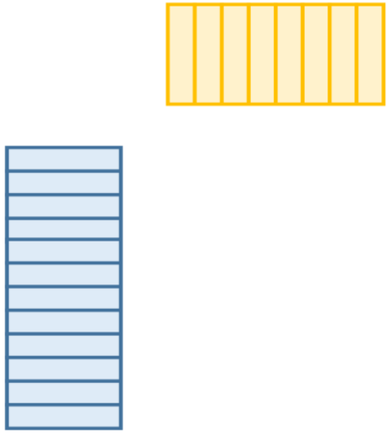
Transformer *(E2E)*

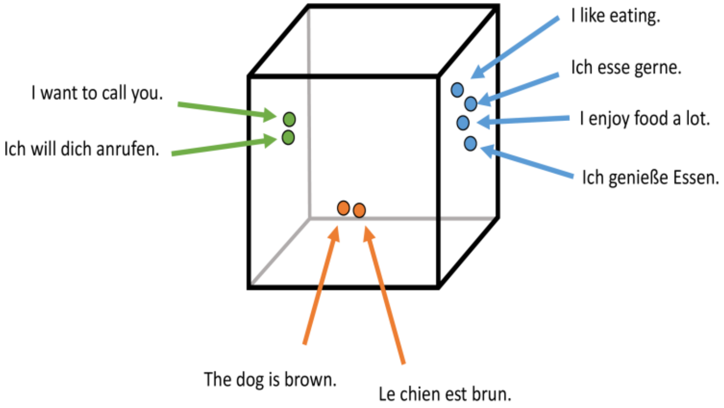
Machine Translation model

Korean *(target text)*

## Data Collection

Use Vecalign to make parallel corpus





## Data Collection

- Conversational parallel corpus configuration

| Language | Number of Sentences | Number of words |
|---|---|---|
| Korean | 11,878,865 | 49,532,476 |
| Vietnam | 11,878,865 | 96,670,925 |

- Conversational parallel corpus example

| Korean | Vietnam |
|---|---|
| 그런데 그걸 집에 와서 봤더니 갑자기 어떤 여자가 나와서. | Vậy mà khi anh ta bật lên để xem lại, thì nhìn thấy một người phụ nữ trong cuốn băng. |
| 그런데 지방이라 동경하고 채널이 다르잖아. | Nhưng không hề có kênh đó phát từ Tokyo. |
| 부인이 병원에 계속 있었어요. | Phải. Vợ ổng đã ở trong bệnh viện. |
| 사과하는 걸 수도 있잖아요 | Đó có thể là một lời xin lỗi theo như chúng ta biết. |

## Experience configuration

: Configure hyper parameters for each model size

### Base model

- Text Encoder/Decoder(MT):

  Transformer (6-block, **256-hidden, 4-head, noam-learning rate)**

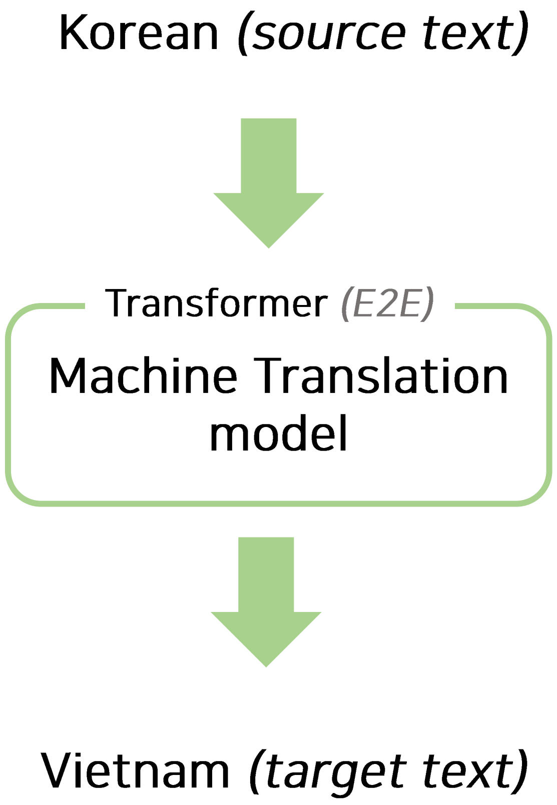
### Large model

- Text Encoder/Decoder(MT):

