Solving Korea Dialect Translation Problem Under Data Scarcity

2024.06.21

Ji Soo Lee | Bo Geon Park | Hyun Jin Lee | Jeong Min Moon
Korea University

About Our Topic

Jeju Dialect

- Perspective of considering as Jeju-language
  - Too unique to be considered as dialects of Korean
  - Separate language within the same language family with Korean
- Retain many medieval vocabulary terms from the creation of Hangul, having high academic value
- UNESCO
  - Classified as 'Definitely Endangered'

About Our Topic

Jeju Dialect

- Training
  - 01.원천데이터
    - TS_01. 충청도_01. 1인발화 따라말하기.zip | 15.97 GB | key: 437904 파일키 복사 다운로드
    - TS_01. 충청도_02. 1인발화 질문에답하기.zip | 27.93 GB | key: 437905 파일키 복사 다운로드
    - TS_01. 충청도_03. 2인발화.zip | 11.58 GB | key: 437906 파일키 복사 다운로드
    - TS_02. 전라도_01. 1인발화 따라말하기.zip | 19.38 GB | key: 437907 파일키 복사 다운로드
    - TS_02. 전라도_02. 1인발화 질문에답하기.zip | 34.94 GB | key: 437908 파일키 복사 다운로드
    - TS_02. 전라도_03. 2인발화.zip | 13.44 GB | key: 437909 파일키 복사 다운로드
    - TS_03. 제주도_01. 1인발화 따라말하기.zip | 4.03 GB | key: 437910 파일키 복사 다운로드
    - TS_03. 제주도_02. 1인발화 질문에답하기.zip | 7.46 GB | key: 437911 파일키 복사 다운로드
    - TS_03. 제주도_03. 2인발화.zip | 2.86 GB | key: 437912 파일키 복사 다운로드
  - 02.라벨링데이터

About Our Topic

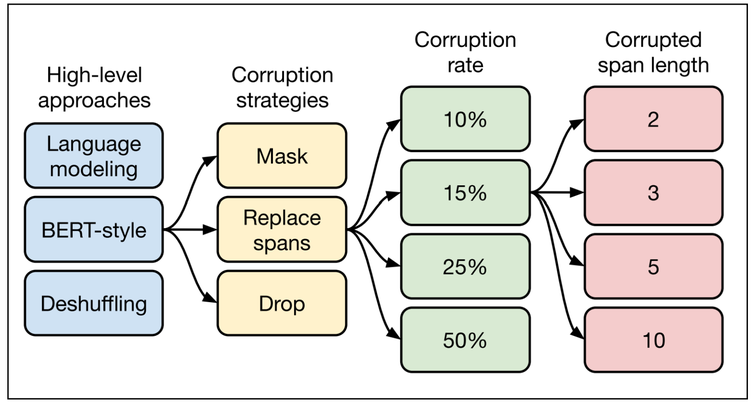
Jeju Dialect

- Training
  - 01.원천데이터
    - TS_01. 충청도_01. 1인발화 따라말하기.zip | 15.97 GB | key: 437904 파일키 복사 다운로드
    - TS_01. 충청도_02. 1인발화 질문에답하기.zip | 27.93 GB | key: 437905 파일키 복사 다운로드
    - TS_01. 충청도_03. 2인발화.zip | 11.58 GB | key: 437906 파일키 복사 다운로드
    - TS_02. 전라도_01. 1인발화 따라말하기.zip | 19.38 GB | key: 437907 파일키 복사 다운로드
    - TS_02. 전라도_02. 1인발화 질문에답하기.zip | 34.94 GB | key: 437908 파일키 복사 다운로드
    - TS_02. 전라도_03. 2인발화.zip | 13.44 GB | key: 437909 파일키 복사 다운로드
    - TS_03. 제주도_01. 1인발화 따라말하기.zip | 4.03 GB | key: 437910 파일키 복사 다운로드
    - TS_03. 제주도_02. 1인발화 질문에답하기.zip | 7.46 GB | key: 437911 파일키 복사 다운로드
    - TS_03. 제주도_03. 2인발화.zip | 2.86 GB | key: 437912 파일키 복사 다운로드
  - 02.라벨링데이터

