Investigating Neural Machine Translation Strategies for Tagalog

Final Presentation | Fall 2023
School of Information | UC Berkeley

DATASCI 266 Project Team

Victoria Hollingshead

Neja Jakkinpali

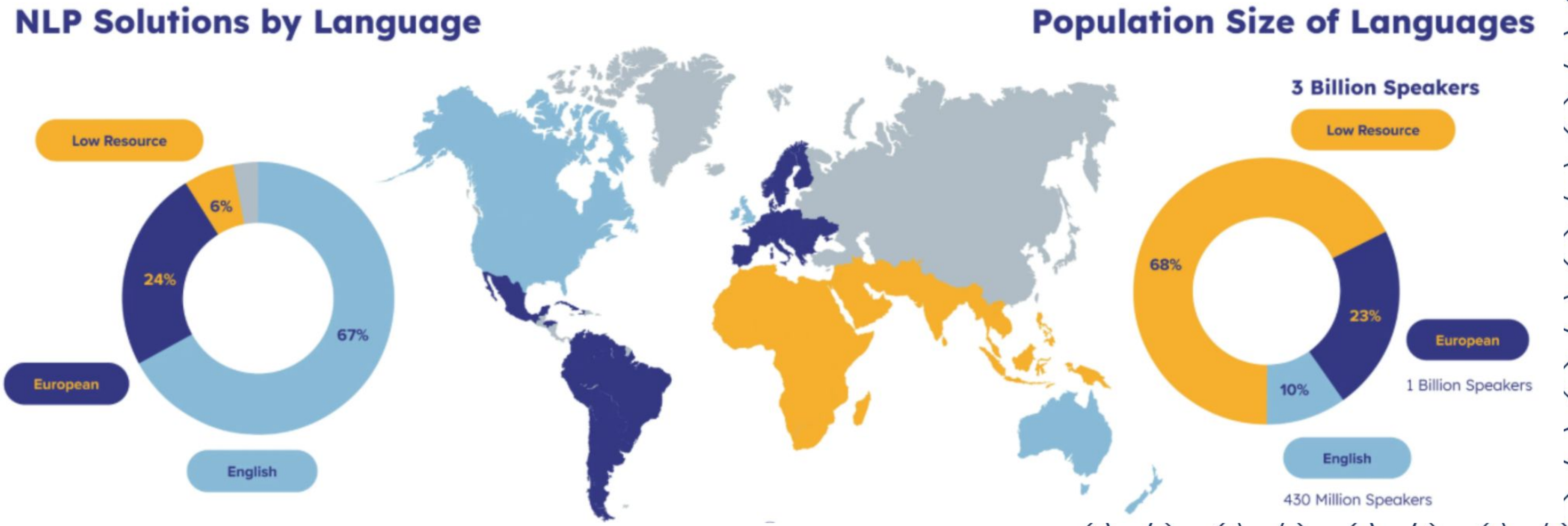
Henry Caldera

Research Questions

1) What is the state-of-the-art model performance for English-to-Tagalog translations?

2) Can model augmentation techniques improve English-to-Tagalog translations?