  Transformer (6-block, **1,024-hidden, 16-head, noam-learning rate)**

### Evaluation matrix

- BLEU score



Korean *(source text)*

Transformer *(E2E)*

Machine Translation model

Vietnam *(target text)*

## Experience configuration

### : Configuration for each experiment

| Experience | Model size | Data | Training data | Validation data |
|---|---|---|---|---|
| Exp1 | Base | Written | 2 million | 3,000 |
| Exp2 | Base | Conversational | 2 million | |
| Exp3 | Base | Conversational | 11 million | |
| Exp4 | Large | Conversational | 11 million | |

### : Evaluation data configuration

| Language | Number of Sentence | Number of words |
|---|---|---|
| Korean | 3,000 | 12,005 |
| Vietnam | 3,000 | 27,002 |

# Result

## Results 1

### : Pre-experimentation

- Quantitative analysis

More difficult

[BLEU Score]

| Model | Written eval set | Conversational eval set |
|---|---|---|
| Written model (2M) | 26.2 | 8.8 |
| Conversational model (2M) | 23.4 | 13.1 |

### : Performance comparison of conversational and written models

- Quantitative analysis

[BLEU Score]

| Model | Conversational eval set |
|---|---|
| Exp1: Written data | 7.13 |
| Exp2: Conversational data | **9.35** |

## Results 1

: Performance comparison of conversational and written models

- Quantitative analysis

[BLEU Score]

| Model | Conversational eval set |
|---|---|
| Exp1: Written data | 7.13 |
| Exp2: Conversational data | **9.35** |

- Static analysis

| Reference | Hypothesish of Exp1 | Hypothesish of Exp2 |
|---|---|---|
| Tay nắm cửa.<br>(The handle is …) | Bạn có tay cầm không? | Tay cầm… |
| tìm bà cô.<br>(I’m helping to find aunt) | Tôi đang giúp bạn tìm dì của tôi. | Tôi đang giúp cô tìm dì. |

# Results 2

## : Pre-experimentation

- Quantitative analysis – 2 million VS 11 million

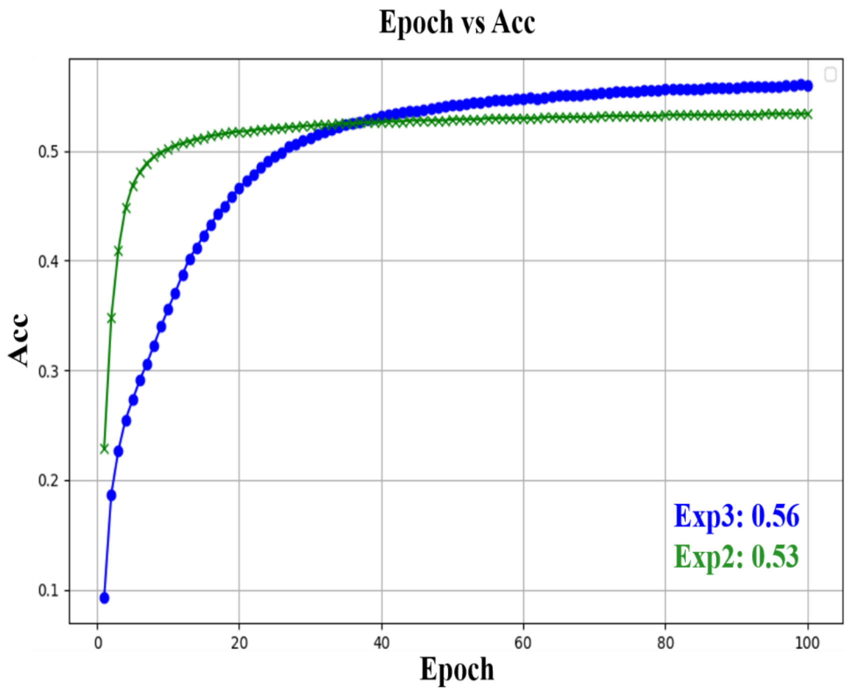
Just increasing the data did not make a difference about quality of translation.