Previous Research

KE-T5

https://arxiv.org/pdf/1910.10683



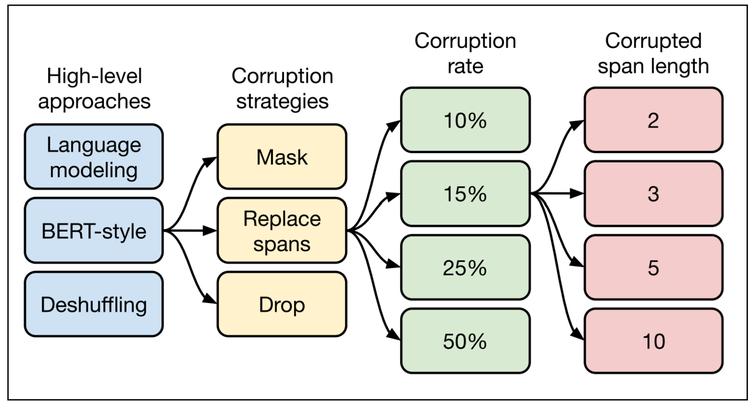
- Korean version of T5 : standard vanilla encoder-decoder transformer
  - Understand Korean structure and knowledge
  - Framework for transfer learning

    ➡ Fine-tune the model : not training the model from the bottom
- Training requires a huge amount of training data

Previous Research

KE-T5

https://arxiv.org/pdf/1910.10683



- Korean version of T5 : standard vanilla encoder-decoder transformer
  - Understand Korean structure and knowledge
  - Framework for transfer learning

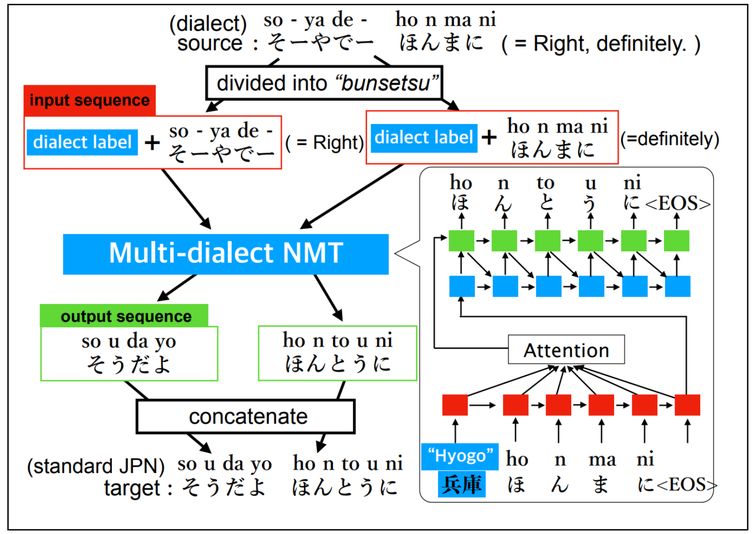
  ➡ Fine-tune the model : not training the model from the bottom

- **Training requires a huge amount of training data ➡ train multiple dialects**

Previous Research

Japanese Multi-Dialect NMT

https://aclanthology.org/Y18-1001.pdf



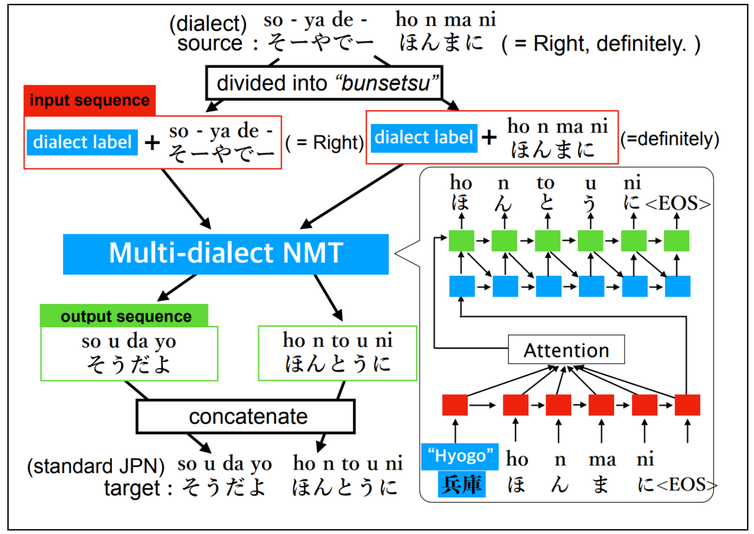
- Multi-Dialect **Neural Machine Translator**
- Japanese dialects ➡ multi dialect Neural MT was better than Statistical MT
  - Lack of dialect dataset

