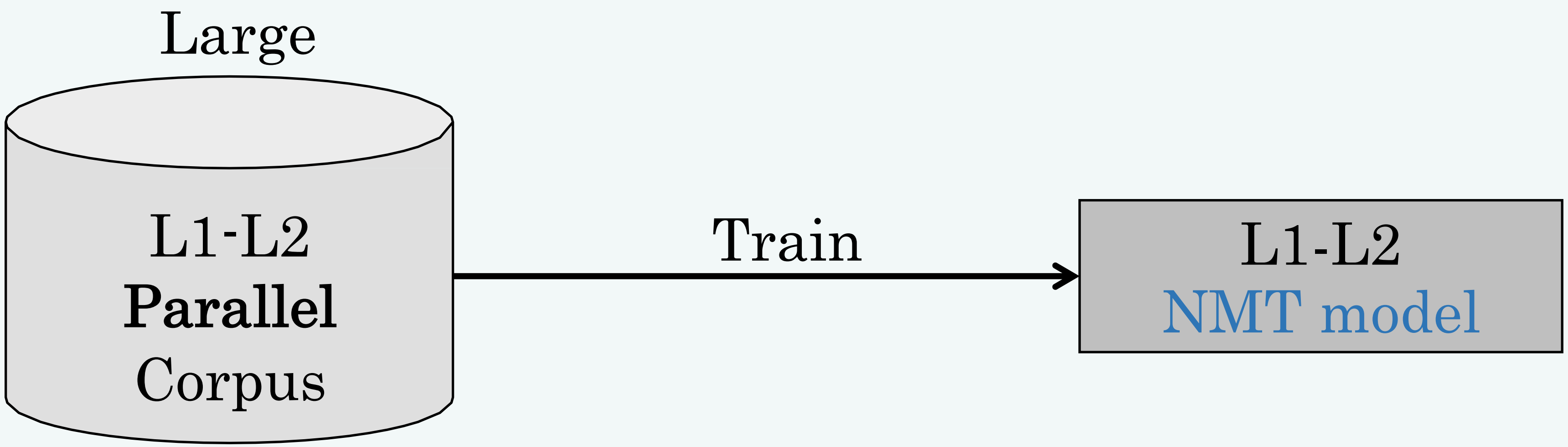
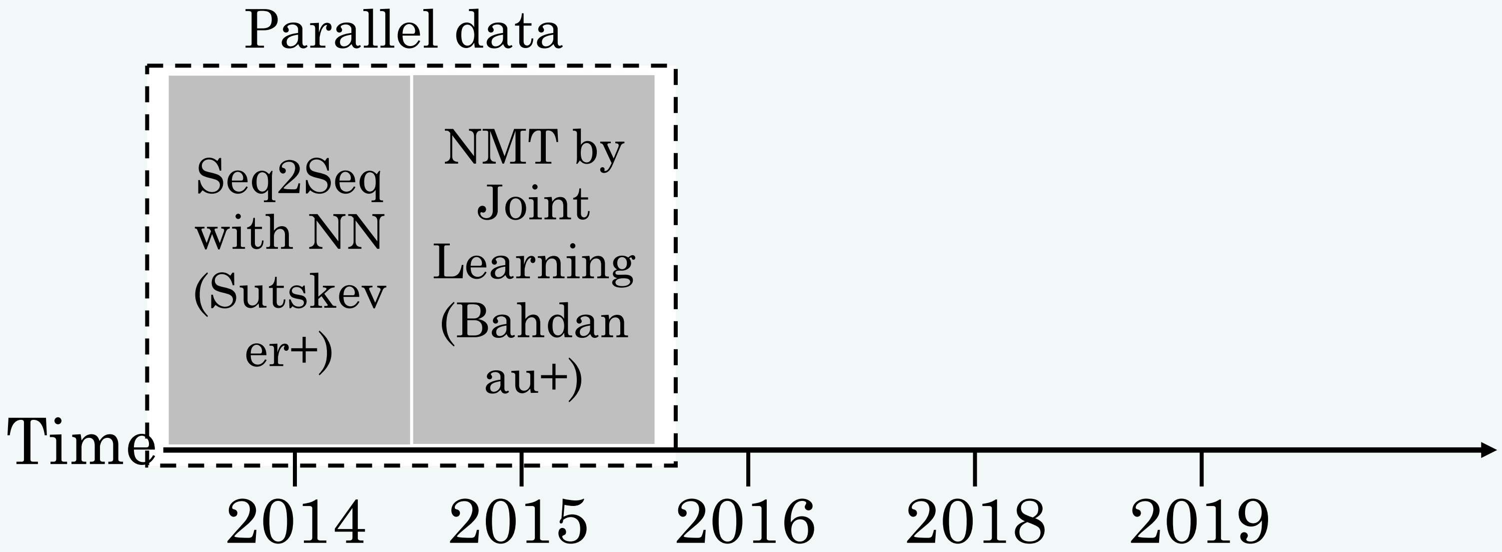
Pre-training via Leveraging Assisting Languages for Neural Machine Translation

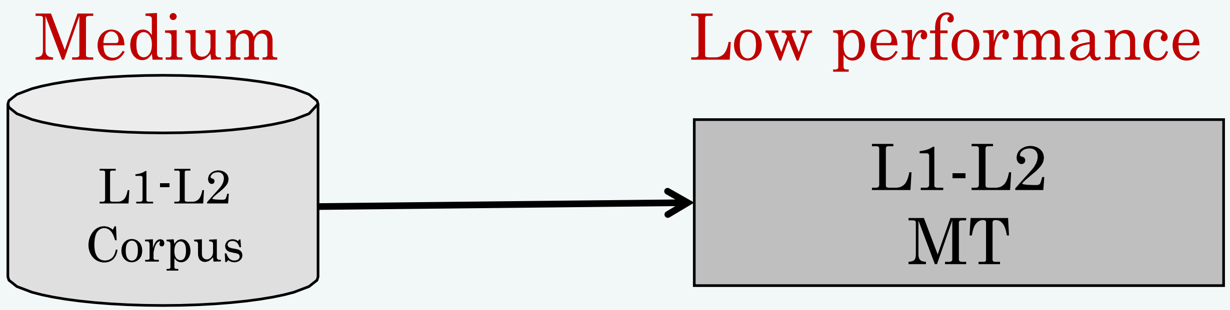
Haiyue Song[1], Raj Dabre[2], Zhuoyuan Mao[1], Fei Cheng[1], Sadao Kurohashi[1], Eiichiro Sumita[2]
[1]Kyoto University [2]NICT

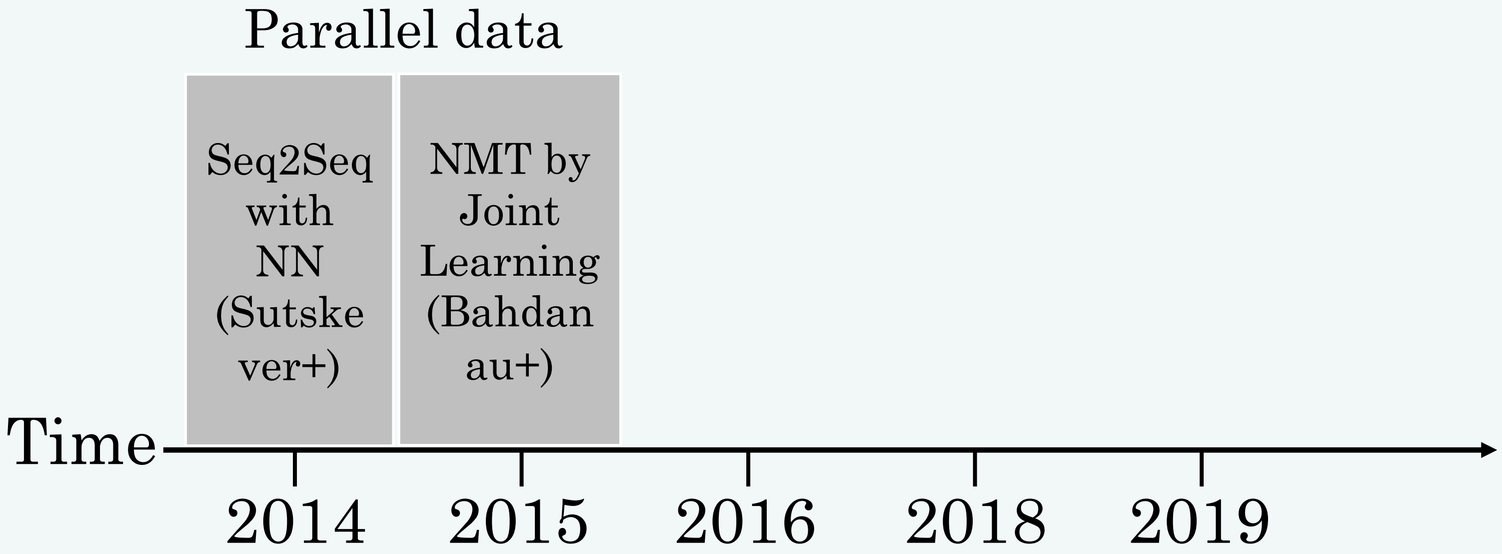
Enough parallel data ← Neural Machine Translation