Low Resource Languages



NLP Solutions by Language

Population Size of Languages

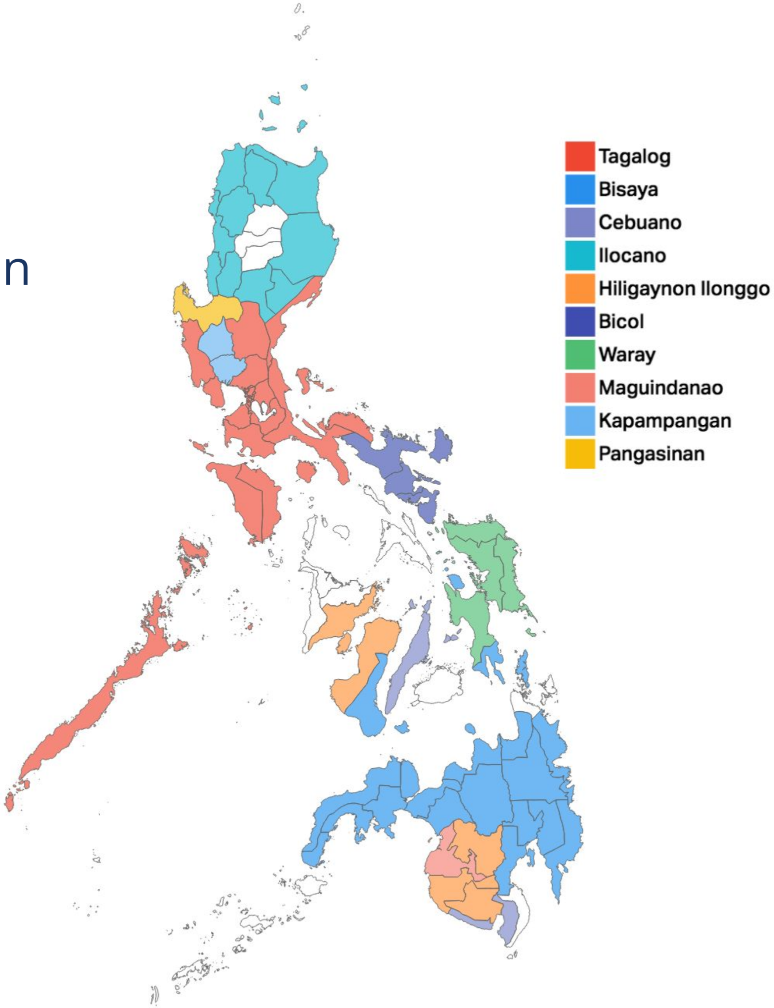
Background

Multilingual Neural Machine Translation (MNMT)

- GPT-3.5 Turbo
- mBART50
- M2M100

Datasets:

- FLORES200
- CulturaX



GPT 3.5 Turbo Methodology

Fine-Tuning with FLORES200

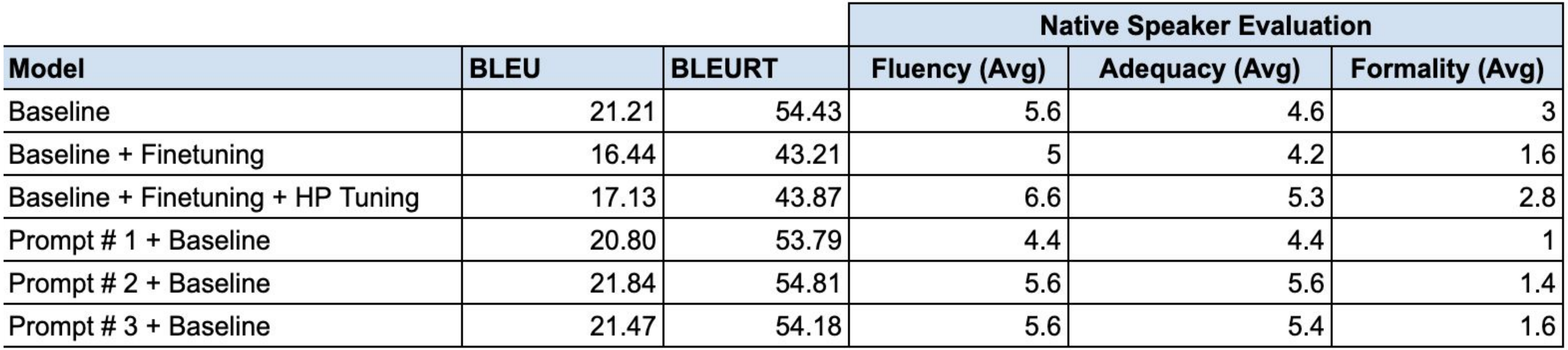
Hyperparameter-Tuning

- *Epochs*
- *Batch Size*
- *Learning Rate Multiplier*

Prompt Engineering

Results & Discussion - GPT 3.5

- **General decrease** in BLEU and BLEURT scores with fine-tuning and hyperparameter tuning applied
- **Prompt engineering** provided marginal improvements over baseline, with the most notable improvement in **adequacy**
- **Relative high** BLEURT and **adequacy** scores compared to other models tested
- GPT 3.5 may benefit from **training on multilingual datasets** comprised of Philippines languages due to their shared linguistic phenology



