LREC-COLING 2024

Advancing Semi-Supervised Learning for Automatic Post-Editing: Data-Synthesis by Mask-Infilling with Erroneous Terms

Wonkee Lee, Seong-Hwan Heo, Jong-Hyoek Lee

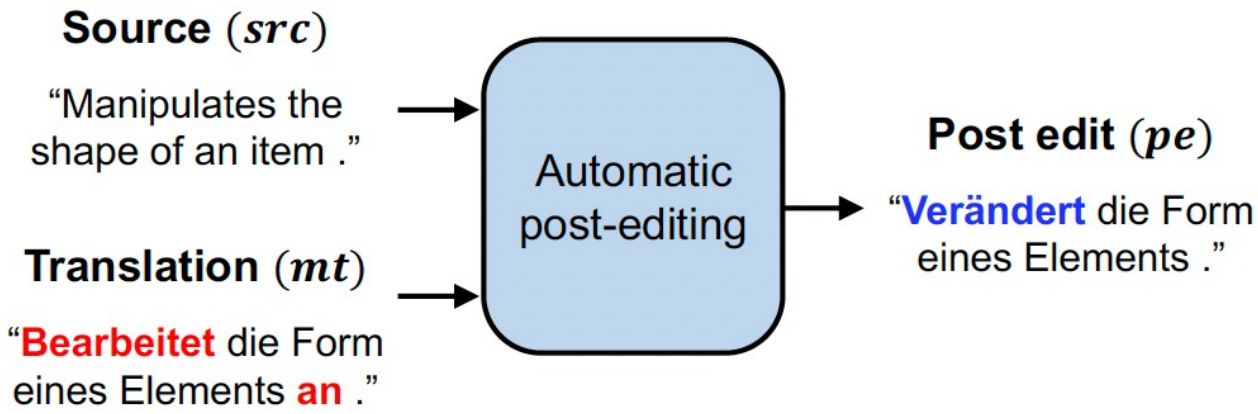
LG AI Research

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Introduction: Overview

- Challenges in Automatic Post-Editing (APE):
  - The primary bottleneck is the scarcity of "gold-standard" human-edited datasets.
- The Role of Semi-Supervised Learning:
  - Use of synthetic datasets for training, offering a promising solution when training data is limited.
- Suggestion of two methodologies for data-synthesis:
  - **MLM Noising**: Leveraging MLM method to simulate the inaccuracies of raw machine translated outputs.
  - **Selective Corpus Interleaving**: Integration of two distinct synthetic datasets by selecting samples that closely resemble the gold data.