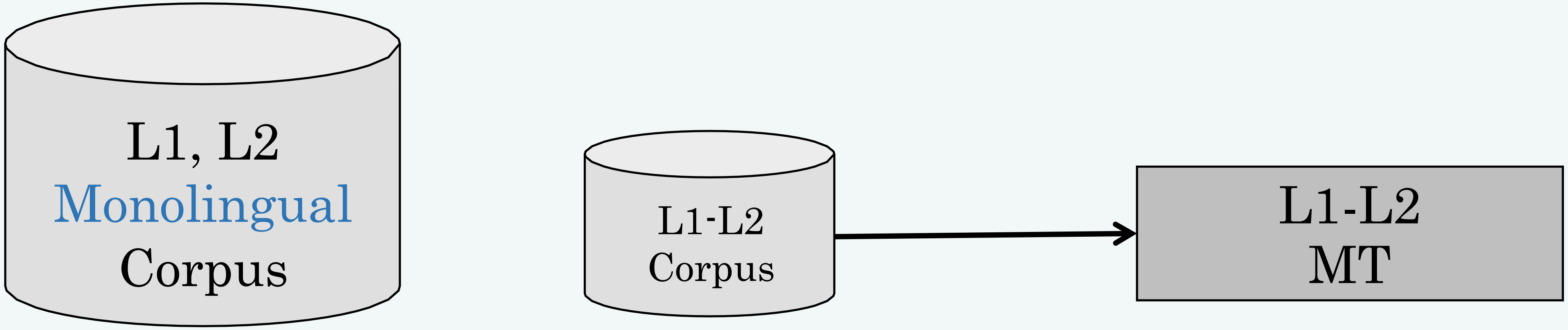


Lack of large parallel corpora → Low performance

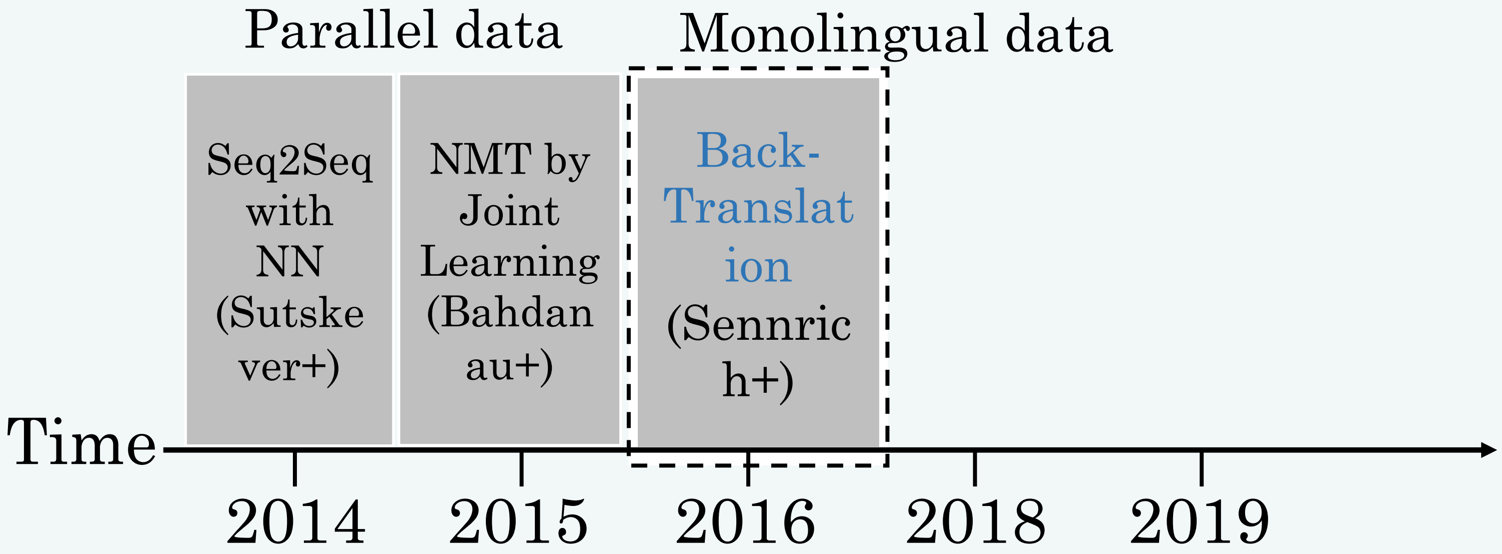




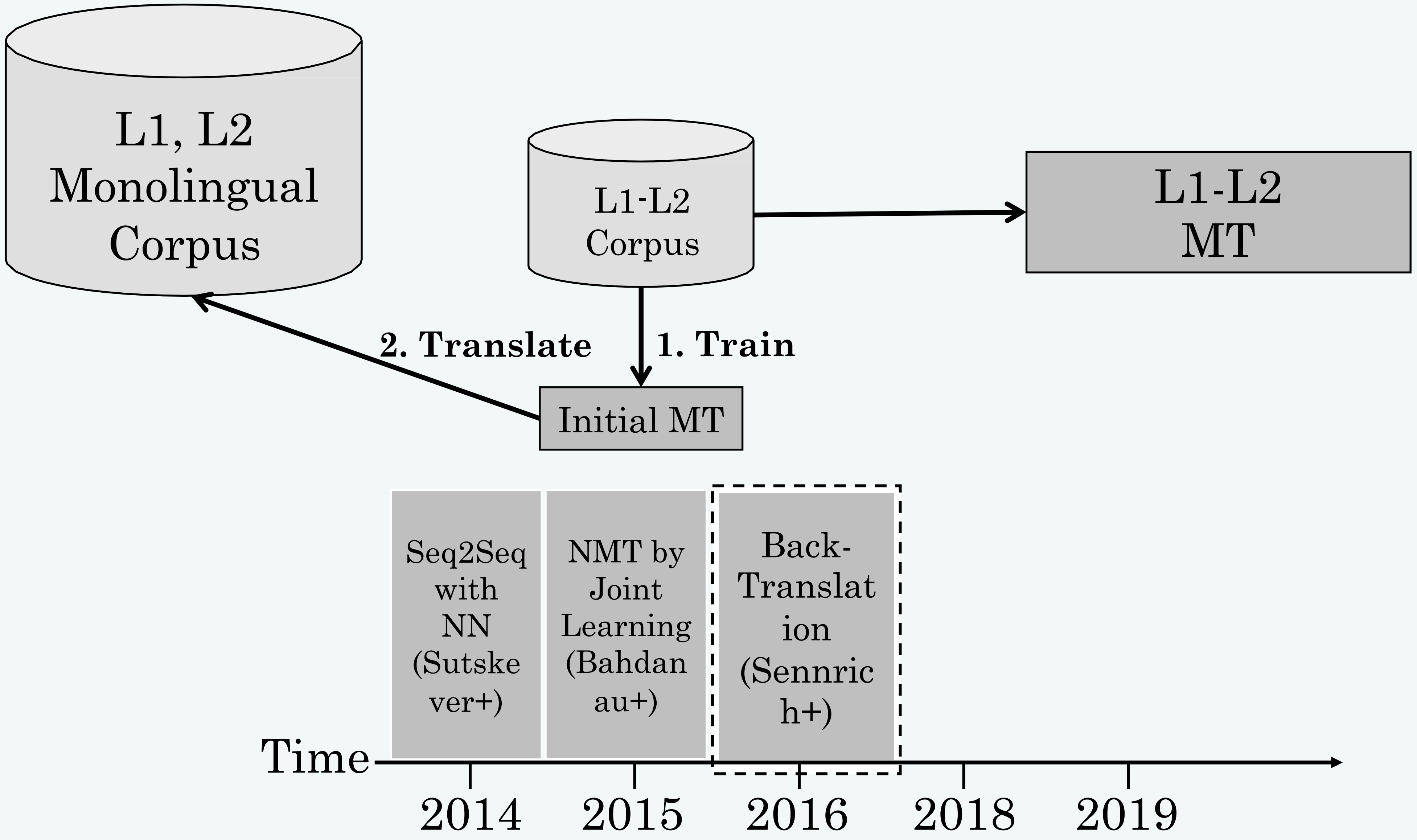
Lack of large parallel corpus ← Back Translation