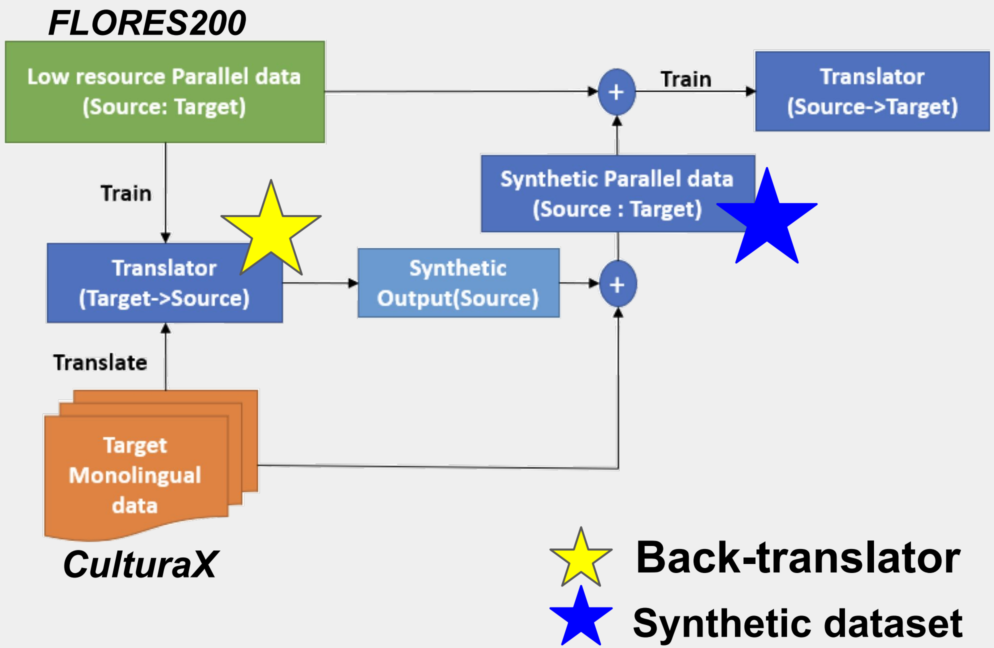
			Native Speaker Evaluation		
Model	**BLEU**	**BLEURT**	**Fluency (Avg)**	**Adequacy (Avg)**	**Formality (Avg)**
Baseline	21.21	54.43	5.6	4.6	3
Baseline + Finetuning	16.44	43.21	5	4.2	1.6
Baseline + Finetuning + HP Tuning	17.13	43.87	6.6	5.3	2.8
Prompt # 1 + Baseline	20.80	53.79	4.4	4.4	1
Prompt # 2 + Baseline	21.84	54.81	5.6	5.6	1.4
Prompt # 3 + Baseline	21.47	54.18	5.6	5.4	1.6

mBART50 & M2M100 Methodology

Setup

- Pre-processing
  - Parsing CulturaX with regular expression (regex).
  - Kept full sentences containing only Latin ASCII characters (hex: x00-x7F)
- Hyperparameters
  - Optimizer: AdamW → RAdam
  - LR: 5e-5 → 2e-4
  - Warmup: 0% → 85% of 1st epoch
  - Batch: 8 → 48
  - Callback: None → Tol: 1e-3, Pat: 5
  - Beam Groups: 0 → 5

Back-Translation Highlights


FLORES200
Low resource Parallel data (Source: Target)
Train
Translator (Source->Target)
Train
Synthetic Parallel data (Source : Target)
Translator (Target->Source)
Synthetic Output(Source)
Translate
Target Monolingual data
CulturaX
Back-translator
Synthetic dataset

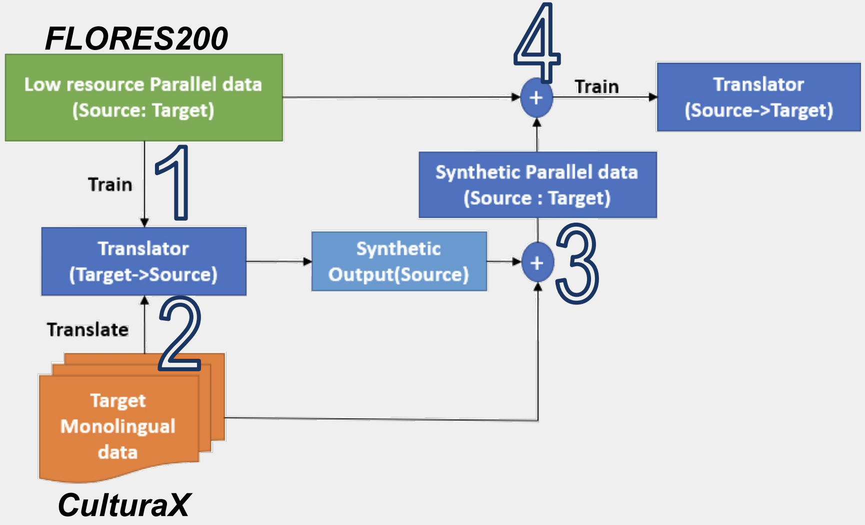
mBART50 & M2M100 Methodology

Train Back-translator → Back-translate → Combine Real+Synthetic → Train on Augmented