    ➡ Train multiple dialects simultaneously for better performance
  - Naturally possible to learn lexical and syntactic similarities by training, through Encoder
- Multi-layer LSTM Encoder-Decoder with Attention
  - Weakness in understanding input context and produce natural translation we want

Previous Research

Japanese Multi-Dialect NMT

https://aclanthology.org/Y18-1001.pdf



- Multi-Dialect **Neural Machine Translator**
- Japanese dialects ➡ multi dialect Neural MT was better than Statistical MT
  - Lack of dialect dataset

    ➡ Train multiple dialects simultaneously for better performance
  - Naturally possible to learn lexical and syntactic similarities by training, through Encoder
- Multi-layer LSTM Encoder-Decoder with Attention
  - **Weakness in understanding input context and produce natural translation we want ➡ use T5**

Previous Research

Cost Effective RLHF

https://aclanthology.org/Y18-1001.pdf

- BLEU : do not strictly align with human preference of translation quality
- Reinforcement Learning with Human Feedback
  - Making high quality preference data : High cost
- Optimizing reward models by distinguishing between human and MT

➡ Reward model learns the deficiencies of MT compared to human

Previous Research

Cost Effective RLHF

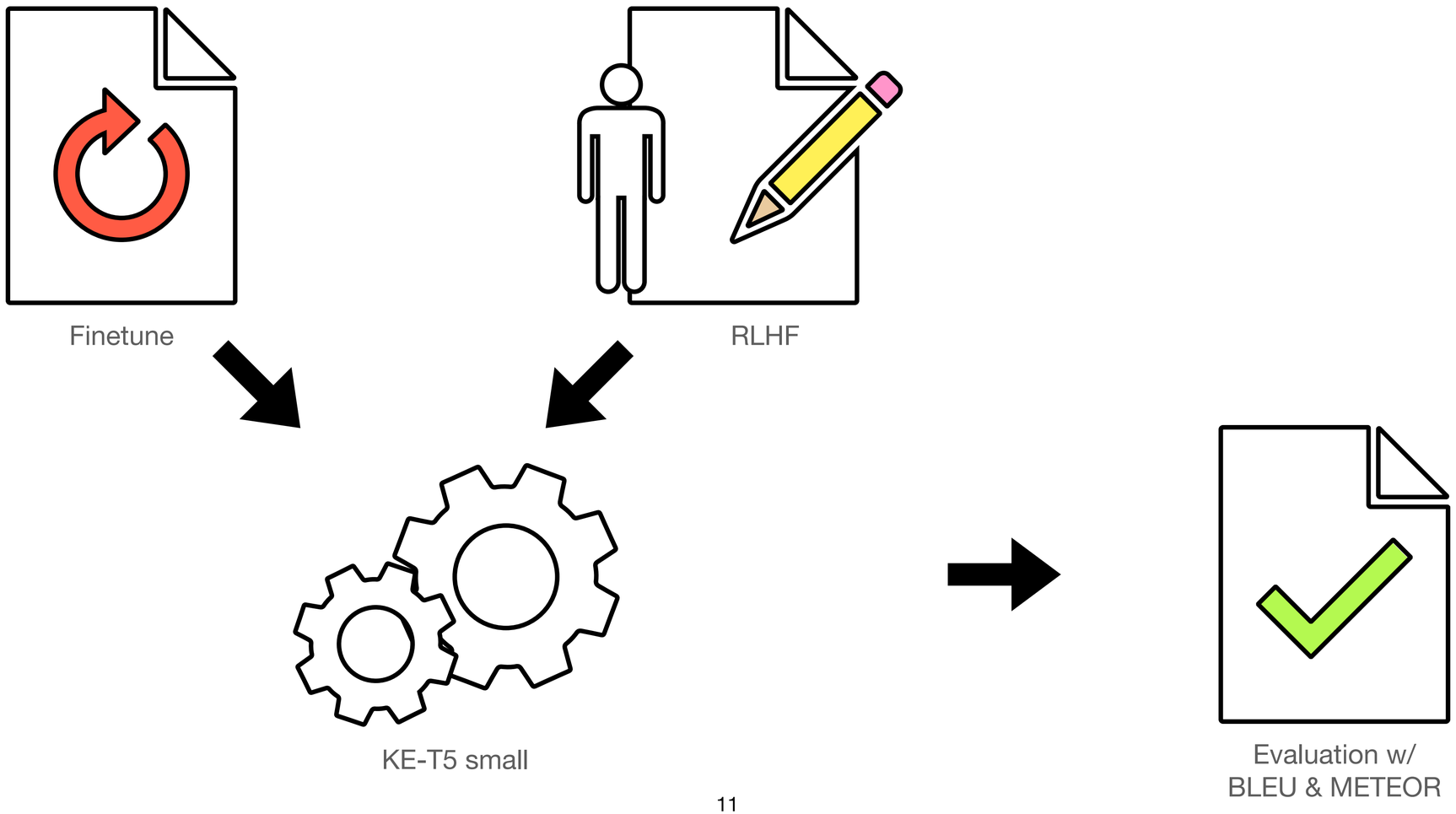
https://aclanthology.org/Y18-1001.pdf

- BLEU : do not strictly align with human preference of translation quality
- Reinforcement Learning with Human Feedback
  - Making high quality preference data : High cost
- **Optimizing reward models by distinguishing between human and MT**