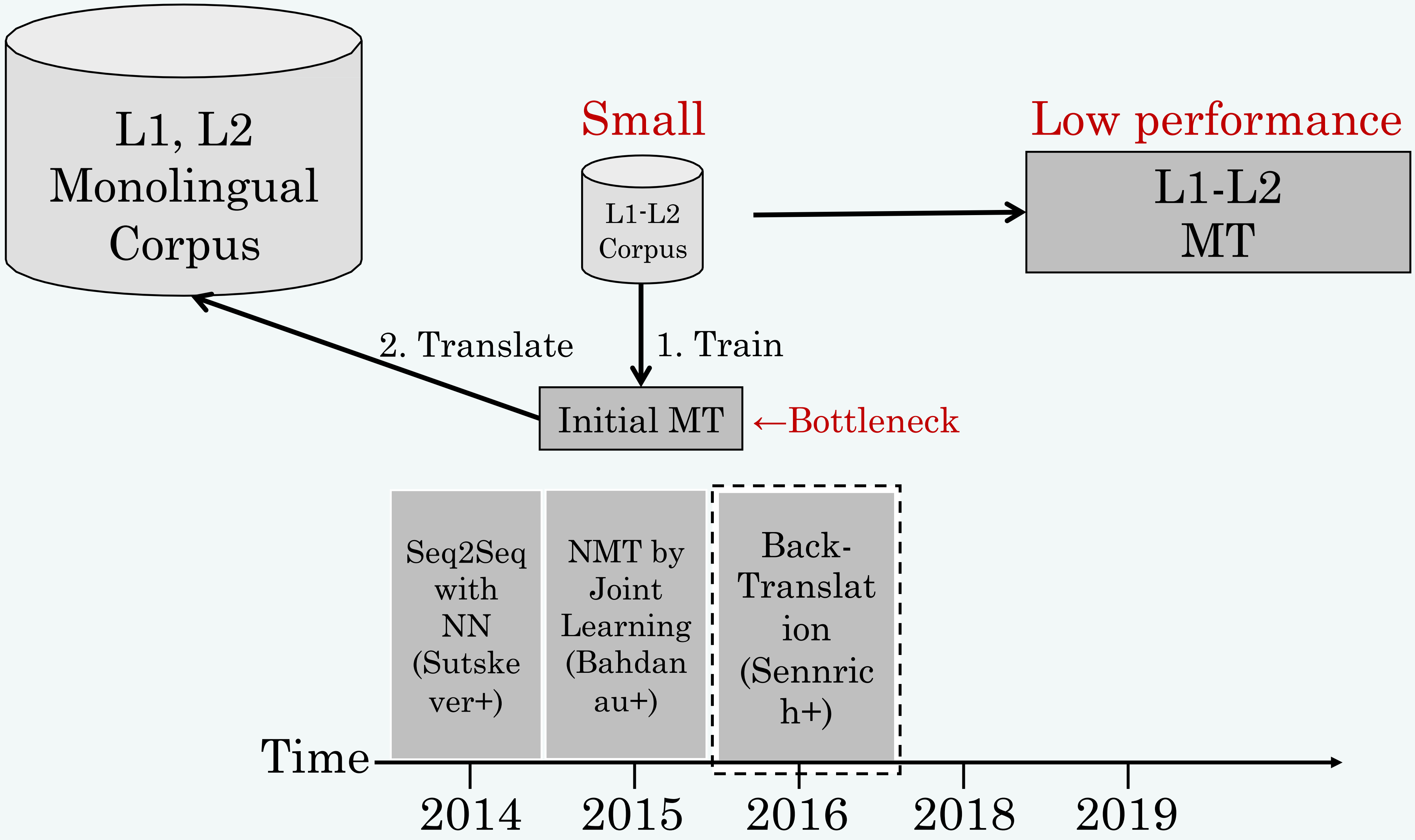
Easier to get



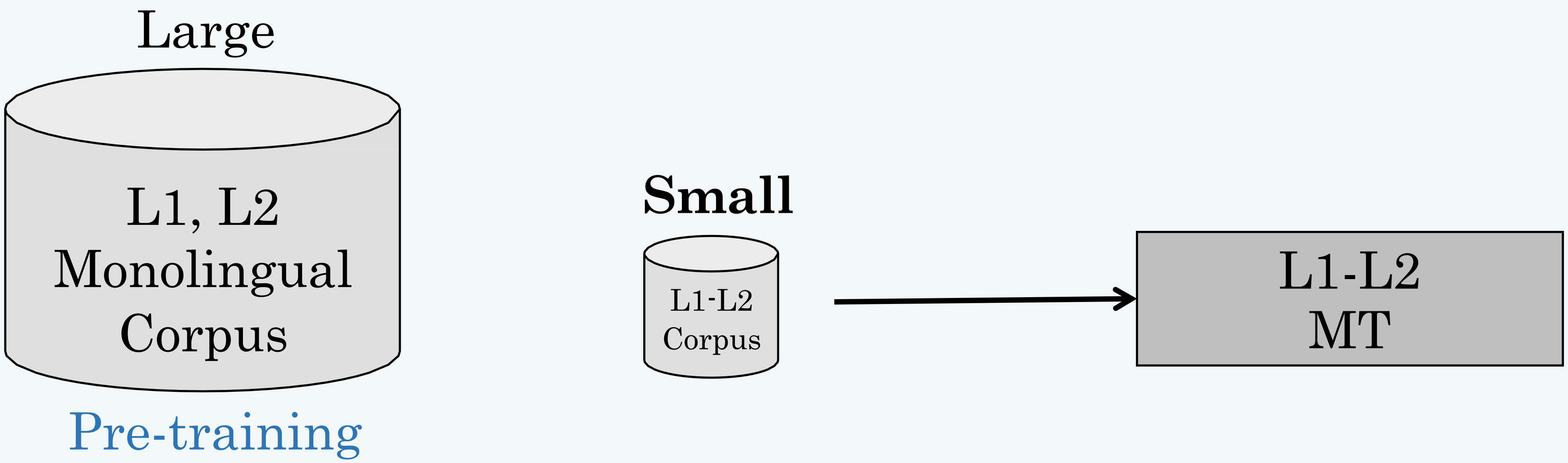
BT: Initial MT → Final MT performance

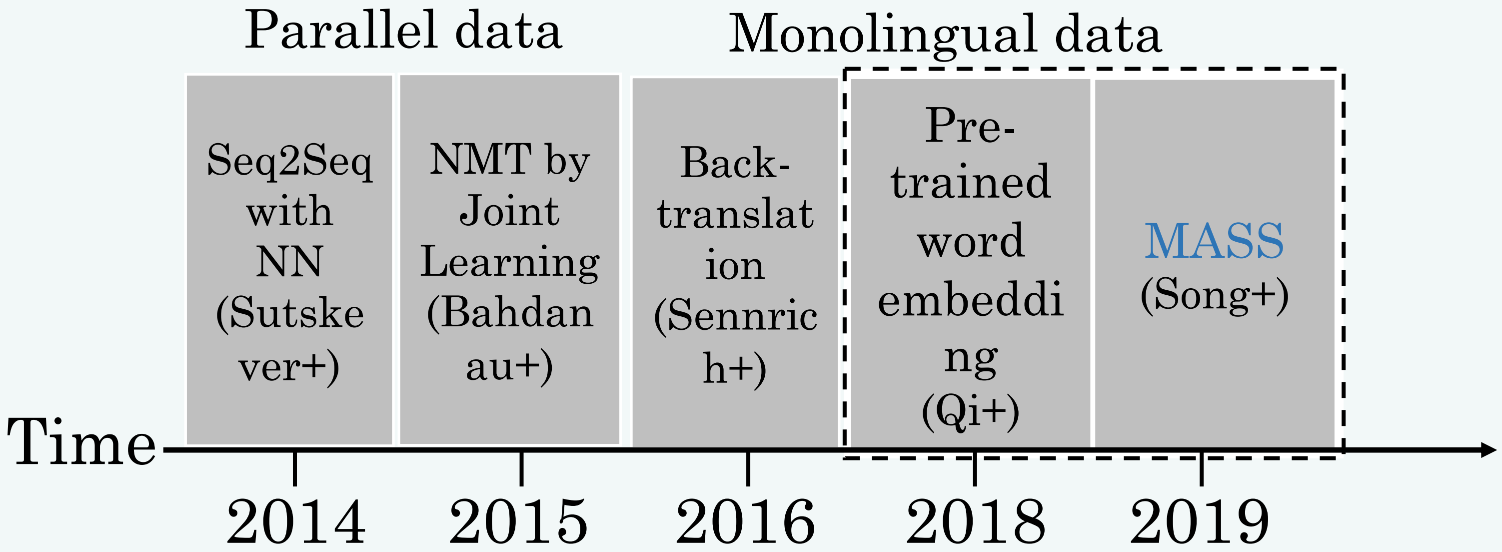


BT: Small parallel corpus → Low performance

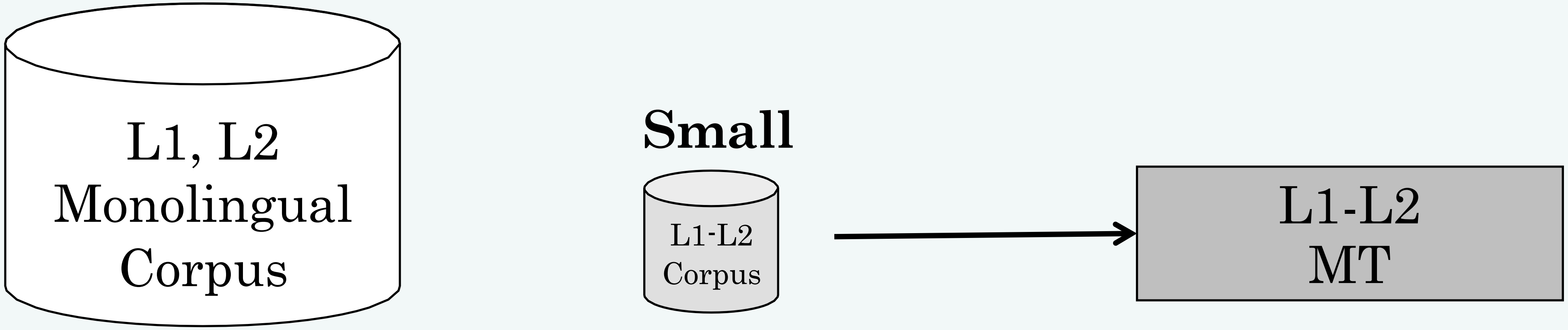


Low-resource situation ← Pre-train

