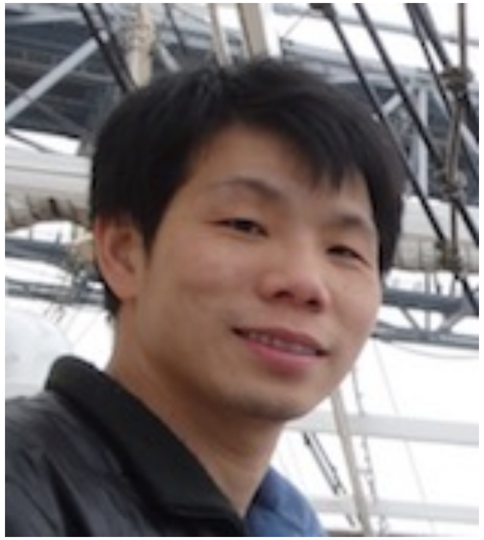
Approaching Sign Language Gloss Translation as a Low-Resource Machine Translation Task

AT4SSL @ MTSummit2021



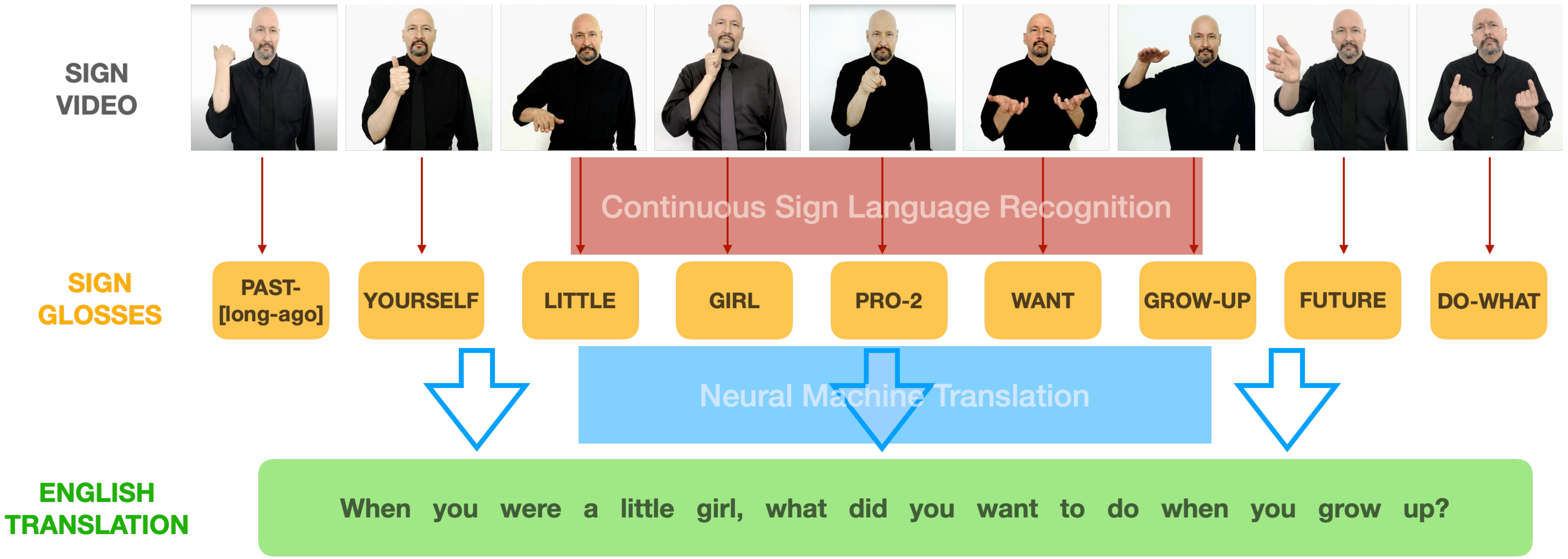
Xuan Zhang

Kevin Duh

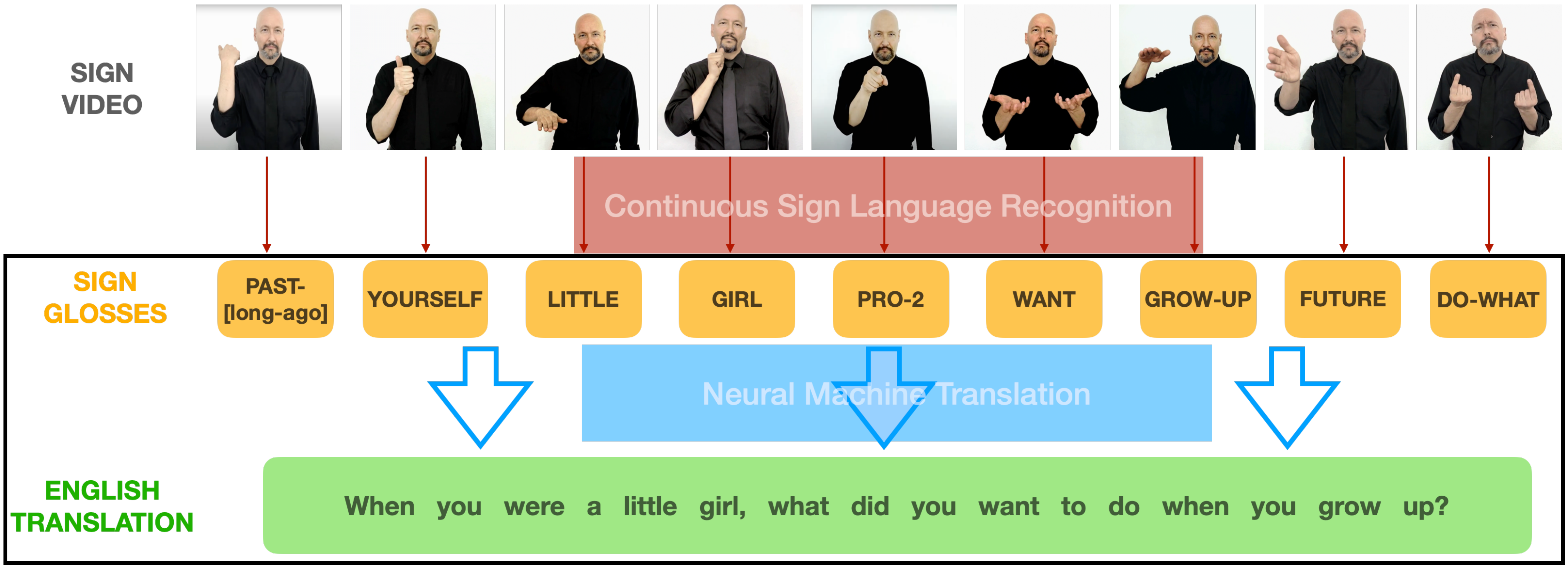
Johns Hopkins Whiting School of Engineering

