Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin

Pin-Jie Lin*[1,2], Muhammad Saeed*[1] , Ernie Chang*[3], Merel Scholman[2,4]

[1]Saarland Informatics Campus, Germany, [2]Language Science and Technology, Saarland University, Germany, [3]Reality Labs, Meta Inc., [4]ILS, Utrecht University, the Netherlands

{pinjie, musaeed, m.c.j.scholman}@lst.uni-saarland.de, erniecyc@meta.com

Motivation

Nigerian Pidgin

- **Over 75 million speakers**
- **Low-resource language**

Contributions:

- Enrich the existing parallel and monolingual datasets to generate a high-quality corpus collection across **10 resources** and **5 domains**
- Two supplementary training approaches for adapting the model to new language and task before fine-tuning on downstream tasks

Data Resources

Table 1: **Overview of Pidgin datasets.**

Corpus	Language	|Train|	Domain
	Parallel		
Bible	En., Pg.	**29,737**	religious
JW300	En., Pg.	**20,218**	religious
Naija Treebank	En., Pg.	**9,240**	misc.
	Monolingual		
NaijaSenti	Pg.	**8,524**	social media
Afri-BERTa	Pg.	**176,843**	news, misc.
BBC Pidgin	Pg.	**4,147**	news
ASR	Pg.	**7,958**	news
PidginUNMT	Pg.	**5,397**	news
IWSLT'15	En.	**143,609**	wiki., misc.
WMT14-En	En.	**4,468,840**	news

- We generated **5 million** synthetic sentence pairs using a transformer-based model, utilizing all available monolingual data
- We released the parallel and synthetic data collection (QR Code)

Monolingual Case: Sentiment Analysis

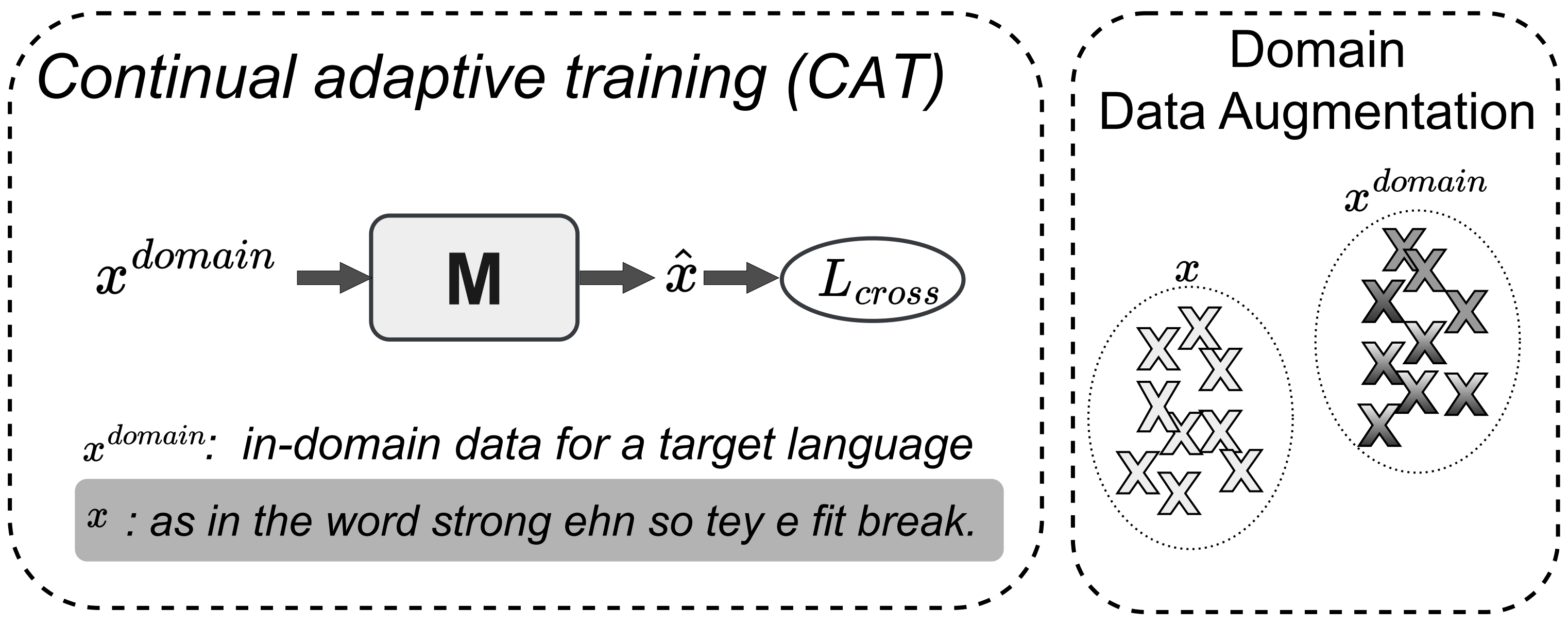
Dataset: NaijaSenti (6.7K/0.6K/1.2K)