Execution

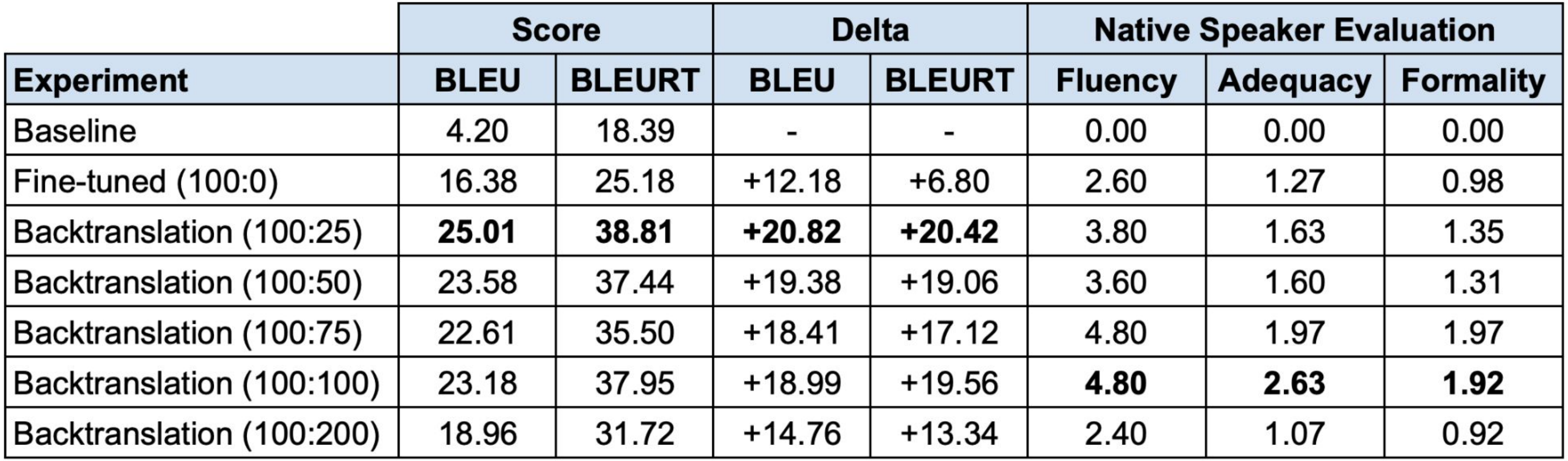
1. Starting with **parallel data**, train baseline model from **target** to **source**
2. Use this **back-translator** to run **inference** on the monolingual **target data**
3. **Combine** the new **synthetic** dataset with the **real** dataset
4. Use the **augmented** dataset to train the baseline model

Back-Translation Schema


FLORES200
Low resource Parallel data (Source: Target)
Train
1
Translator (Target->Source)
Translate
2
Target Monolingual data
CulturaX
Synthetic Output(Source)
+
3
Synthetic Parallel data (Source : Target)
4
+
Train
Translator (Source->Target)

Results - mBART50

- **Training** mBART50 on a **4:1 ratio** of augmented data results in **best BLEU score**.
- **Training** mBART50 on **1:1 ratio** of augmented data results in **best native evaluation**.



	Score		Delta		Native Speaker Evaluation		
Experiment	**BLEU**	**BLEURT**	**BLEU**	**BLEURT**	**Fluency**	**Adequacy**	**Formality**
Baseline	4.20	18.39	-	-	0.00	0.00	0.00
Fine-tuned (100:0)	16.38	25.18	+12.18	+6.80	2.60	1.27	0.98
Backtranslation (100:25)	**25.01**	**38.81**	**+20.82**	**+20.42**	3.80	1.63	1.35
Backtranslation (100:50)	23.58	37.44	+19.38	+19.06	3.60	1.60	1.31
Backtranslation (100:75)	22.61	35.50	+18.41	+17.12	4.80	1.97	1.97
Backtranslation (100:100)	23.18	37.95	+18.99	+19.56	**4.80**	**2.63**	**1.92**
Backtranslation (100:200)	18.96	31.72	+14.76	+13.34	2.40	1.07	0.92