Department of Computer Science,
Johns Hopkins University

Cascaded Sign Language Translation (SLT) System

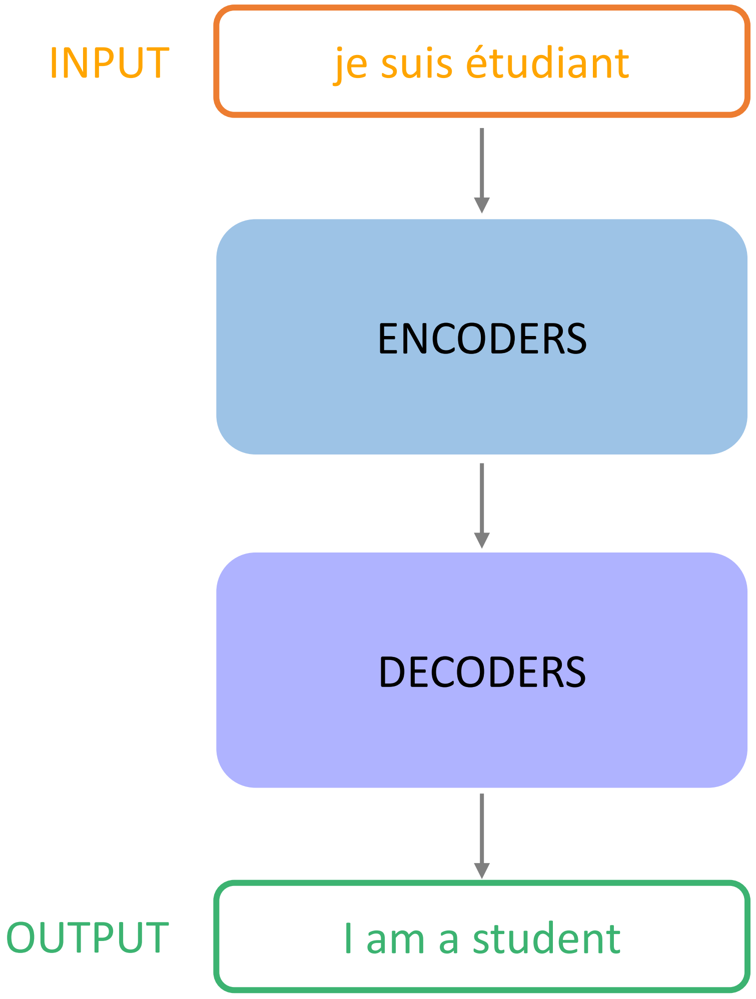


Cascaded Sign Language Translation (SLT) System

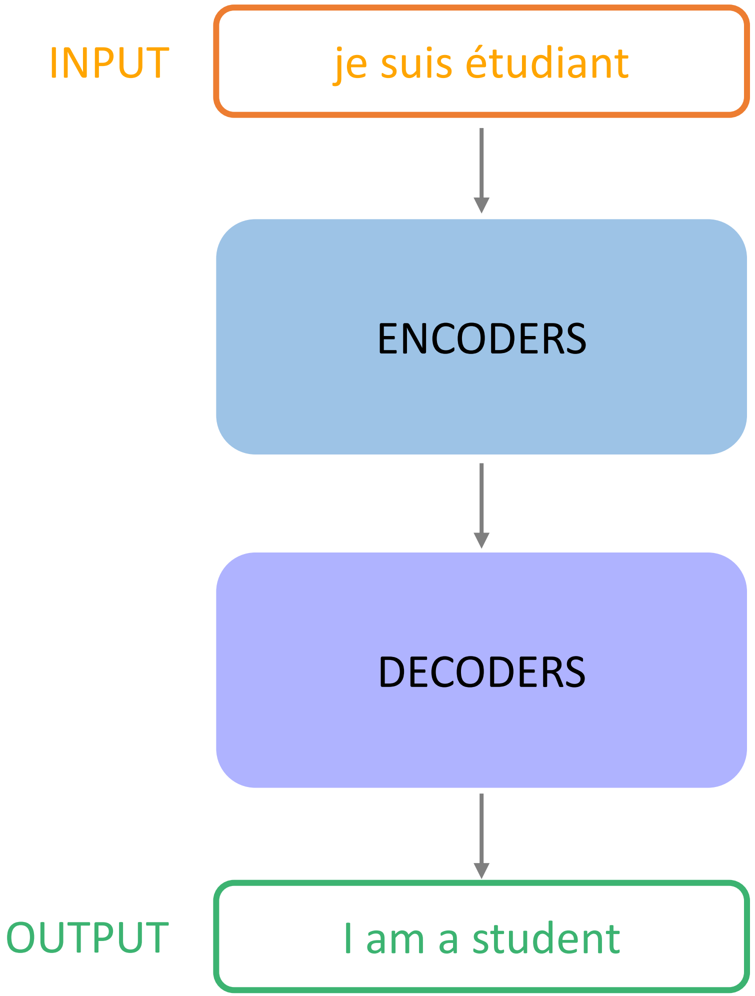


this work

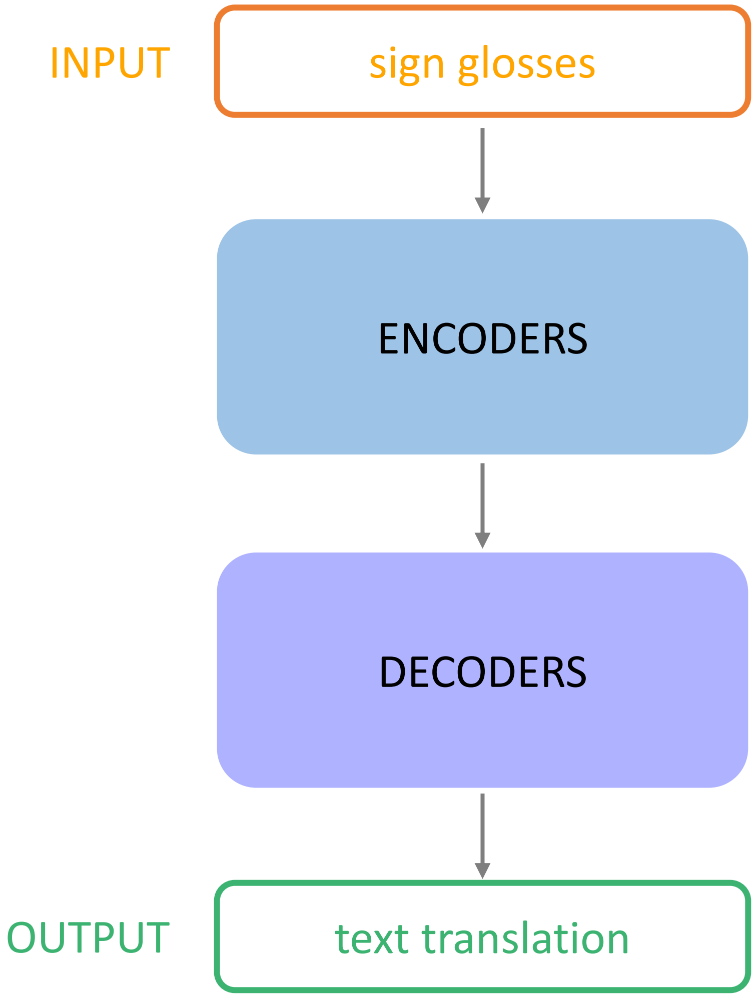
Neural Machine Translation (NMT)



Neural Machine Translation (NMT)



this work:



Challenges of NMT in SLT

- **NMT systems are data-hungry** and usually require millions of training examples to obtain a good translation performance.

- The publicly available annotated parallel **sign gloss–text data are scarce.** The popular continuous SLT dataset, "RWTH-PHOENIX-Weather2014" contains only 7,096 gloss-text examples in training set.

Our Solution:

Approaching gloss-text translation as a **Low-Resource** NMT Task

Low-resource NMT Techniques

Hyperparameter Search

Hyperparameter Search is crucial especially for the low-resource scenarios.

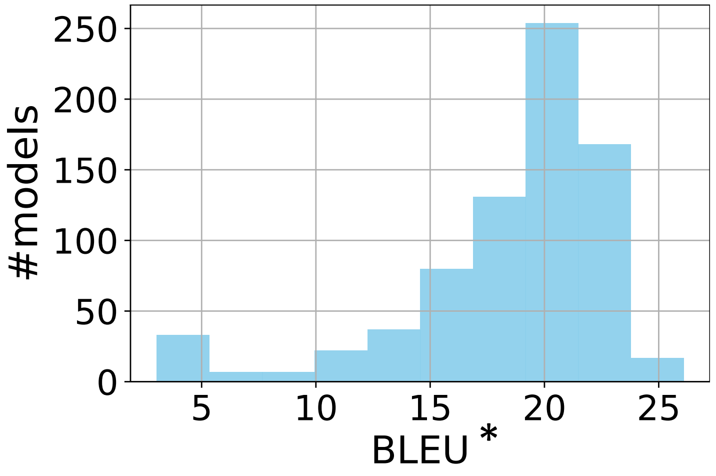


Figure*: Wide variance in performance of a low-resource NMT system with different hyperparameter settings.

Back-translation

Back-translation leverages the abundant monolingual data to enhance the translation performance.

*BLEU: The most commonly used scoring tool to evaluate the machine translation performance. Higher is better.

*Figure borrowed from *Reproducible and Efficient Benchmarks for Hyperparameter Optimization of Neural Machine Translation Systems, Zhang and Duh, TACL 2020*.

Experiment Setup

- **Parallel Data:** RWTH-PHOENIX-Weather 2014T

  - records the weather forecast airings of the German public tv-station PHOENIX.
  - a continuous SLT corpus with both gloss annotations and German translations.
  - train/dev/test: 7,096/519/642 sentences
  - vocabulary size: gloss 1,066; German translation 2,887

- **Monolingual Data** (for back-translation)**:** TED Talks

  - German TED Talk subtitles
  - contains 151,627 sentences

- **NMT model:** Transformer

Experiment: Hyperparameter Search

Search space:

Hyperparameter	Settings
BPE* merge operations	1k, 2k
Number of layers	1, 2, 4
Embedding size	256, 512
Initial learning rate	5e-5, 2e-4, 5e-4

BPE: a word segmentation approach that combines frequent sequence of characters so that out-of-vocabulary words are handled.