Baselines: We set two baselines **Init** where the weights are randomly initialized and fine-tuning (**FT**) which directly transfers the pre-trained language model

Our Approach:

- **C**ontinual **A**daptive **T**raining(**CaT**) provides supplementary training for adapting a model M to a new language via the unlabeled Pidgin corpus
- Continually train BERT and RoBERTA on Bible

Figure 1



x^{domain}: *in-domain data for a target language*

x *: as in the word strong ehn so tey e fit break.*

CaT on monolingual data enables significant performance gains

Table 2: **Results of sentiment classification.**

Model Type	Init	FT	CaT
BERT	71.8	79.7	**80.7**
RoBERTa	68.4	80.1	**82.5**

Parallel & Synthetic Case: Machine Translation

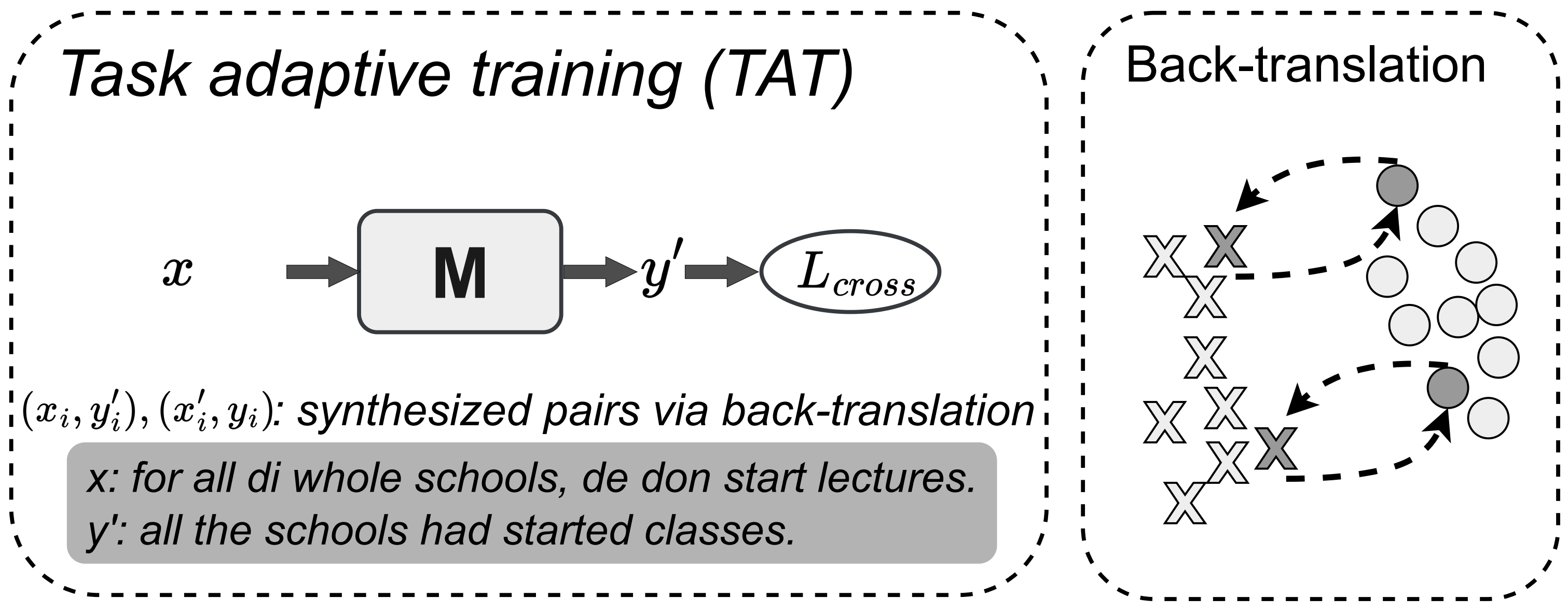
Dataset: JW300 (20K/1K/2.1K)

Baselines: Transformer - FT

Our Approach:

- Data Aug.: data augmentation with Bible
- **TaT** (**T**ask **A**daptive **T**raining) allows the model to adapt to the translation task through synthesized pairs

Figure 2



$(x_i, y_i'), (x_i', y_i)$: *synthesized pairs via back-translation*

x: for all di whole schools, de don start lectures.
y': all the schools had started classes.