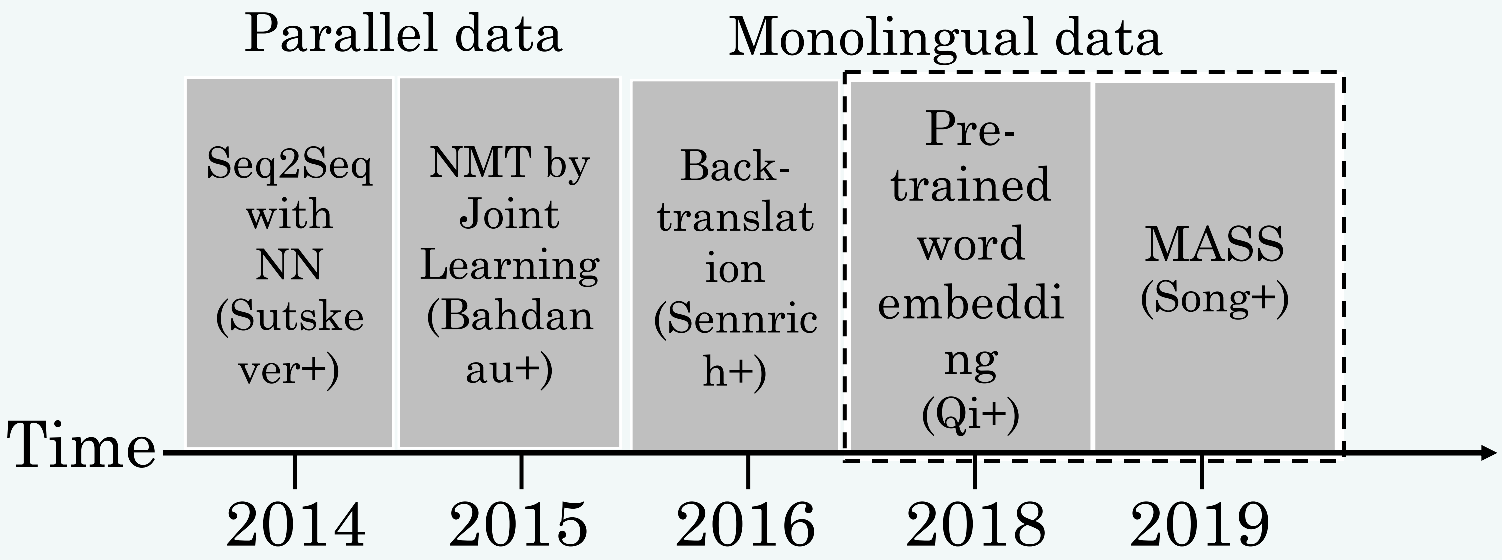




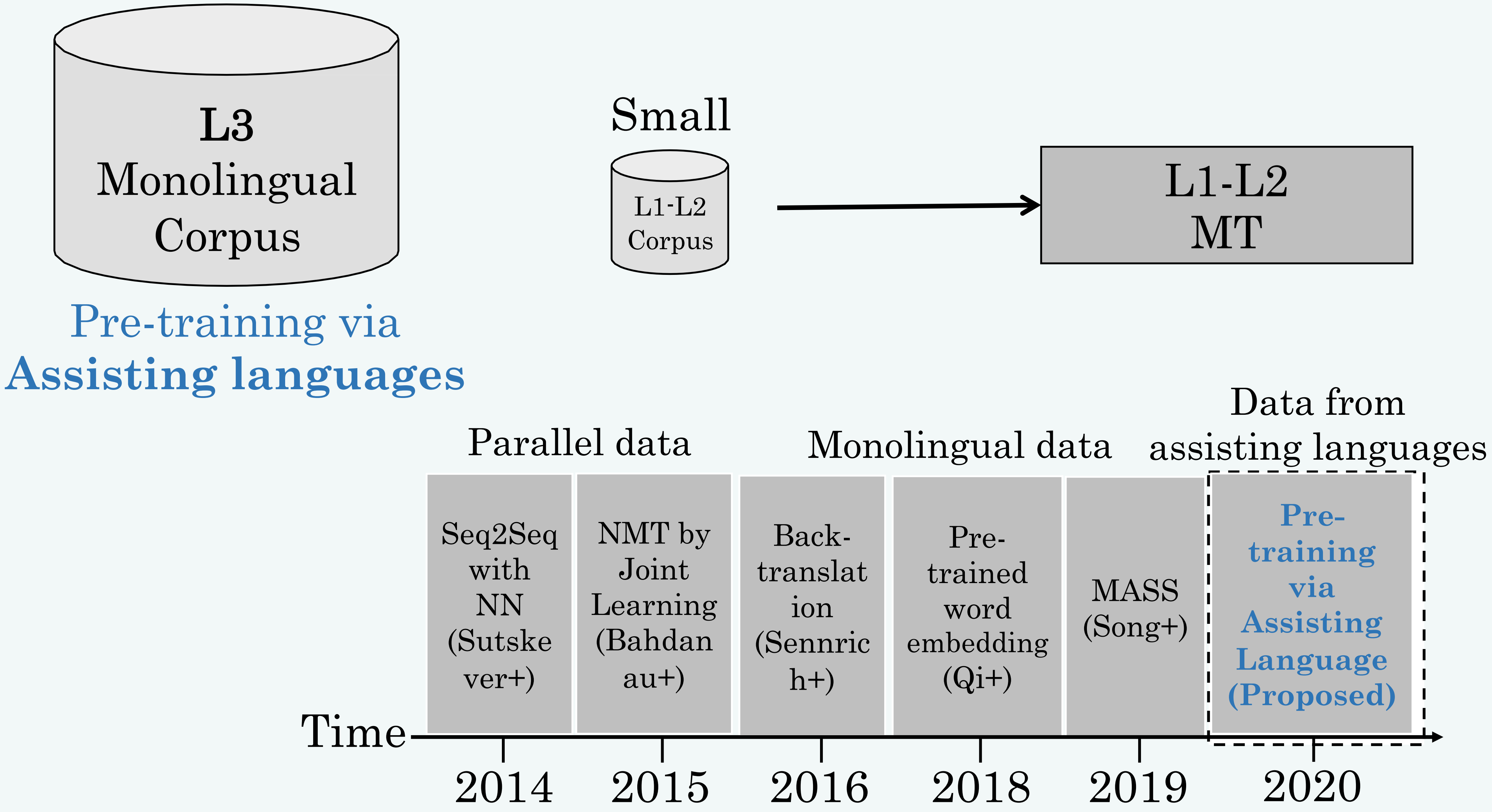
Extreme Low-resource: Lack both parallel and monolingual



Lack of monolingual data?



Lack of monolingual data ← Proposed method

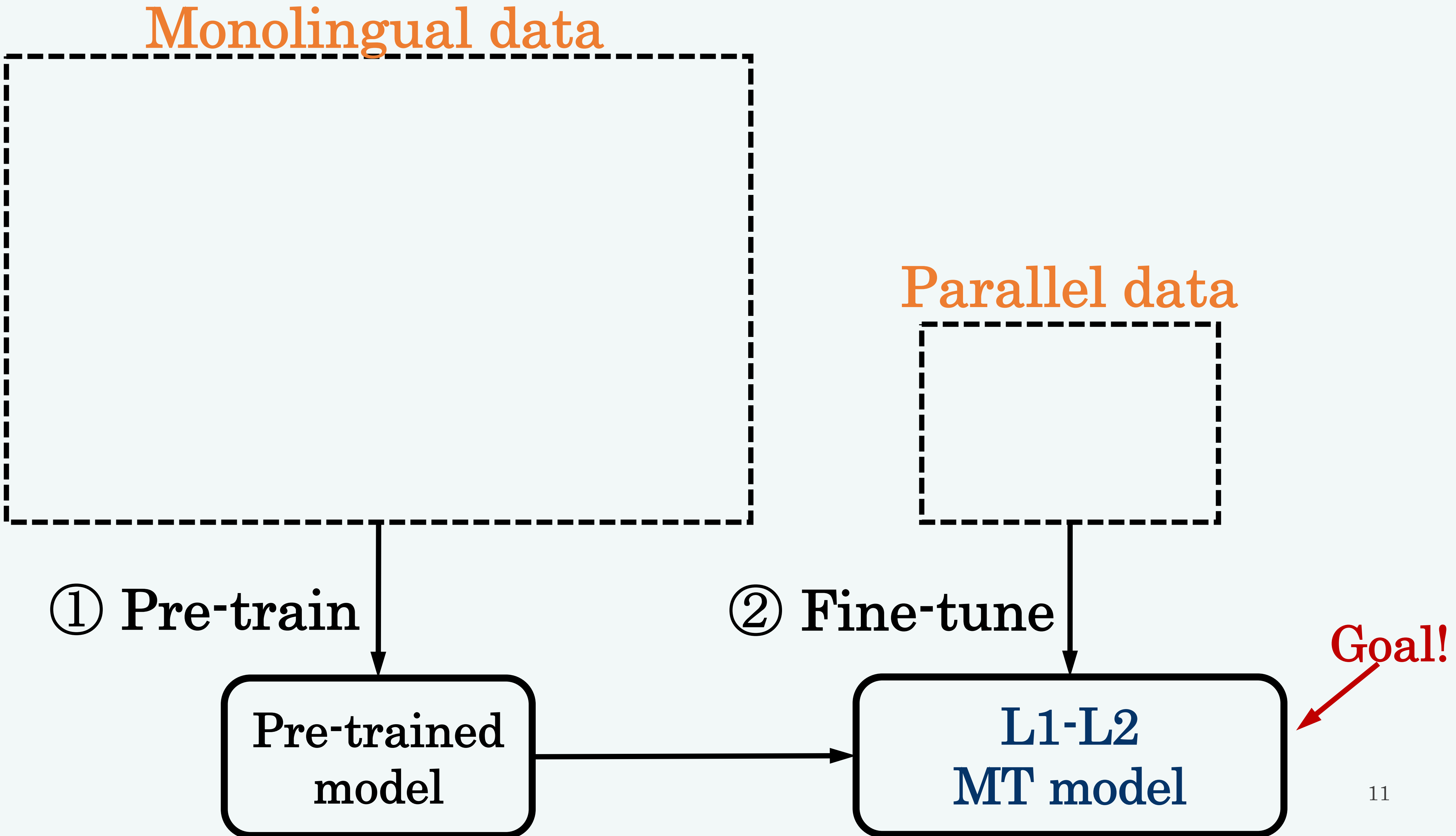


Proposed method: Overview

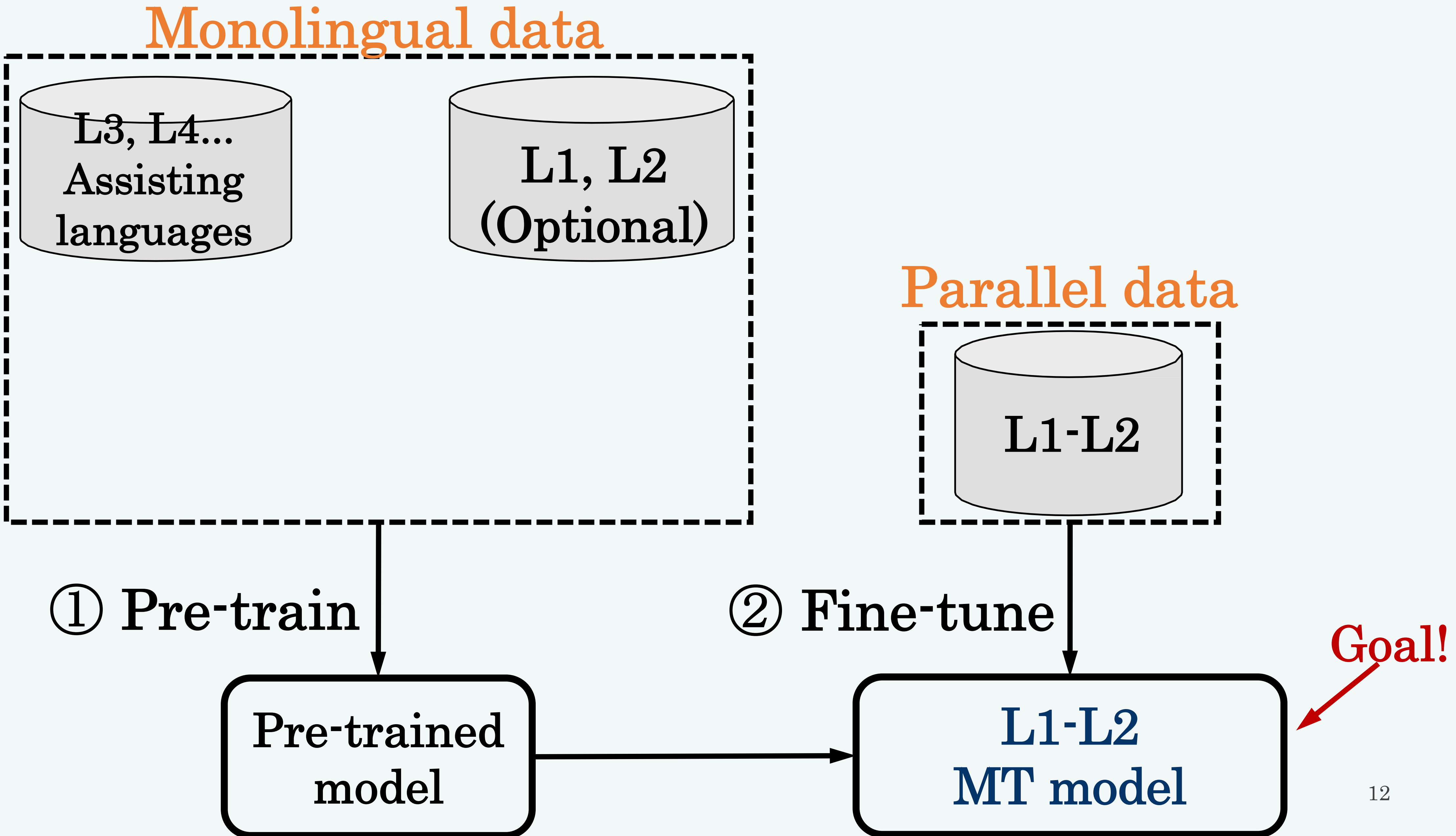
Goal!