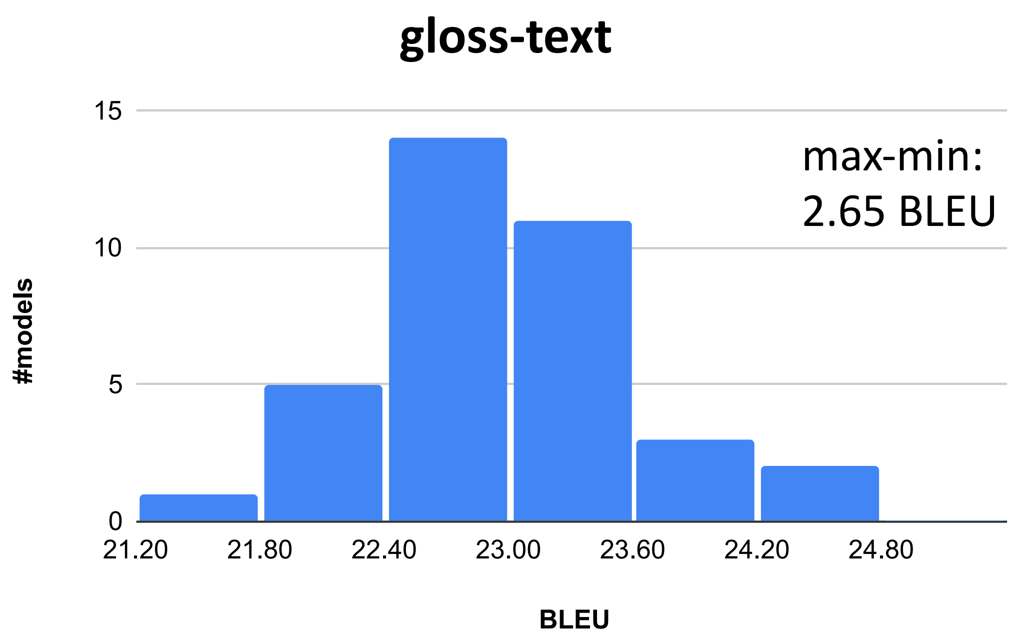
Experiment: Hyperparameter Search

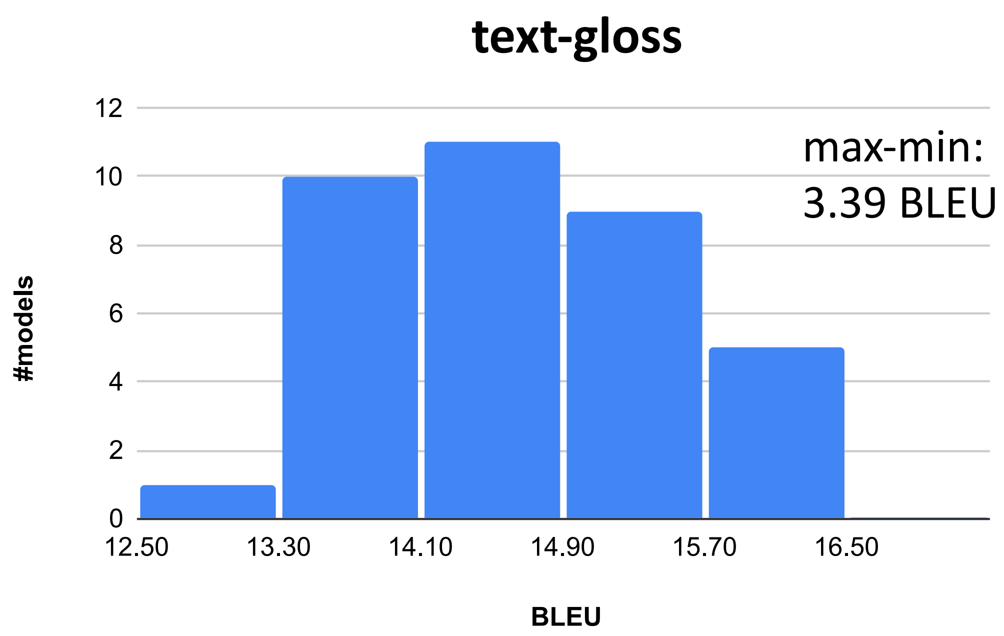
Search space:

Hyperparameter	Settings
BPE merge operations	1k, 2k
Number of layers	1, 2, 4
Embedding size	256, 512
Initial learning rate	5e-5, 2e-4, 5e-4

Compare to existing work:

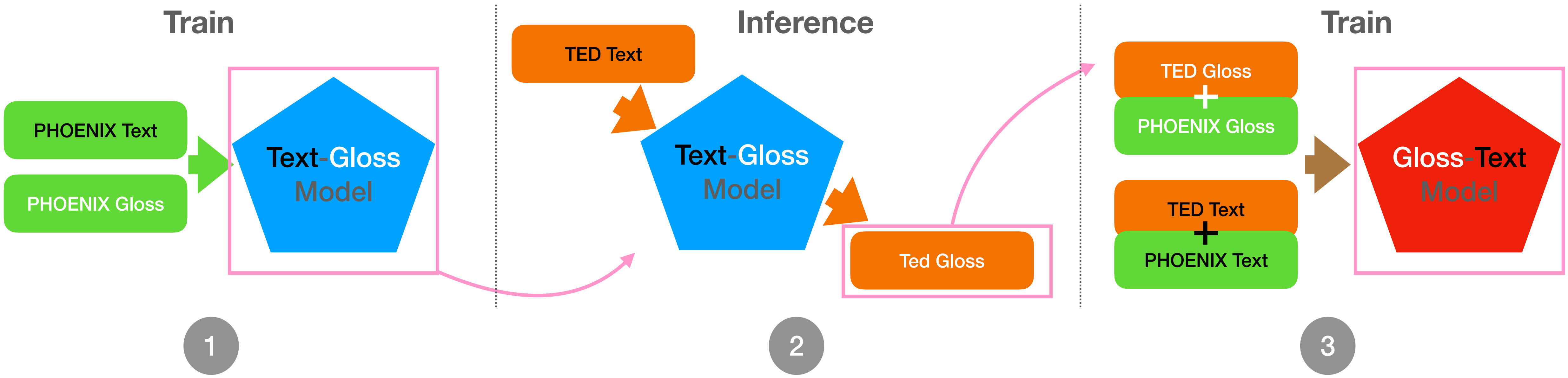
Gloss-text system	Best BLEU
This work	24.38
Camagoz et al. (2020)	24.54
Yin and Read (2020)	24.9





Experiment: Back-translation

Back-translation workflow:



Experiment: Back-translation

One issue: PHOENIX14T and TED data are from different domains, which makes the translation task more challenging.

Domain adaptation leverages out-of-domain data to improve the domain-specific translation. We adopt two domain adaptation methods to aid back-translation.

Experiment: Back-translation

One issue: PHOENIX14T and TED data are from different domains, which makes the translation task more challenging.

Domain adaptation leverages out-of-domain data to improve the domain-specific translation. We adopt two domain adaptation methods to aid back-translation.

- **Data Selection**
- **Fine-tuning**

Experiment: Back-translation

Domain Adaptation – Data Selection

I: in-domain data (PHEONIX14T)
N: out-of-domain data (TED Talks)

Goal: select top n training examples from N which are most similar to I.

Experiment: Back-translation

Domain Adaptation – Data Selection

I: in-domain data (PHEONIX14T)
N: out-of-domain data (TED Talks)

Goal: select top n training examples from N which are most similar to I**.**

Each sentence s in N is assigned a score:

$$H_I(s) - H_N(s)$$

$H_I(s)$: the per-word cross-entropy of s according to a language model trained on I.