## : Performance comparison of increase data

- Quantitative analysis

[BLEU Score]

| Model | Conversational eval set |
|---|---|
| Exp2: 2 million | 9.35 |
| Exp3: 12 million | **11.83** |



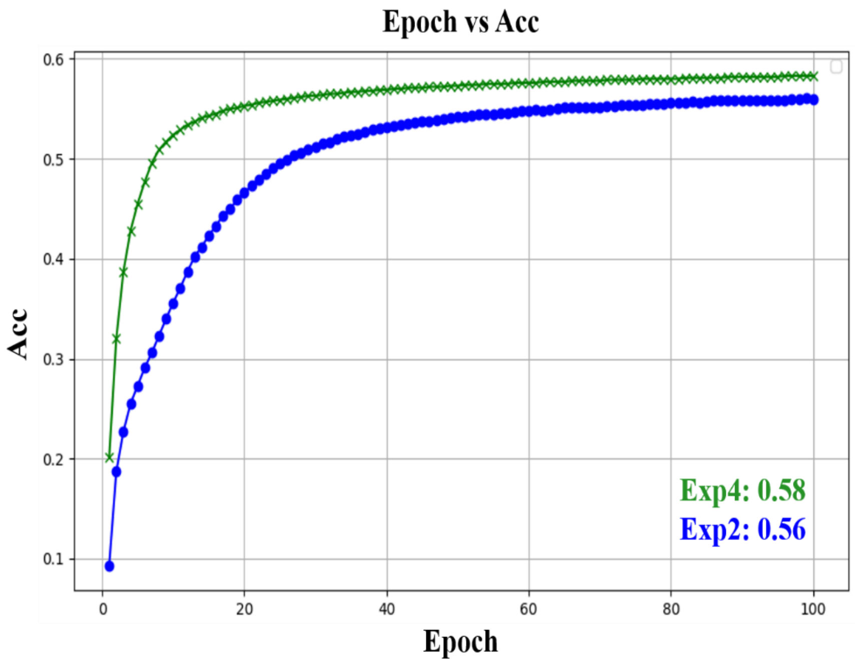
## Results 3

: Performance comparison of increase data & model size

- Quantitative analysis

[BLEU Score]

| Model | Conversational eval set |
|---|---|
| Exp2: 2 million, Base | 9.35 |
| Exp3: 12 million, Base | 11.83 |
| Exp4: 12 million, Large | **13.02** |



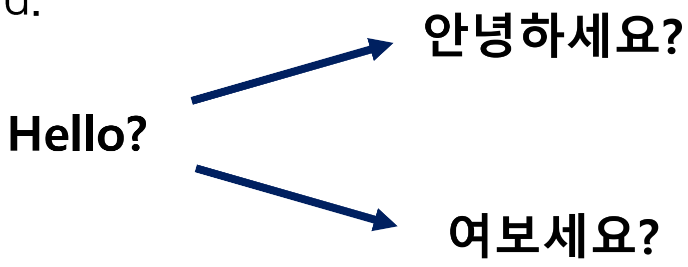
- Static analysis

| Reference | Hypothesis of Exp2 | Hypothesis of Exp4 |
|---|---|---|
| chính là nó. | Đây rồi. | Chính là nó. |
| Tôi đang ở đâu? | Chúng ta đang ở đâu? | Tôi đang ở đâu? |
| Đây không hẳn là kiếp sau. | Đây không phải là thế giới sau cái chết. | Đây không hẳn là thế giới sau khi chết. |

## Analysis

### : Feature of conversational data

- Conversational data is difficult to translate,
  because it contains paraphrases.
- Conversational data increases, the ambiguity of the translation increases.
- Compared to other translators, the result is a bit easier to understand.



### : Improving massively parallel corpus about MT

- Modified hyperparameters to be suitable for huge data.
- Reduce the likelihood of output
  - Learning contextual information from translation models.

# Side Experience

## Data configuration

: Korean – English machine translation

- Source language: English
- Target language: Korean

: Parallel corpus

- Training data: Conversational Data, 40 million
- Evaluating data: 4,000

| Model | Data info | Domain |
|---|---|---|
| Korean -> English | 38 million sentence | Conversational<br>(Drama, movie scripts) |
| | 10 million sentence | Written<br>(Trip, daily) |
| | 6.9 million sentence | Written<br>(Trip, daily) |

## Experience configuration

: Korean – English machine translation

- Text Encoder (MT):