L1-L2
MT model

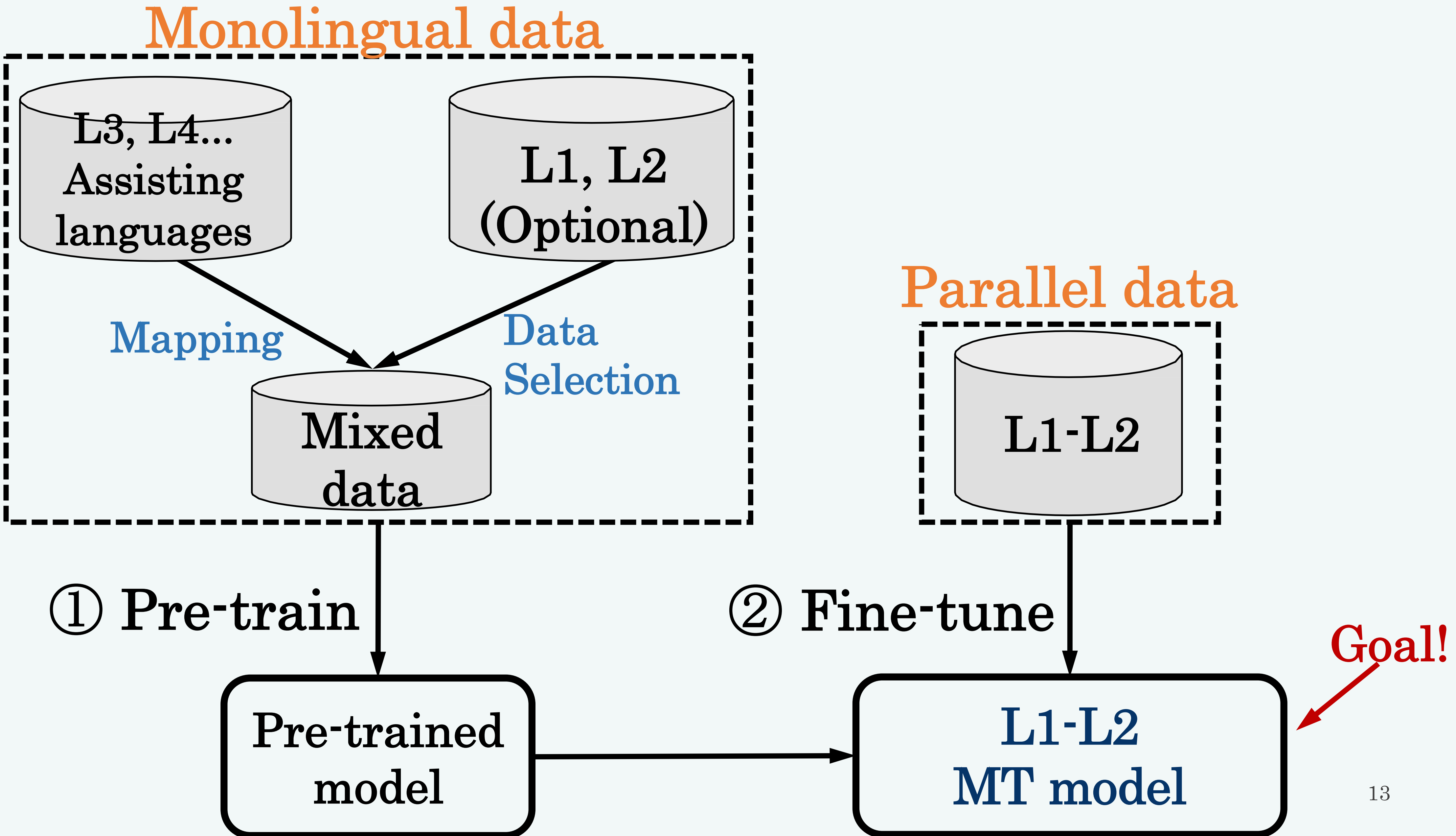
Proposed method: Overview



Proposed method: Overview



Proposed method: Overview



Proposed method: Mapping

Goal:

Maximize the cognate sharing

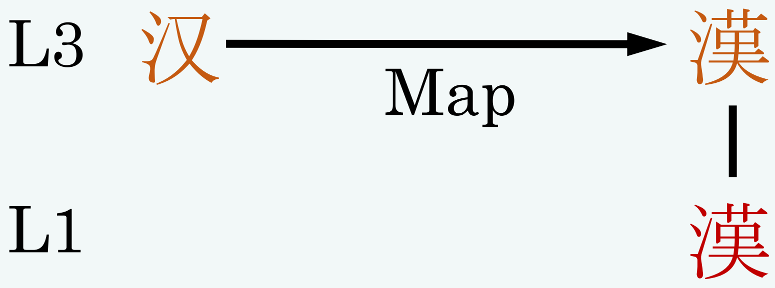
L3 汉

L1 漢

Proposed method: Mapping

Goal:

Maximize the cognate sharing



Proposed method: Mapping

Goal:

Maximize the cognate sharing

Example:

Chinese Hanzi and Japanese Kanji

	Chinese Hanzi	Japanese Kanji
Early ages	漢	漢

Proposed method: Mapping

Goal:

Maximize the cognate sharing

Example:

Chinese Hanzi and Japanese Kanji

Background:

Kanji borrowed from Hanzi

	Chinese Hanzi		Japanese Kanji
Early ages	漢	→	漢

Proposed method: Mapping

Goal:

Maximize the cognate sharing

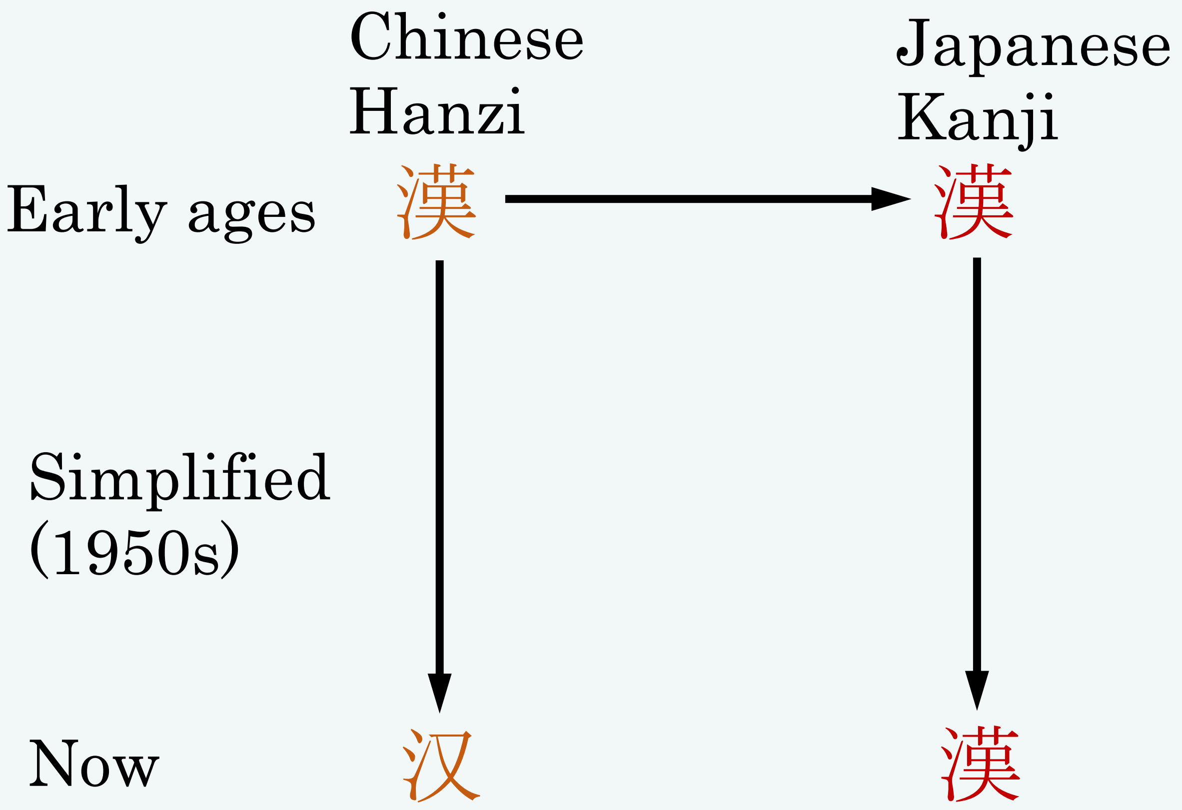
Example:

Chinese Hanzi and Japanese Kanji

Background:

Kanji borrowed from Hanzi

Over time the written scripts diverged



Proposed method: Mapping

Goal:

Maximize the cognate sharing

Example:

Chinese Hanzi and Japanese Kanji

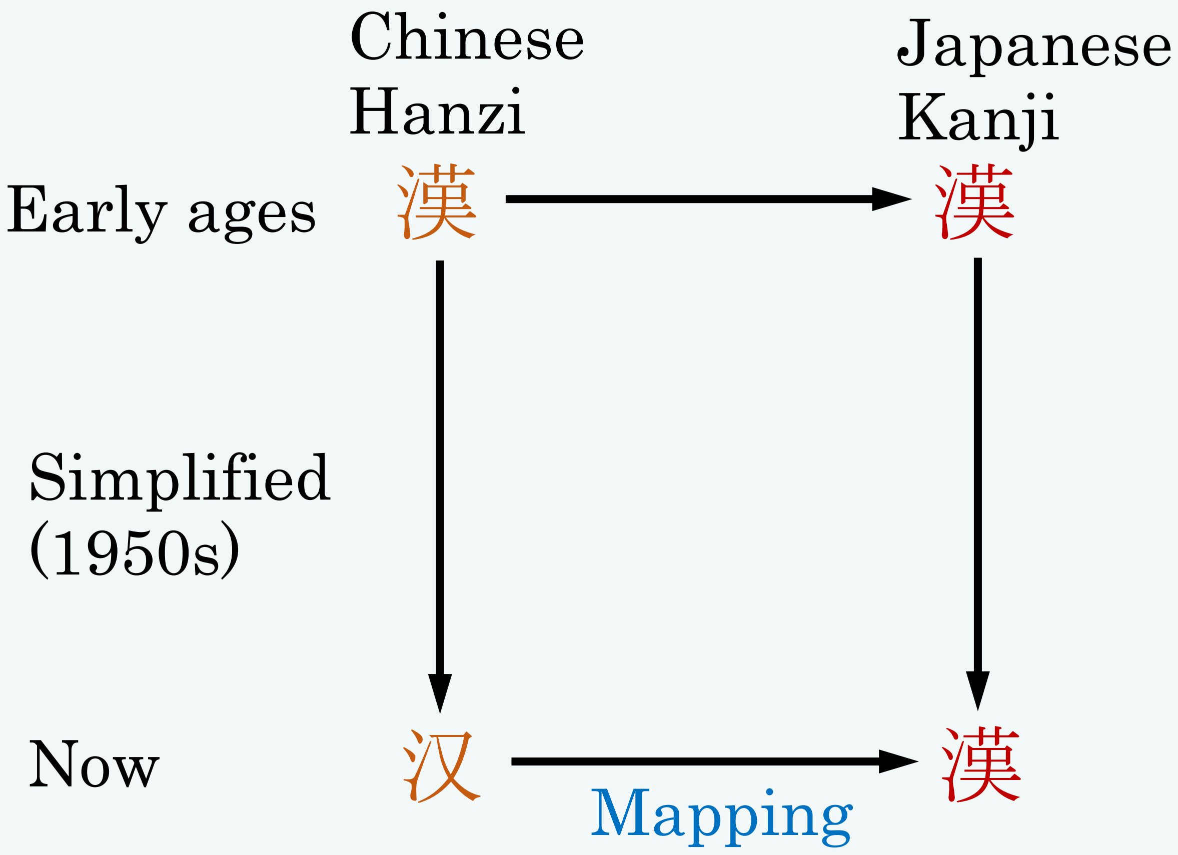
Background:

Kanji borrowed from Hanzi

Over time the written scripts diverged

Method:

Map Hanzi to Kanji by a mapping table (Chu et al., 2012)



Proposed method: Mapping

Goal:

Maximize the cognate sharing

Method: Map Hanzi to Kanji

One Hanzi may map to many Kanji

Hanzi		Kanji
机	→	[机, 機]
构	→	[構, 搆]

Proposed method: Mapping

Goal:

Maximize the cognate sharing

Method: Map Hanzi to Kanji

One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Hanzi		Kanji
机	→	[机, 機]
构	→	[構, 搆]

Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Hanzi		Kanji
机	→	[机, 機]
构	→	[構, 搆]

Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Method 2: many-to-many mapping

Hanzi → Kanji

机 → [机, 機]

构 → [構, 搆]

Chinese word → Japanese word (Synthetic)

机构 → [机構, 机搆, 機構,機搆]

Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Method 2: word-to-word mapping

Hanzi		Kanji
机	→	[机, 機]
构	→	[構, 搆]

Chinese word		Japanese word (Synthetic)
机构	→	[机構, 机搆, 機構,機搆]

↑

Japanese LM

Proposed method: Mapping

Goal:
Maximize the cognate sharing

Method: Map Hanzi to Kanji
One Hanzi may map to many Kanji

Method 1: one-to-one mapping

Method 2: word-to-word mapping

Hanzi		Kanji
机	→	[机, 機]
构	→	[構, 搆]

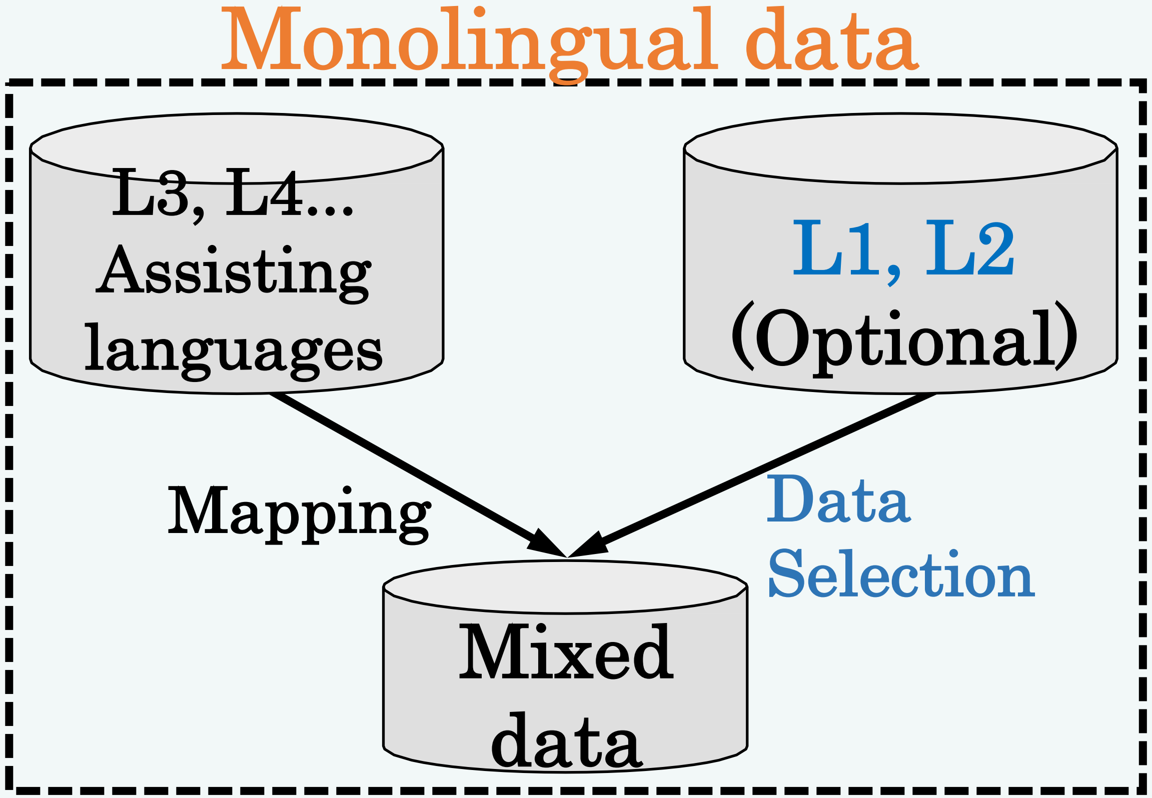
Chinese word		Japanese word (Synthetic)
机构	→	[机構, 机搆, 機構, 機搆]

↑
Japanese LM

Proposed method: Data Selection

Goal:

Reduce difference between train and test



Proposed method: Data Selection

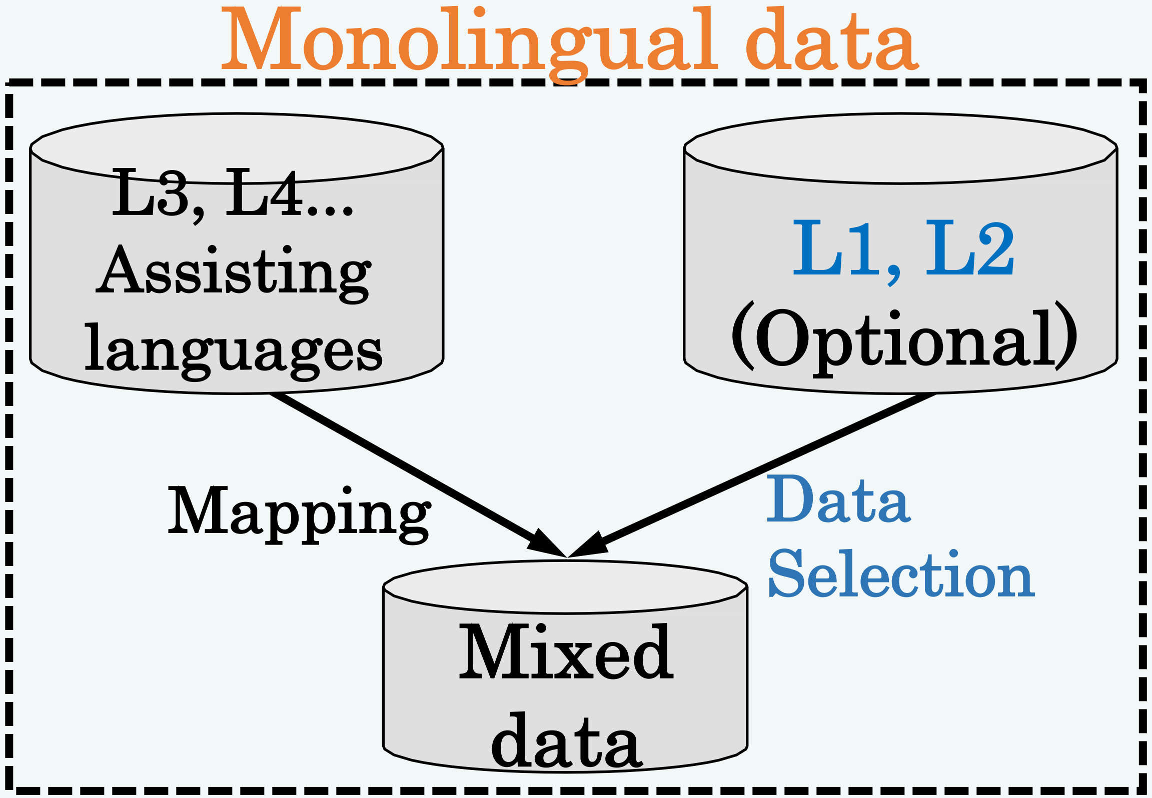
Goal:

Reduce difference between train and test

Method:

Data Selection

Method 1: LM based data selection



Proposed method: Data Selection

Goal:

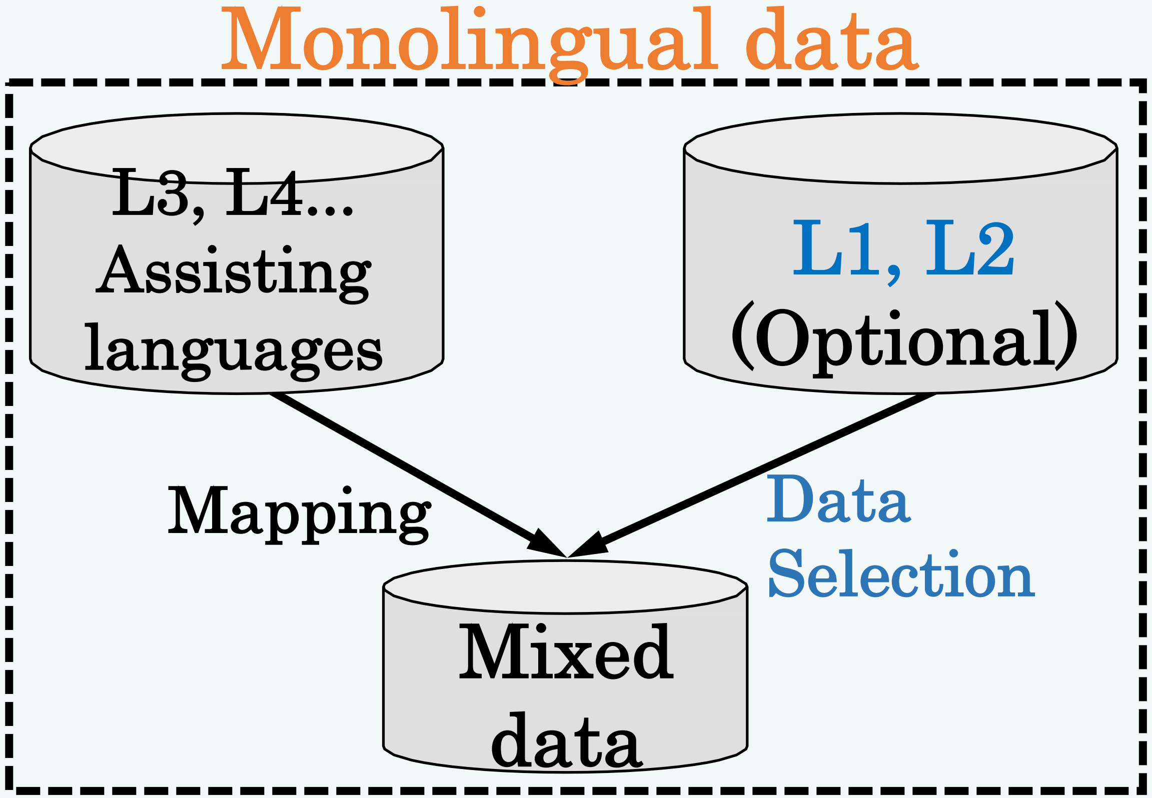
Reduce difference between train and test

Method:

Data Selection

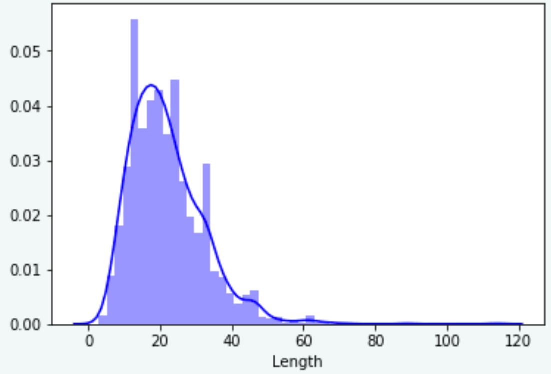
Method 1: LM based data selection

Method 2: Length based data selection

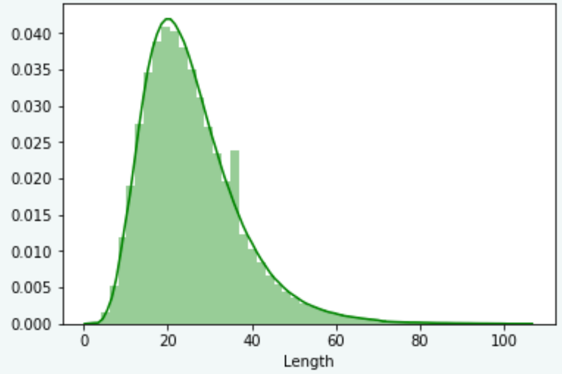


Proposed method: Data Selection

Length Distribution (LD) of target data



LD of train data



Goal:

Reduce difference between train and test

Method:

Data Selection

Method 1: LM based data selection

Method 2: Length based data selection

Proposed method: Data Selection

Goal:

Reduce difference between train and test

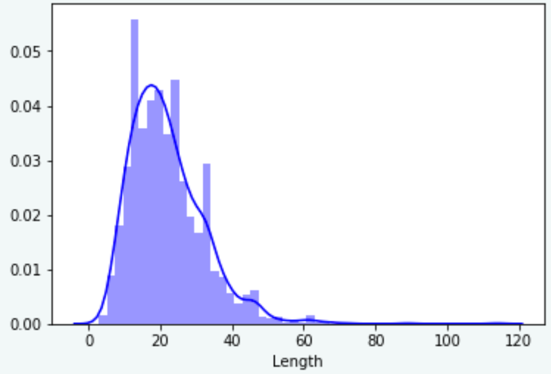
Method:

Data Selection

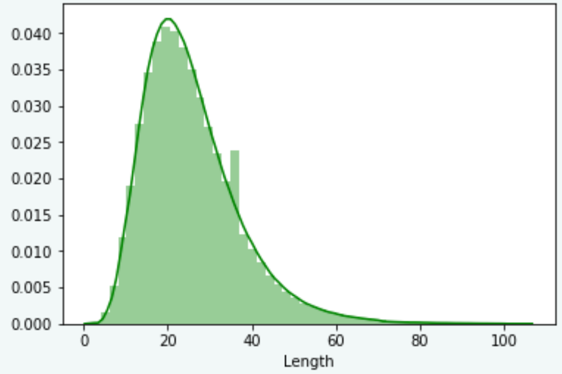
Method 1: LM based data selection

Method 2: Length based data selection

Length Distribution (LD) of target data



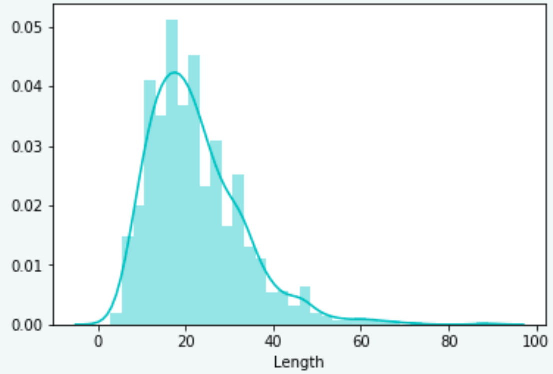
LD of train data



LD

Data selection

LD of selected train data



Proposed method: Data Selection

Goal:

Reduce difference between train and test

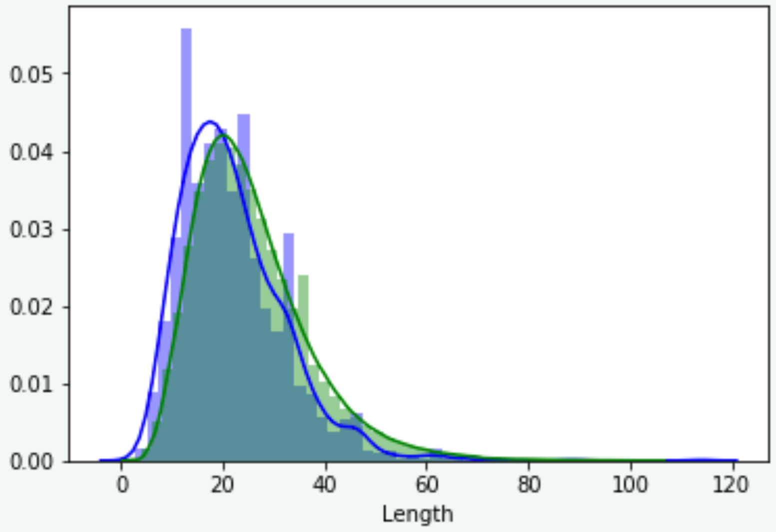
Method:

Data Selection

Method 1: LM based data selection

Method 2: Length based data selection

LD of target data and original train data



Proposed method: Data Selection

Goal:

Reduce difference between train and test

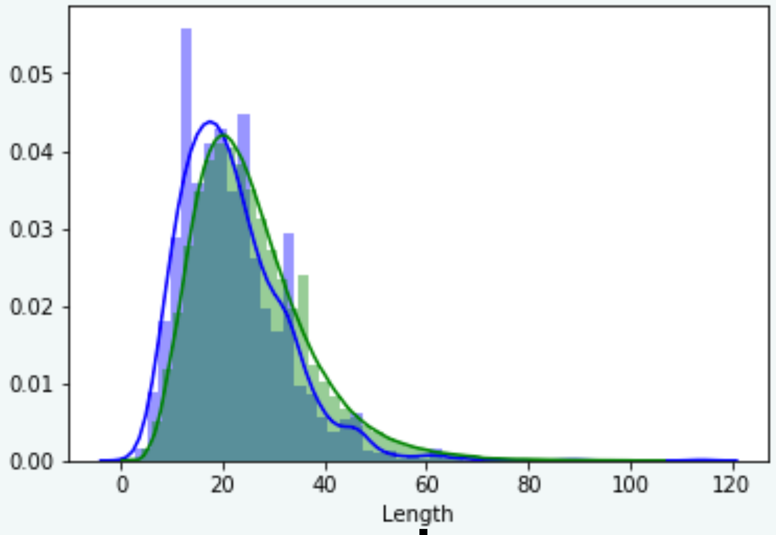
Method:

Data Selection

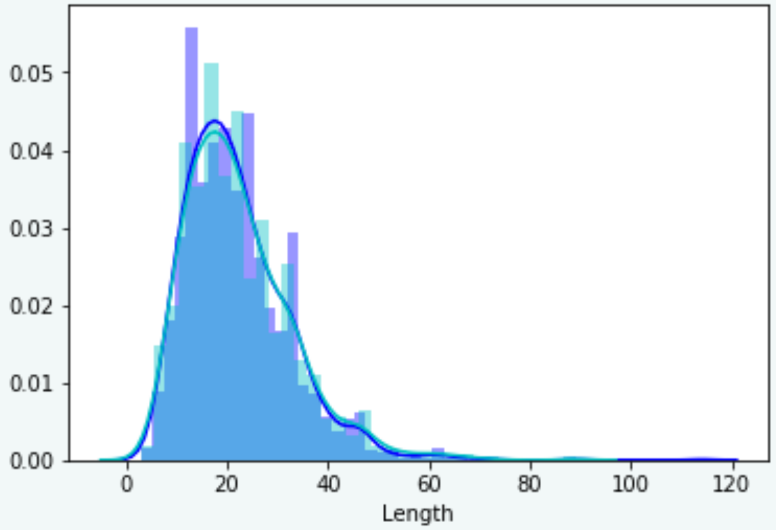
Method 1: LM based data selection

Method 2: Length based data selection

LD of target data and original input data



LD of target data and selected input data

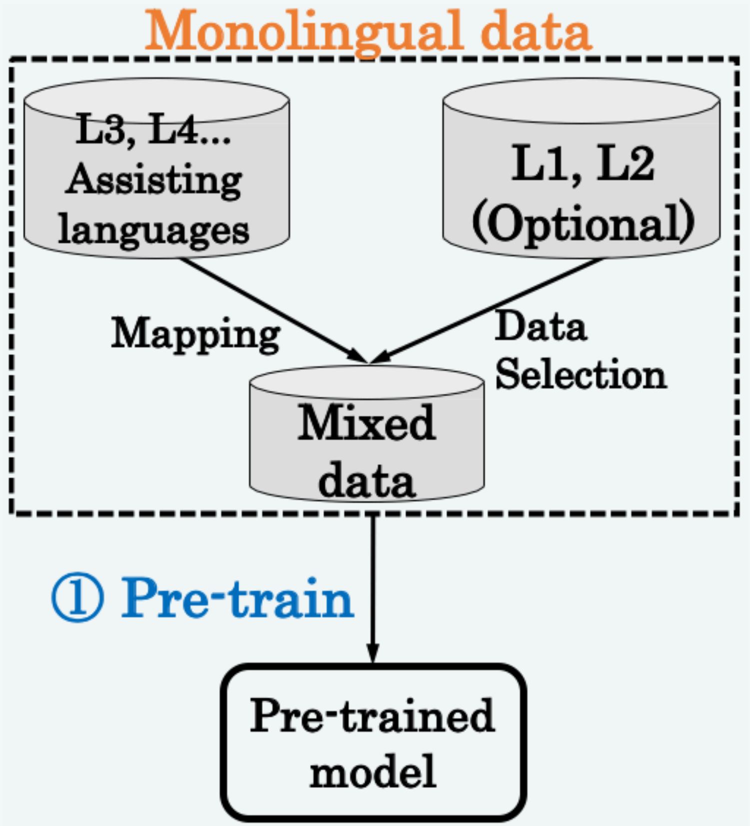


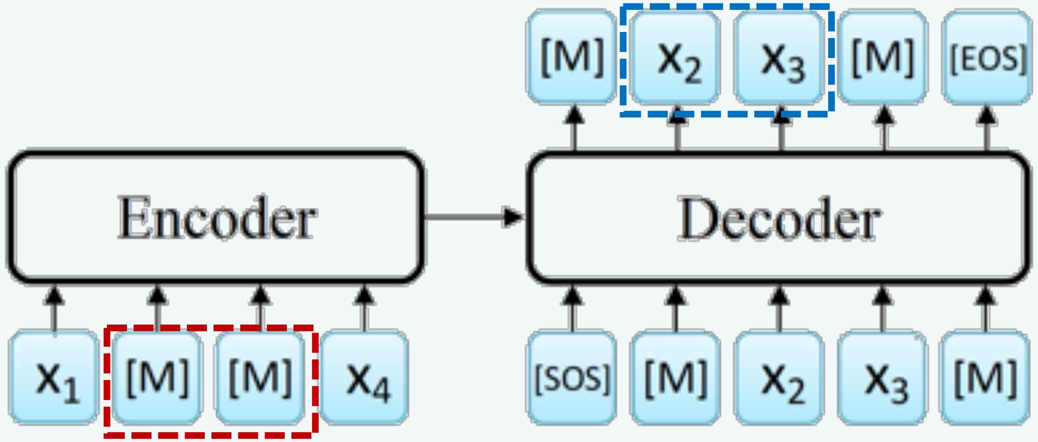
Pre-train: MASS (Song+, 2019)

Method:

Input: Monolingual sentence
with tokens [MASK]ed

Target: [MASK]ed tokens





Experiment settings:

Interested languages:

Japanese and English

Assisting languages:

Chinese, French, Arabic and Russian

Experiment settings:

Interested languages:

Japanese and English

Assisting languages:

Chinese, French, Arabic and Russian

Dataset:

Pre-train:

Ja, En: ASPEC (Nakazawa+, 2016)

Others: Common Crawl*

Fine-tune:

Ja-En: ASPEC(Nakazawa+, 2016)

No overlap with pre-train data

Data for LM: News commentary*

Data pre-processing:

Normalization and filtering

Script mapping for Zh->Ja

KenLM to train LM

*http://data.statmt.org/ngrams/

*http://data.statmt.org/news-commentary/v14/

Experiment settings:

Interested languages:

Japanese and English

Assisting languages:

Chinese, French, Arabic and Russian

Dataset:

Pre-train:

Ja, En: ASPEC (Nakazawa+, 2016)

Others: Common Crawl*

Fine-tune:

Ja-En: ASPEC(Nakazawa+, 2016)

No overlap with pre-train data

Data for LM: News commentary*

Data pre-processing:

Normalization and filtering

Script mapping for Zh->Ja

KenLM to train LM

Train and evaluate:

- Tensor2tensor (Vaswani+, 2018) with 'transformer_big' setting
- Shared vocab of 64k, using SentencePiece (Kuro+, 2018)
- sacreBLEU

*http://data.statmt.org/ngrams/

*http://data.statmt.org/news-commentary/v14/

Results:

#	Pre-training					Fine-tuning							
	Data pre-processing	Zh	Ja	En	Fr	En→Ja				Ja→En			
						3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
B1	1-to-1 Zh→Ja mapping + LM	20M	-	-	-	**5.3**	**14.5**	**20.0**	**26.1**	**3.7**	**11.2**	**15.6**	**20.5**
B2	LM	-	-	-	20M	3.4	9.1	14.9	23.4	2.1	6.3	11.3	17.7
B3	1-to-1 Zh→Ja mapping + LM	20M	-	-	20M	2.1	6.7	12.6	21.9	2.2	6.3	10.7	16.8

1. Extreme Low Resource Situation

   Compared with baseline, using monolingual data from assisting languages helps.

   There may be conflicts between data of different assisting languages.

Results:

#	Pre-training: Data pre-processing	Zh	Ja	En	Fr	Fine-tuning: En→Ja 3K	10K	20K	50K	Ja→En 3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
C1	LD	-	1M	1M	-	7.7	15.8	**20.7**	26.3	7.2	**12.7**	15.7	19.6
C2	1-to-1 Zh→Ja mapping + LD	20M	1M	1M	-	**8.3**	**16.4**	20.2	**26.9**	**7.5**	12.5	**16.3**	**20.7**
C3	LD	-	1M	1M	20M	**8.3**	15.3	19.3	26.7	6.8	12.3	15.4	20.4
C4	1-to-1 Zh→Ja mapping + LD	20M	1M	1M	20M	7.1	15.2	19.4	26.5	6.6	12.0	15.4	19.9

2. Low Resource Situation

   Compared with baseline, using monolingual data from assisting languages helps.

   There may be conflicts between data of different assisting languages.

Results:

#	Pre-training					Fine-tuning							
	Data pre-processing	Zh	Ja	En	Fr	En→Ja				Ja→En			
						3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
D1	LD	-	15M	15M	-	9.6	**17.2**	21.5	**28.0**	**8.6**	**13.5**	**16.8**	**20.9**
D2	1-to-1 Zh→Ja mapping + LD	20M	15M	15M	-	**9.7**	17.1	**21.6**	27.2	8.3	13.3	16.7	20.6
D3	LD	-	15M	15M	20M	7.7	15.0	19.8	26.3	6.3	11.7	15.1	20.2
D4	1-to-1 Zh→Ja mapping + LD	20M	15M	15M	20M	7.7	14.9	19.7	26.1	6.5	11.4	15.4	19.8

3. Rich Resource Situation

Data from assisting languages does not help.

Results:

#	Pre-training					Fine-tuning							
	Data pre-processing	Zh	Ja	En	Fr	En→Ja				Ja→En			
						3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
E1	1-to-1 Zh→Ja mapping	20M	20M	20M	20M	**7.0**	**13.4**	**19.3**	**25.7**	**5.9**	**11.1**	**15.0**	**19.8**
E2	LM-scoring Zh→Ja mapping	20M	20M	20M	20M	6.3	12.7	18.1	24.7	5.7	10.3	13.5	18.9

Mapping:

1-to-1 Zh->Ja mapping is better than many-to-many mapping
Japanese LM cannot directly apply to Chinese mapped data
Segmentation granularity of Chinese and Japanese data is different

Results:

#	Pre-training					Fine-tuning							
						En→Ja				Ja→En			
	Data pre-processing	Zh	Ja	En	Fr	3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
D1	LD	-	15M	15M	-	9.6	**17.2**	21.5	**28.0**	**8.6**	**13.5**	**16.8**	**20.9**
F1	LM-scoring	-	20M	20M	-	4.7	11.7	16.6	23.9	4.5	9.1	12.9	18.3

Data Selection:

Sentence length distribution selection is better than LM score method

Maybe the data used to train the LM is not in-domain.

Results:

#	Pre-training					Fine-tuning							
	Data pre-processing	Zh	Ja	En	Fr	En→Ja				Ja→En			
						3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
F1	LM-scoring	-	20M	20M	-	4.7	11.7	16.6	23.9	4.5	9.1	12.9	18.3
F2	1-to-1 Zh→Ja mapping + LM-scoring	20M	20M	20M	20M	**7.0**	**13.4**	**19.3**	**25.7**	**5.9**	**11.1**	**15.0**	**19.8**
F3	LM-scoring + Ar20M + Ru20M	-	20M	20M	-	4.8	12.1	18.1	25.1	4.4	10.2	13.5	18.9

Different assisting languages:

Similar languages performs better than randomly selected languages

Conclusions:

- Leveraging monolingual data from other languages to improve NMT is possible.

- Script mapping is a good way to improve data similarity thus improve performance.

Future work:

- Explore data selection methods

- Experiments with more challenging language pairs such as Japanese-Russian

Thanks for listening!

Pre-training via Leveraging Assisting Languages for Neural Machine Translation

Haiyue Song[1], Raj Dabre[2], Zhuoyuan Mao[1], Fei Cheng[1], Sadao Kurohashi[1], Eiichiro Sumita[2]
[1]Kyoto University [2]NICT