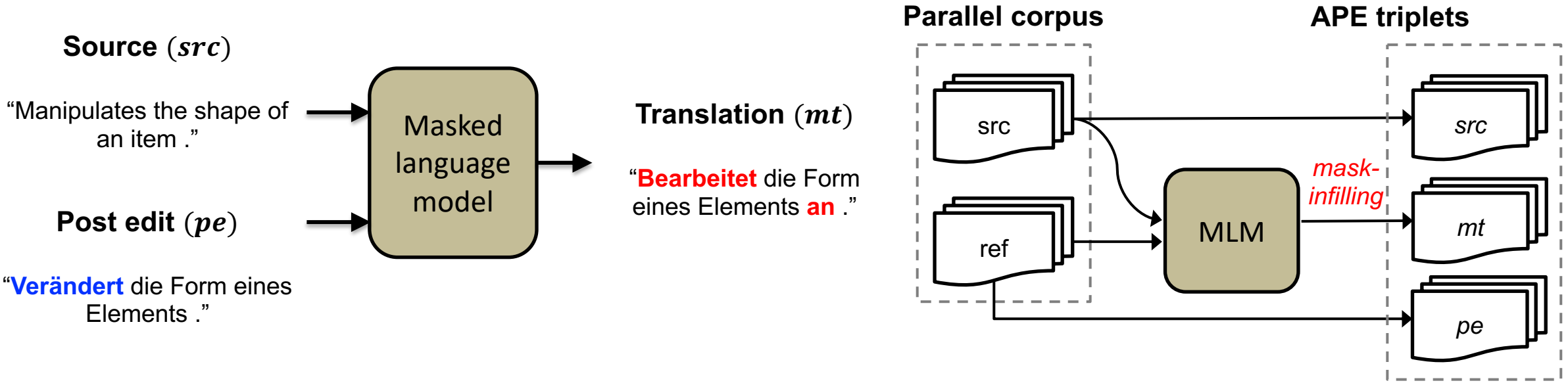
Introduction: Automatic Post Editing



- Multi-source sequence-to-sequence problem
  - $(src, mt) \rightarrow pe$
- APE data (Underlying assumption)
  - *src* and *pe* should be **error-free** and **semantically equivalent** to each other.
  - *pe* is the **minimal correction** of *mt*.

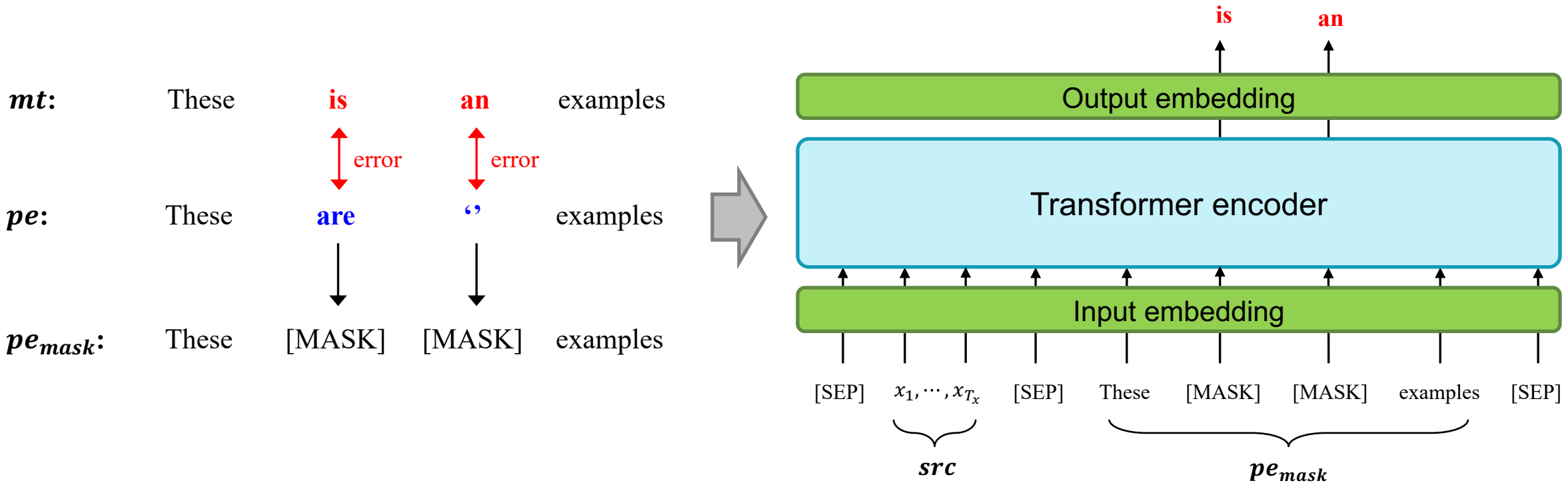
Method

- MLM Noising
  - ✓ Employ masked language model (MLM) to replicate translation errors observable in (src, pe) pairs: $(src, pe) \rightarrow mt$.
  - ✓ Apply mask-infilling of MLM to parallel texts (src, ref), yielding our synthetic dataset.
  - ✓ Additionally, process synthetic data (TRANS) for training the MLM noising model.