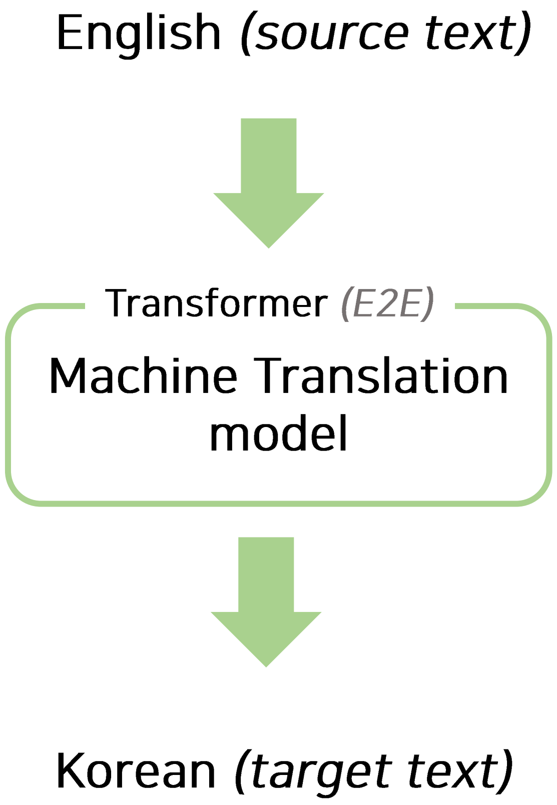
  Transformer (6-block, 512-hidden, 8-head, noam-optimizer)

- Text Decoder (MT):

  Transformer (6-block, 512-hidden, 8-head, noam-optimizer)

: Eval matrix

- DeepL, Papago
- BLEU score



# Result

## : My model vs DeepL vs Papago

- Quantitative analysis

**[BLEU Score]**

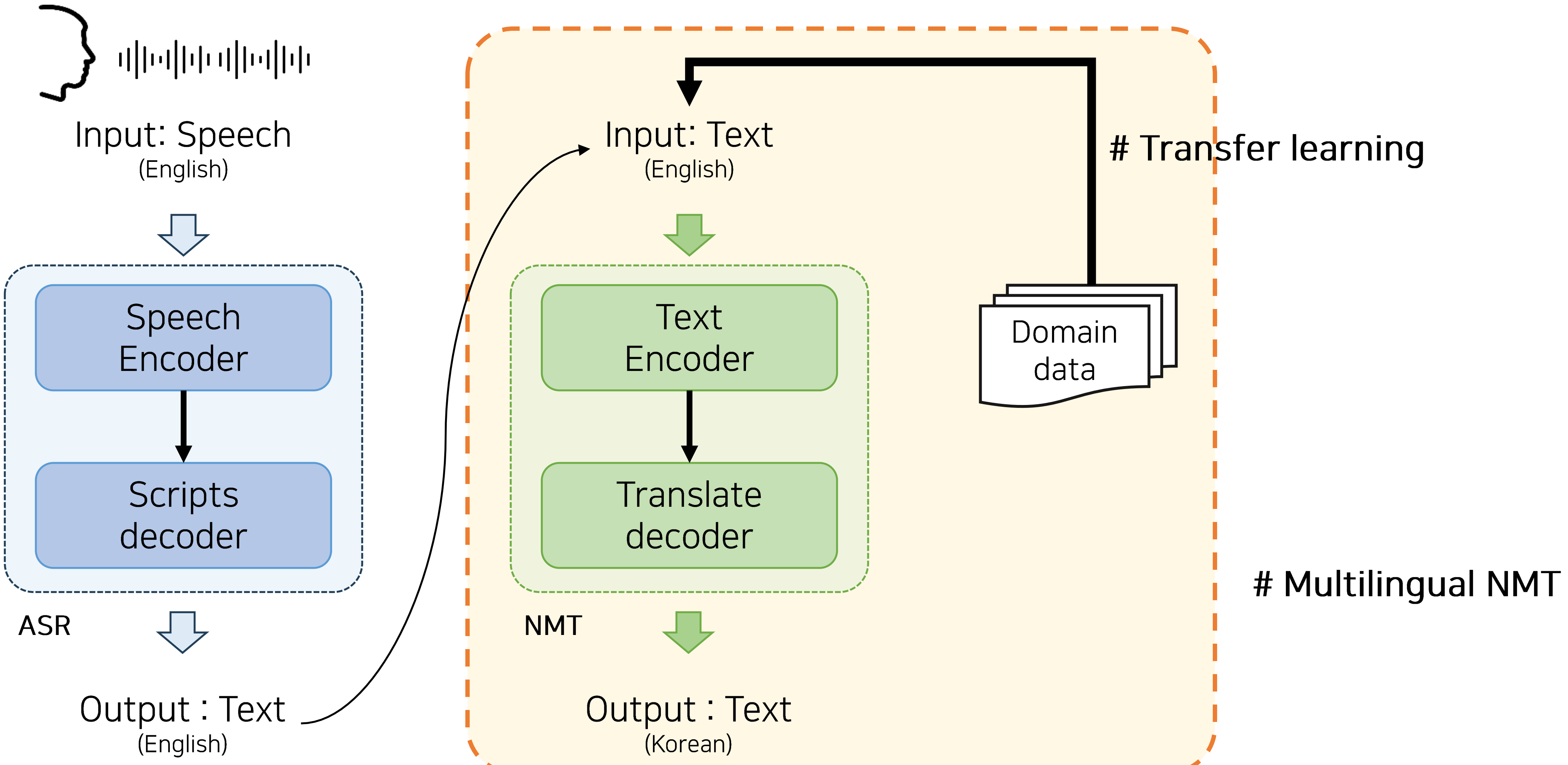
| Model | **Conversational eval set** | **Written eval set** |
|---|---|---|
| My model | 34.0 | 32.3 |
| DeepL | 36.6 | 28.6 |
| Papago | 32.6 | 33.9 |