Discussion - mBART50

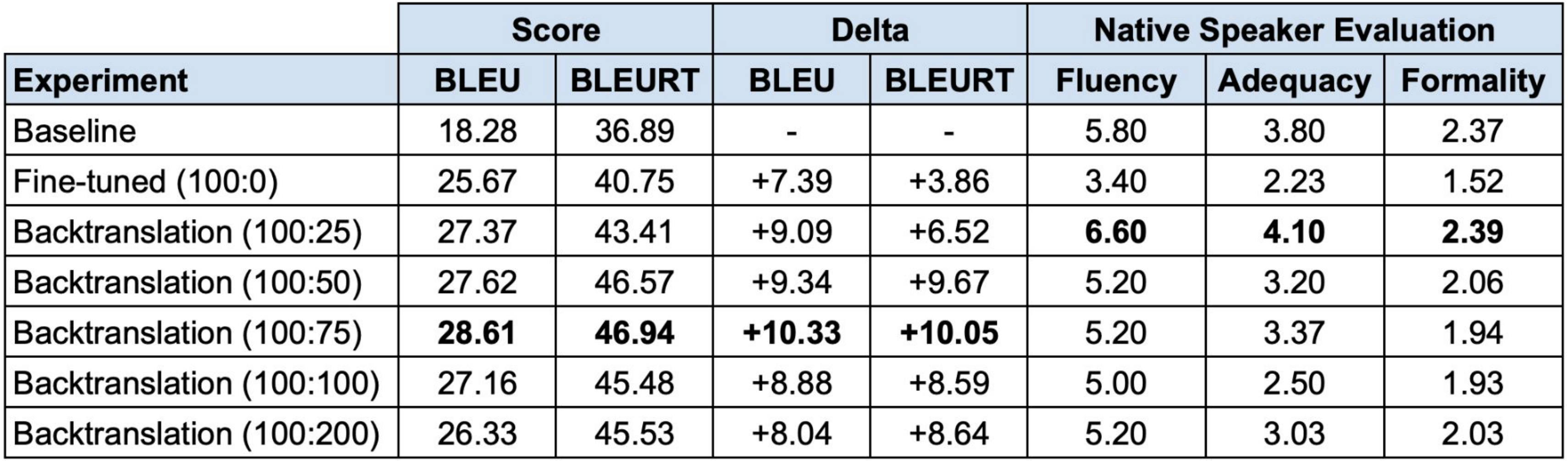
- More augmented data, more accurate translation.
  - ‘Botella’ means bottle in Spanish.

Ground Truth (eng)	`'A bottle fell onto the floor and shattered.'`
Ground Truth (tgl)	`'Bote isang nahulog papunta sa sahig at nabasag.'`
Baseline	**'ini: A bottle fell on the floor and shattered.'**
Back-translation (4:1)	'Ang isang bote ay **tumaklong** sa **ilahok** at **bitigil**.' (A bottle takes refuge in the joint and stops.)
Back-translation (1:1)	'Isang **botella** ay nahulog sa **floor** at **nahulog**.' (A bottle fell on the floor and fell.)

Related language, **Errors**, **Failed to translate**

Results - M2M100

- Training M2M100 on a 4:1 ratio of augmented data results best native speaker evaluation.
- Training M2M100 on 4:3 ratio of augmented data results in best BLEU score.



	Score		Delta		Native Speaker Evaluation		
Experiment	**BLEU**	**BLEURT**	**BLEU**	**BLEURT**	**Fluency**	**Adequacy**	**Formality**
Baseline	18.28	36.89	-	-	5.80	3.80	2.37
Fine-tuned (100:0)	25.67	40.75	+7.39	+3.86	3.40	2.23	1.52
Backtranslation (100:25)	27.37	43.41	+9.09	+6.52	**6.60**	**4.10**	**2.39**
Backtranslation (100:50)	27.62	46.57	+9.34	+9.67	5.20	3.20	2.06
Backtranslation (100:75)	**28.61**	**46.94**	**+10.33**	**+10.05**	5.20	3.37	1.94
Backtranslation (100:100)	27.16	45.48	+8.88	+8.59	5.00	2.50	1.93
Backtranslation (100:200)	26.33	45.53	+8.04	+8.64	5.20	3.03	2.03

Discussion - M2M100

- ‘Butila’ is approximating words meaning bottle, and ‘botol’ means bottle in Malay.