➡ Reward model learns the deficiencies of MT compared to human

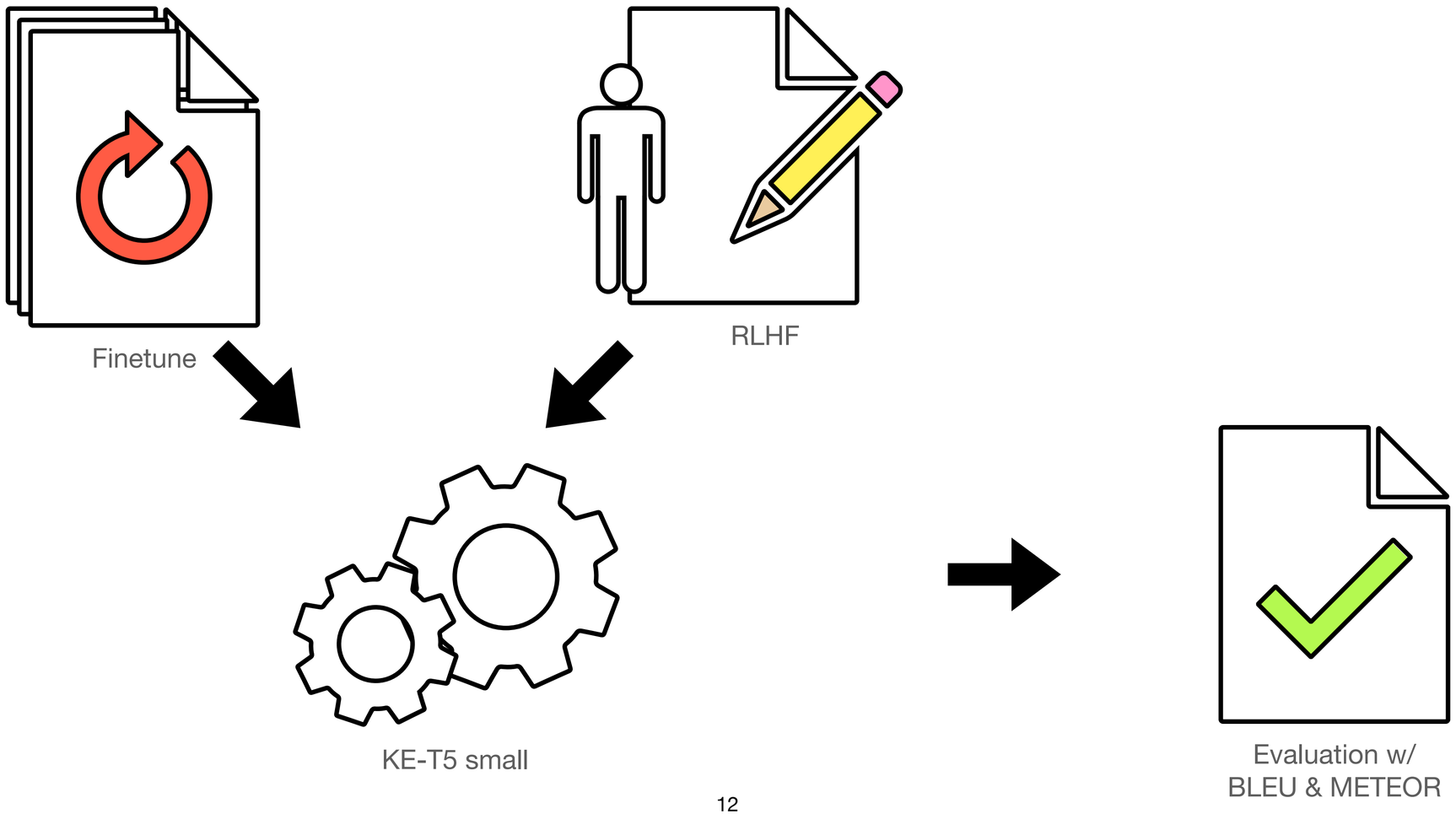
Our Approach

Model Overview



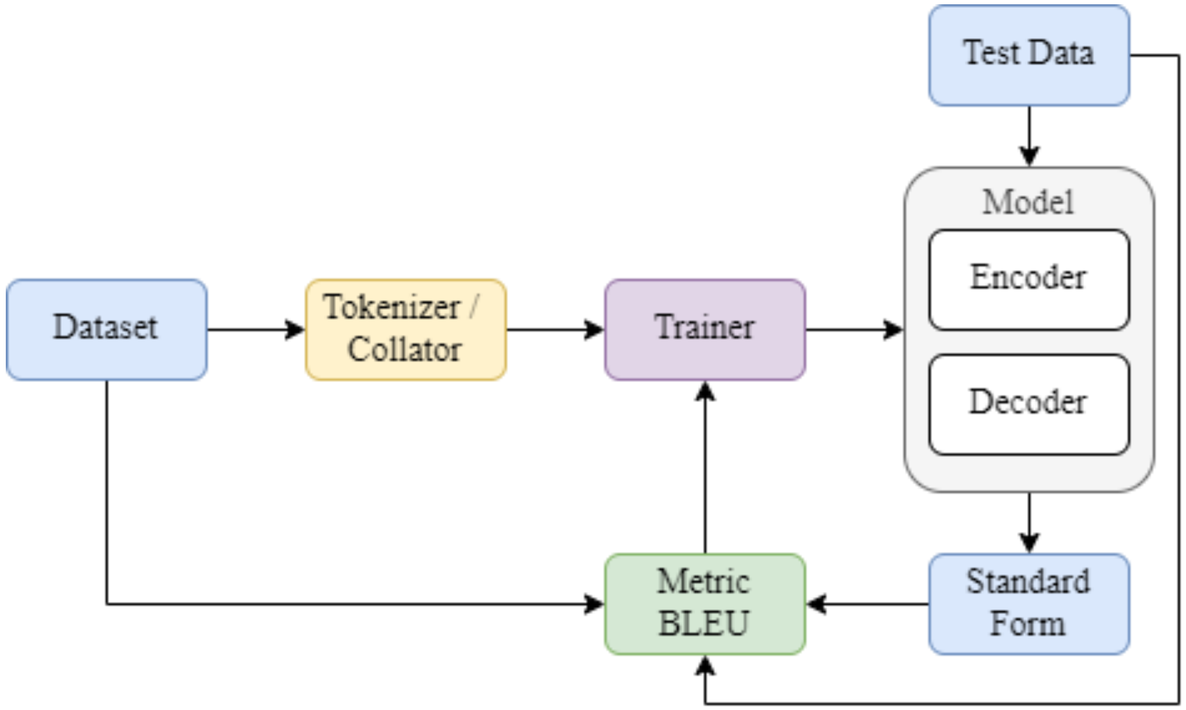
Our Approach

Model Overview



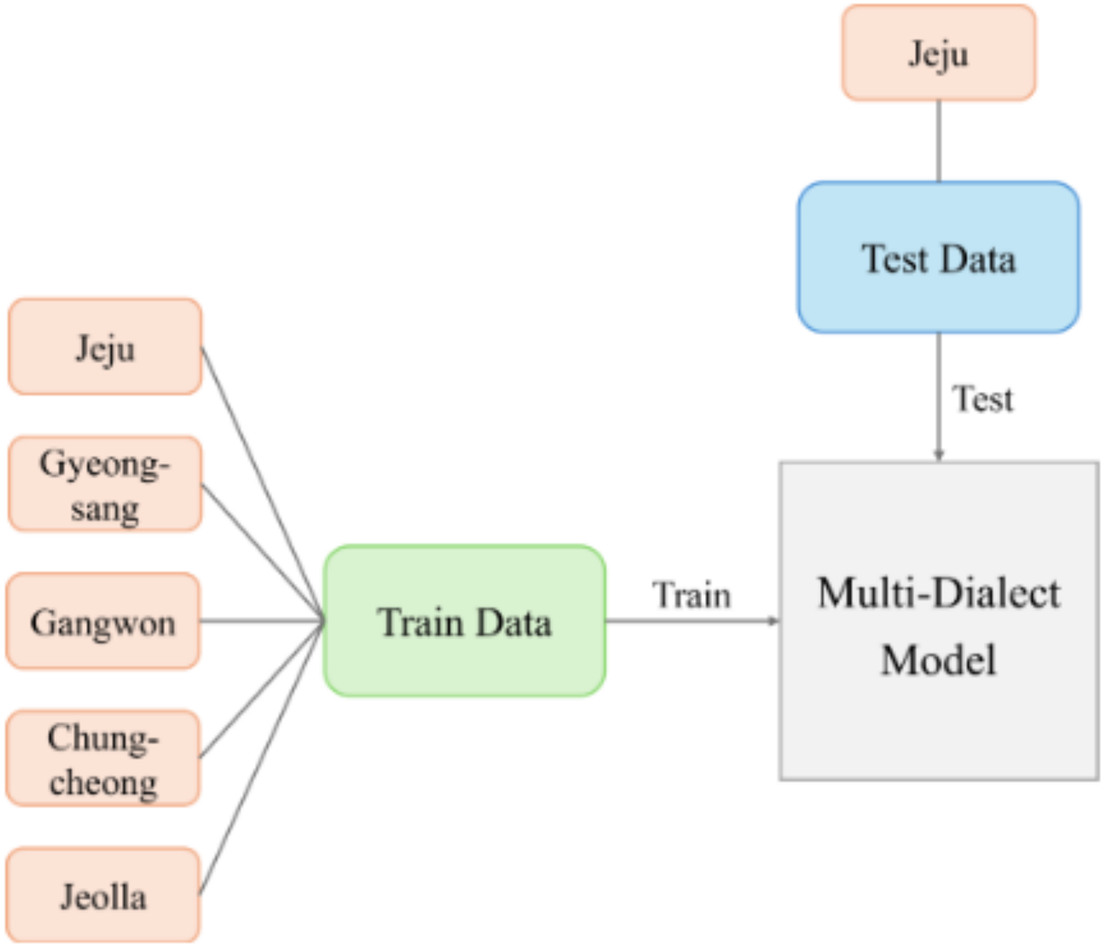
Our Approach

KE-T5 small



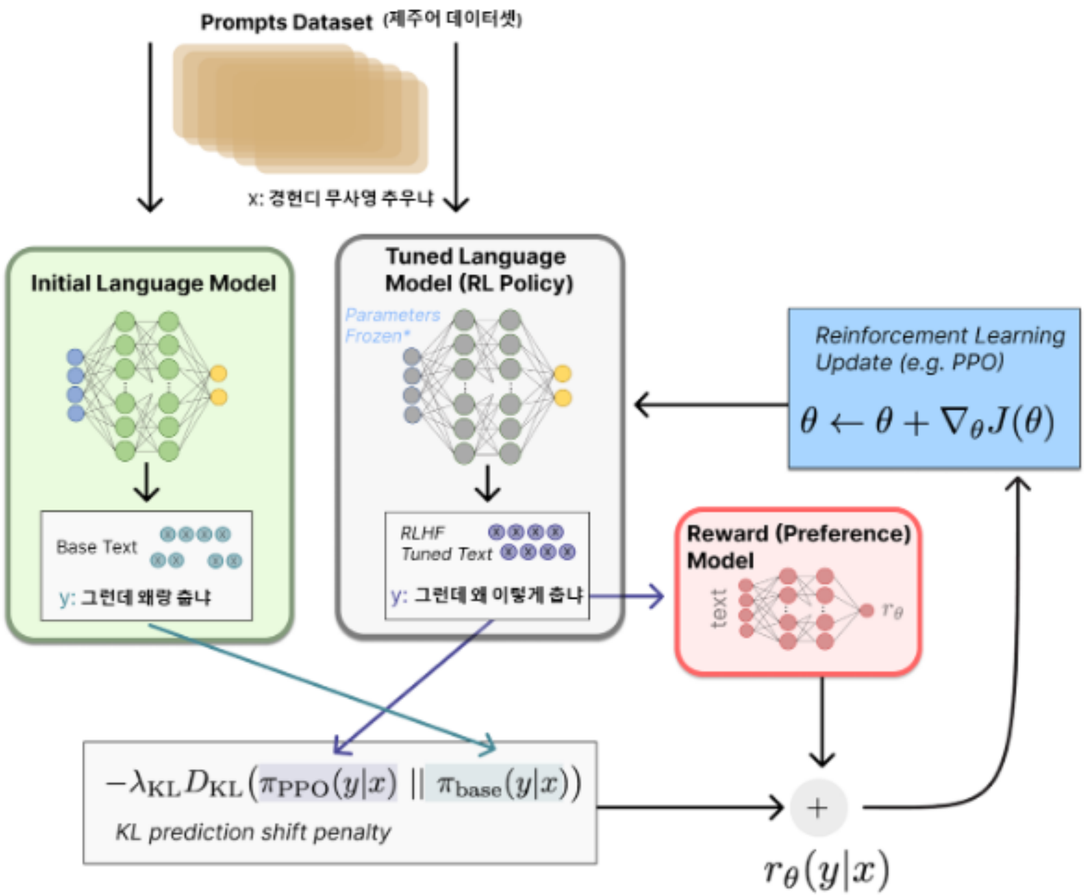
Our Approach

Multi-dialect



Our Approach

RLHF



Our Approach

Dataset

- From AI-hub (https://www.aihub.or.kr/)
  - 한국어 방언 발화 (제주도)
  - 중·노년층 한국어 방언 데이터 (충청도, 전라도, 제주도, 강원도, 경상도)

Our Approach

Dataset

- Voice data : Source data
- Text data : Labeled data
  - Transcribed to include dialect text and corresponding standard language pairs