Experiment: Back-translation

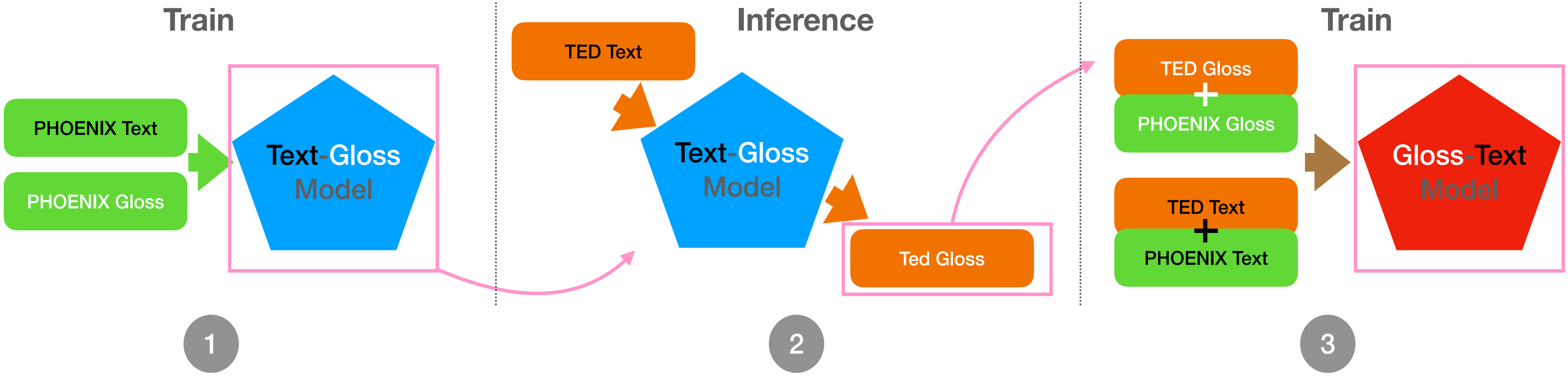
Domain Adaptation – Fine-tuning

1. First, train on the mix of out-of-domain and in-domain data till convergence.

2. Then, continue training or fine-tuning on the in-domain data.

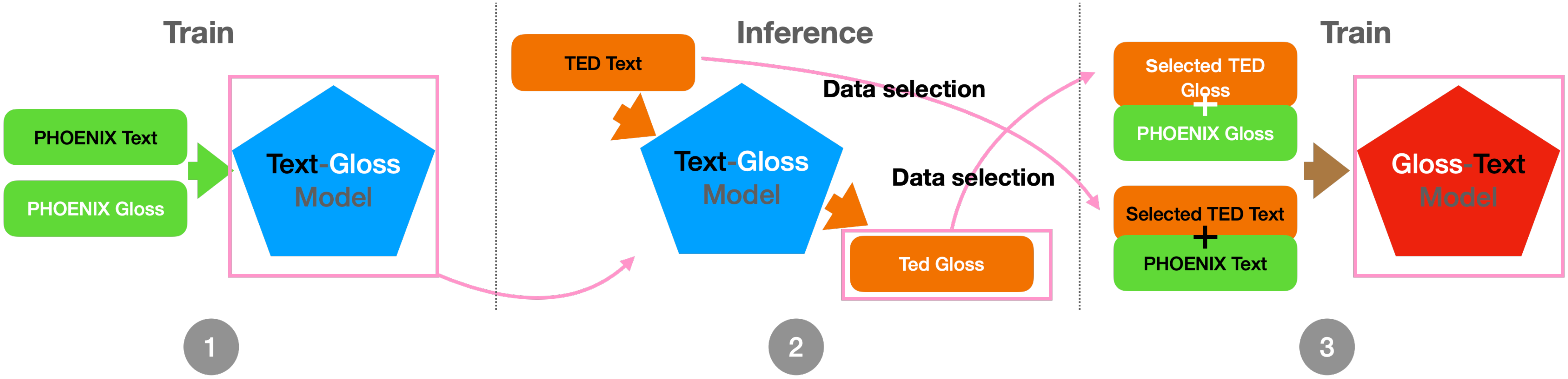
Experiment: Back-translation

Back-translation workflow review:



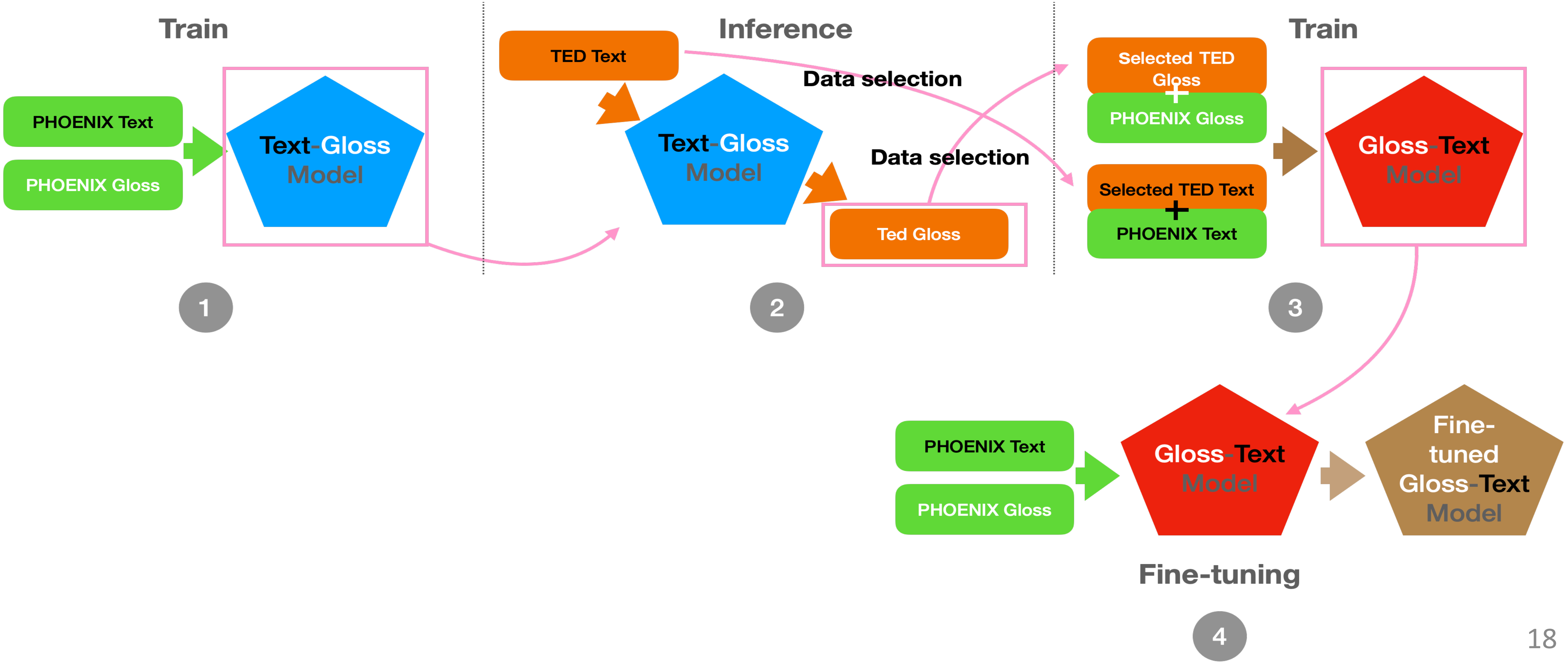
Experiment: Back-translation

Back-translation + data selection workflow:



Experiment: Back-translation

Back-translation + data selection + fine-tuning workflow:



Experiment: Back-translation

Figure: Performance of NMT systems on gloss-text translation.
Systems vary in whether fine-tuned on PHOENIX14T (bt vs. ft), the size of selected TED data (10k, 50k, 100k) and the number of copies of PHOENIX14T added into the training data (0, 1, 2, 5, 10).

Experiment: Back-translation

issue: The best BLEU score of back-translation + domain adaptation is slightly lower than that of hyperparameter search (24.31 vs. 24.38).

possible cause: the domain adaptation techniques fail to overcome the side-effect of introducing the out-of-domain data into training.

Experiment: Back-translation

incorporating in-domain monolingual data

issue: The best BLEU score of back-translation + domain adaptation is slightly lower than that of hyperparameter search (24.31 vs. 24.38).

possible cause: the domain adaptation techniques fail to overcome the side-effect of introducing the out-of-domain data into training.

We thus consider a simpler situation in order to evaluate back-translation:
incorporating in-domain monolingual data.

- Divide PHOENIX14T into two sets, with the first half acting as parallel data, the second as monolingual data containing only the German text side of the data.

Experiment: Back-translation

incorporating in-domain monolingual data

NMT Systems	BLEU
w/o back-translation	19.13
w/ back-translation	21.57

Back-translation has a good potential to improve the translation performance when in-domain monolingual data are available.

Summary

- The translation between sign language glosses and written languages is a challenging task in SLT.
- The obstacle lies in the sparsity of parallel data.
- Approaching the translation task with low-resource MT techniques like hyperparameter search and back-translation is promising.
- Back-translation is more likely to contribute when abundant in-domain monolingual data are available.
- We urge the sign language processing community to put in extra efforts in creating more annotated parallel data.

Q & A