Ground Truth (eng)	'A bottle fell onto the floor and shattered.'
Ground Truth (tgl)	'Bote isang nahulog papunta sa sahig at nabasag.'
Baseline	'Ang isang **bottle** ay **lumabas** sa **floor** at **lumabas**.' (A bottle came off the floor and came out.)
Back-translation (4:1)	'Ang isang **butila** ay **bumabalik** sa **lupa** at **bumabalik**.' (A ~~particle~~ returns to earth and returns.)
Back-translation (4:3)	'Ang isang **botol** ay nahuhulog sa **lupa** at **nanirahan**.' (A bottle falls to the ground and settles.)

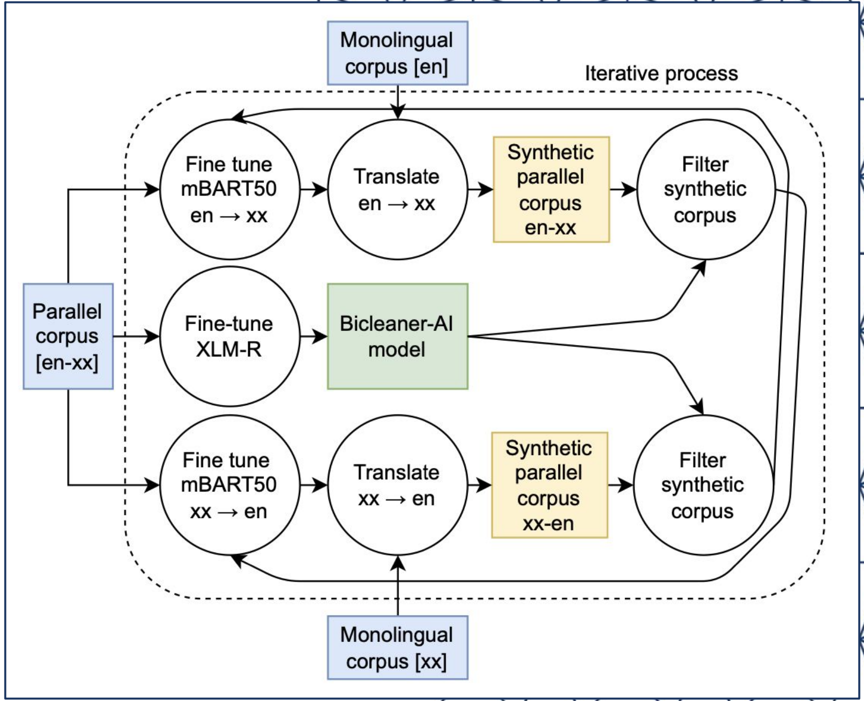
Related language, **Errors**, **Failed to translate**

Conclusion

- Individual improvements
  - **mBART50** has the **largest relative** increase in performance with **BLEU increases** between **300%-500%**.
  - **M2M100** has the **next largest** relative increase in performance with **BLEU increases** between **40%-60%**.
  - **GPT-3.5 Turbo** has the **smallest relative** increase in performance with **BLEU changes** between **-23% to +3%**.
- Although **M2M100** attained the **highest BLEU scores**, **GPT-3.5** turbo was able to reach the **highest native evaluation scores** of 5.6/7 in **fluency** and **adequacy**. This corroborates literature findings that **BLEURT scores** are most **indicative** of NMT **performance**.
- Regarding **decoder-encoder models**, those with the **strongest baseline scores** stand to gain the most from the back-translation **augmentation technique** because the back-translations that make up the **synthetic** dataset will have **higher quality**.

Next Steps

- **More resources**, more training
  - Training on **45K rows of data** for 4 epochs would take **nearly 3 days** to train on a **A100 GPU**.
- Use an **intermediary** model to **clean back-translations** before adding them to **synthetic** dataset.
  - Improved schema shown.


Monolingual corpus [en]
Iterative process
Fine tune mBART50 en → xx
Translate en → xx
Synthetic parallel corpus en-xx
Filter synthetic corpus
Parallel corpus [en-xx]
Fine-tune XLM-R
Bicleaner-AI model
Fine tune mBART50 xx → en
Translate xx → en
Synthetic parallel corpus xx-en
Filter synthetic corpus
Monolingual corpus [xx]