Our Approach

Evaluation

https://aclanthology.org/P02-1040.pdf

- BLEU (**B**i**L**ingual **E**valuation **U**nderstudy)
  - Evaluate quality of MT : Compare similarity between MT and human translation

➡ Limitation due to emphasis on precision without considering recall

  - Can miss important aspects of how much of the reference translation is captured by MT

Our Approach

Evaluation

https://aclanthology.org/W05-0909.pdf

- METEOR (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering)
  - Harmonic mean of precision and recall
    - Precision : Accuracy of the translation
    - Recall : Completeness
  - Evaluate word-to-word matches between translated & reference text
    - Adjust for stemming and synonymy
    - Incorporates a measure of sentence structure alignment

➡ More nuanced evaluation of translation quality

Experiment

Details

- Single-dialect KE-T5 Small
  - 350,000 train & validation + 10,000 test : Jeju dialect sentences
  - About 40 minutes with L4 GPU
- Multi-dialect KE-T5 Small
  - 525,000 train & validation + 10,000 test : multi-dialect sentences
  - About an hour with L4 GPU

Experiment

Details

- Reward Model
  - Over 300,000 pairs of human translation
  - About 3 hours with L4 GPU
- Fine-tuning via PPO
  - 50,000 Jeju dialect sentences
  - About 3.5 hours with T4 GPU

Experiment

Results

Dataset	BLEU	METEOR
Single Jeju Dialect	68.26	0.83
Multi-Dialect	**73.38**	**0.86**
Single Jeju + RL	66.27	0.81
Multi-Dialect + RL	72.33	0.85
GPT4o	23.20	0.37

Table 1: BLEU and METEOR scores for our methods and GPT4o for reference.

Experiment

Analysis : Multi-dialect

Dataset	BLEU	METEOR
Single Jeju Dialect	68.26	0.83
Multi-Dialect	**73.38**	**0.86**
Single Jeju + RL	66.27	0.81
Multi-Dialect + RL	72.33	0.85
GPT4o	23.20	0.37

Increase

Table 1: BLEU and METEOR scores for our methods and GPT4o for reference.