TaT yields further improvement on the translation quality

- Data Aug. significantly improves the baseline's performance by **6.45** and **15.76** points
- TaT on synthetic data leads to noticeable improvements in translation coherence, showing enhancements of **1.69** and **2.28**

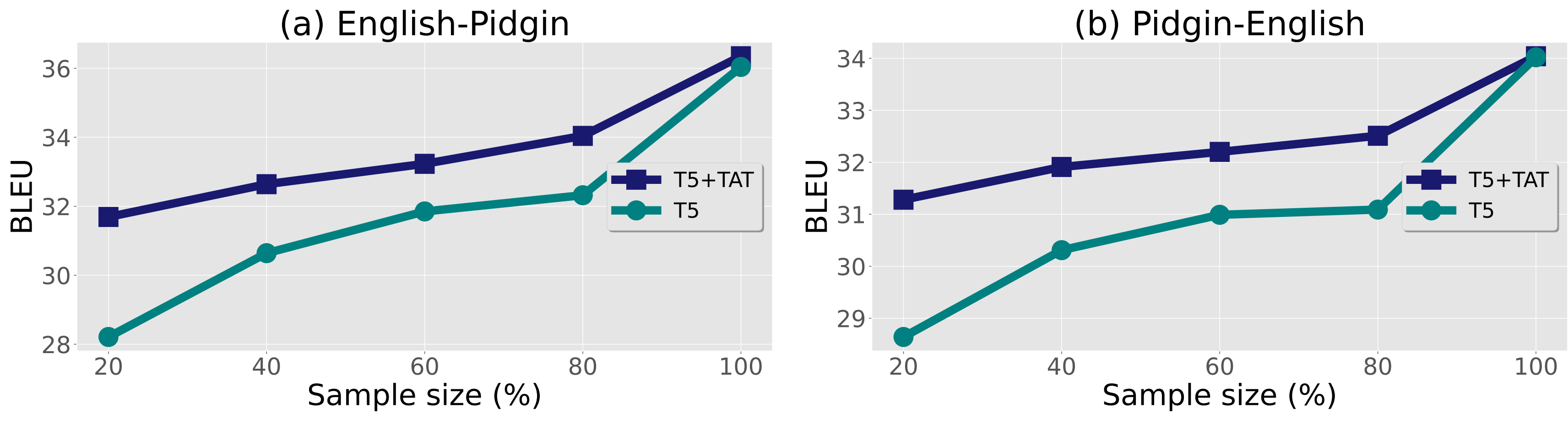
Table 3: **BPE Results on JW300 translation benchmark. BLEU is reported**

	English-Pidgin	Pidgin-English
Transformer - FT	24.29	13
Transformer - FT+ Data Aug.	30.74	28.76
Transformer - FT + Data Aug.+TaT	**32.43**	**31.04**

Low-data setting

- Obtaining strong performance by **+3.48** and **+2.64** BLEU improvement for Pidgin-English and English-Pidgin respectively when only 20% of the data is available for training

Figure 3: BLEU scores on **20**%, **40**%, **60**%, **80**% of sample size and full dataset



Q: Are multilingual models better for low-resource language?

- English model outperforms the multilingual counterparts by a large margin of **2.14** and **1.27** points
- Nigerian Pidgin is English lexified

Table 4: **Results on JW300 translation benchmark using T5 and mT5.**

Model Type	English-Pidgin	Pidgin-English
Data Aug.⋆		
mT5 (base)	33.92	32.75
T5 (base)	**36.04**	**34.02**

Conclusions

- Largest English-Pidgin corpus, performed large-scale data augmentation, and proposed a framework of cross-lingual adaptive training for low-resource language
- Surprisingly, our studies show that English-based models outperforms multilingual models and significantly improves model performance
- **Future work**: challenge of orthographic variations in Pidgin

https://github.com/muhammed-saeed/CLaT