**Conversational eval set: Papago < My model < DeepL**

**Written eval set: DeepL < My model < Papago**

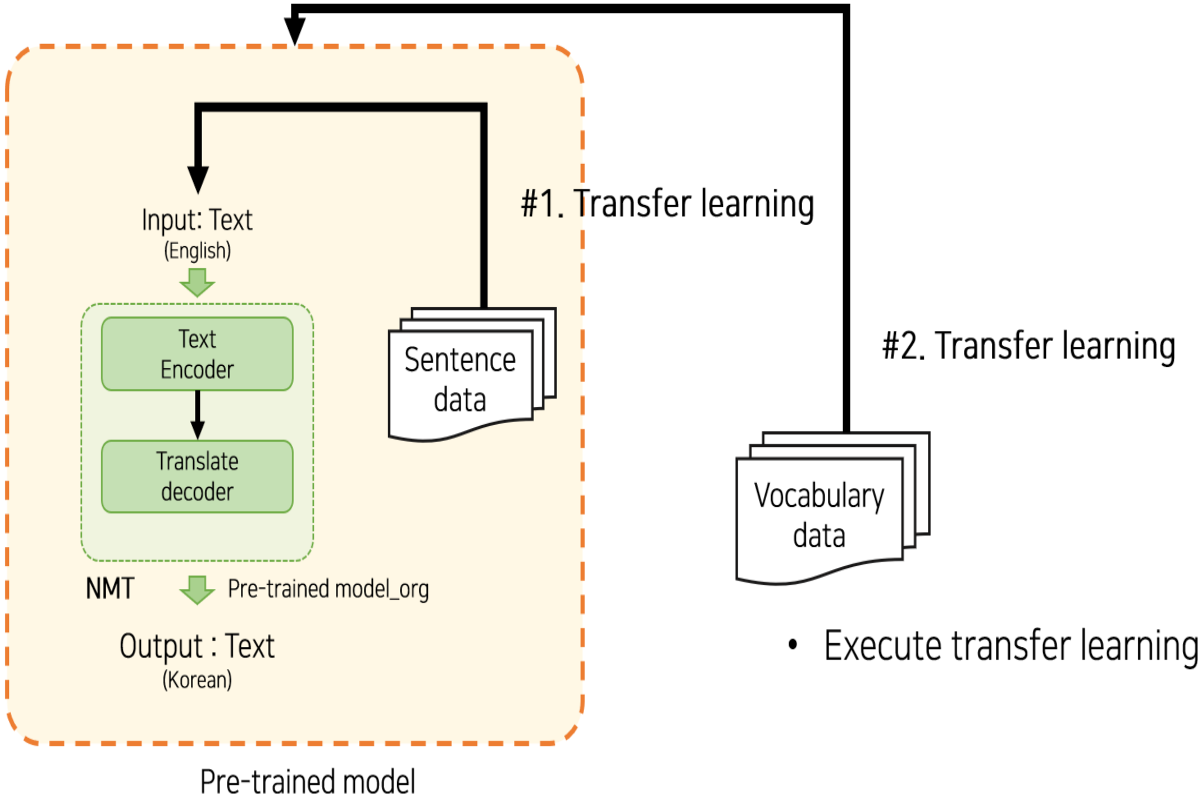
# Future work

# Conversational NMT model applicable in real-life scenarios- structure



# 03. Motivation

## Domain adaptation



- Execute transfer learning in two stages

Transfer learning

- Domain: conference (CV)
- Consider: copy-mechanism
- Use data: sentences & words

# Future work

# Q & A

2023.11.09
Presenter: Seonhui, Kim