Experiment

Analysis : Multi-dialect

Dataset	BLEU	METEOR
Single Jeju Dialect	68.26	0.83
Multi-Dialect	**73.38**	**0.86**
Single Jeju + RL	66.27	0.81
Multi-Dialect + RL	72.33	0.85
GPT4o	23.20	0.37

Decrease

Decrease

Table 1: BLEU and METEOR scores for our methods and GPT4o for reference.

Experiment

Analysis : Multi-dialect

Original	[‘몸이 자꾸 고료운 거 보난 니가 신 셍이여’]
Reference	[‘몸이 자꾸 **가려운** 거 보니 **이가 있는** 모양이야’]
Single T5	[‘몸이 자꾸 **고료운** 거 보니 **이 있는** 모양이야’]
Multi T5	[‘몸이 자꾸 **가려운** 것을 보니 **이가 있는** 모양이야’]
Single T5 BLEU score	19.13
Multi T5 BLEU score	50.00

Table 2: Comparative translations showing the effectiveness of multi-dialect training.

Experiment

Analysis : Multi-dialect

Original	['몸이 자꾸 고료운 거 보난 니가
Reference	['몸이 자꾸 **가려운** 거 보니 **이가**
Single T5	['몸이 자꾸 **고료운** 거 보니 **이 있는** 모양이야']
Multi T5	['몸이 자꾸 **가려운** 것을 보니 **이가 있는** 모양이야']
Single T5 BLEU score	19.13
Multi T5 BLEU score	50.00

Frequently used subject marker (조사)

Table 2: Comparative translations showing the effectiveness of multi-dialect training.

Experiment

Analysis : RLHF

Original	[‘집 앞이 손 매는 건 집에 아픈 사름이 시나 굿을 허나 다 이유가 싯주게’]
Reference	[‘집 앞에 손 매는 **건** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지’]
RLHF Model	[‘집 앞에 손 매는 **것은** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지’]
Multi T5	[‘집 앞에 손 매는 **건** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지’]
RLHF BLEU score	78.25
Multi T5 BLEU score	100.00

Table 3: Comparative translations showing the impact of RLHF on translation quality.

Experiment

Analysis : RLHF

Output	Reward
집 앞에 손 매는 **건** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지	3.49
집 앞에 손 매는 **것은** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지	3.59

Table 5: Output of reward model

Higher score for formal expressions

Original	['집 앞이 손 매는 [illegible] 굿을 허나 다 이유가 싯주게']
Reference	['집 앞에 손 매는 **건** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지']
RLHF Model	['집 앞에 손 매는 **것은** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지']
Multi T5	['집 앞에 손 매는 **건** 집에 아픈 사람이 있거나 굿을 하거나 다 이유가 있지']
RLHF BLEU score	78.25
Multi T5 BLEU score	100.00

Table 3: Comparative translations showing the impact of RLHF on translation quality.

Experiment

Analysis : RLHF

Original	[‘옛날 수도 엇인 땐 집이 큰일 나민 동네 사름덜이 다 물 질어다 어’]
Reference	[‘옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어’]
RLHF Model	[‘옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어’]
Multi T5	[‘옛날 수도 없을 **땐 집이** 큰일 나면 동네 사람들이 다 물 길어다 줬어’]
RLHF score	100.00
Multi T5 score	67.03

Table 7: Comparison of RLHF and Multi T5 model translations with performance scores.

Experiment

Analysis : RLHF

If reference is formal, increase in score

Original	['옛날 수도 엇인 땐 집이 큰일 나민 동네 사름덜이 다 물 질어다 어']
Reference	['옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
RLHF Model	['옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
Multi T5	['옛날 수도 없을 **땐 집이** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
RLHF score	100.00
Multi T5 score	67.03

Table 7: Comparison of RLHF and Multi T5 model translations with performance scores.

Experiment

Analysis : RLHF

If reference is formal, increase in score

Original	['옛날 수도 엇인 땐 집이 큰일 나민 동네 사름덜이 다 물 질어다 어']
Reference	['옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
RLHF Model	['옛날 수도 없을 **때는 집에** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
Multi T5	['옛날 수도 없을 **땐 집이** 큰일 나면 동네 사람들이 다 물 길어다 줬어']
RLHF score	100.00
Multi T5 score	67.03

Table 7: Comparison of RLHF and Multi T5 model translations with performance scores.

Thank You! ☺️