Method: MLM Noising

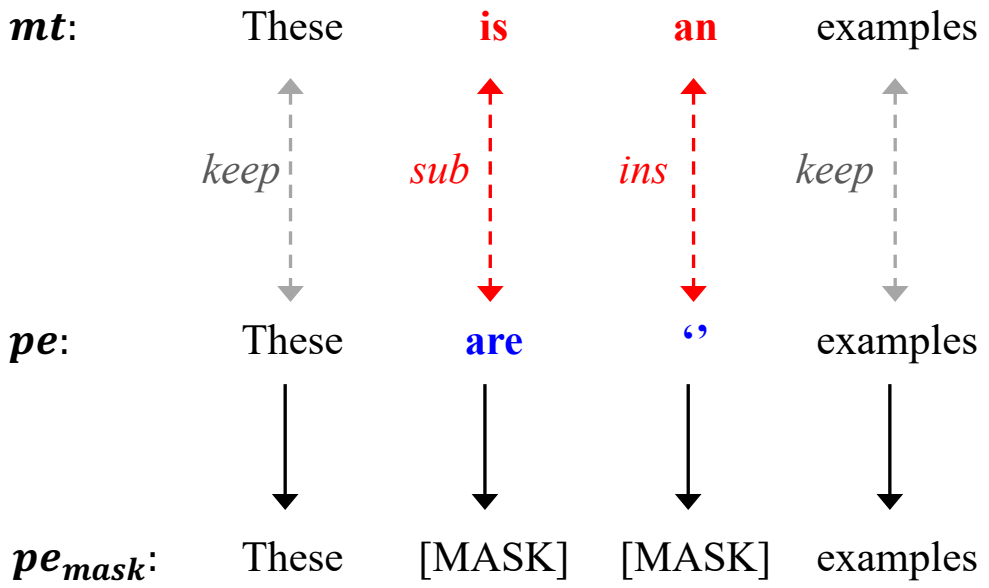
1. Mask *pe* tokens that are aligned to **mistranslated** *mt* tokens ➔ pe_{mask}
2. Provide the **Transformer encoder** with (*src*, pe_{mask}) pairs, and let the model learn to predict the erroneous *mt* token from [MASK]



Method: MLM Noising

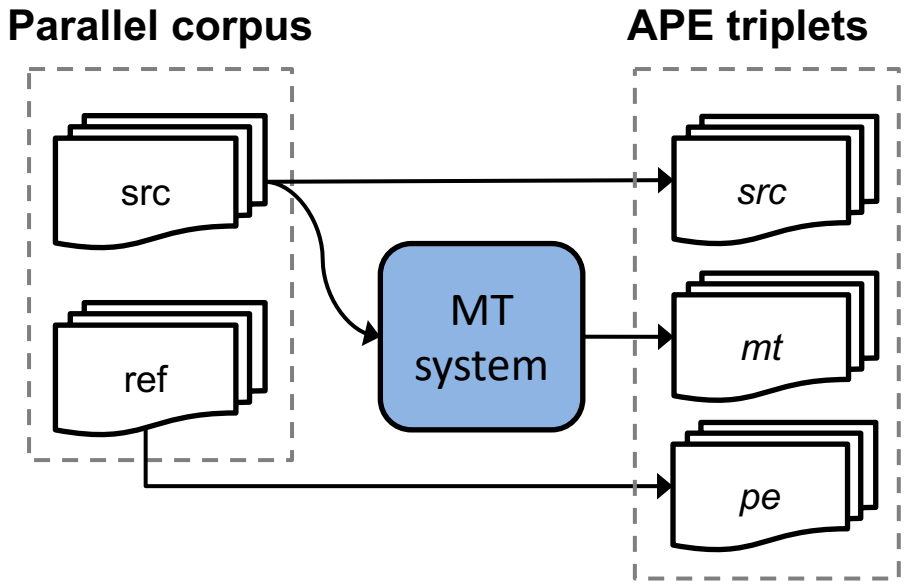
- Employing Edit-distance alignment (a.k.a Levenshtein distance)
  - Replace or add [MASK] to the pe position where the aligned words between pe and mt are NOT identical.

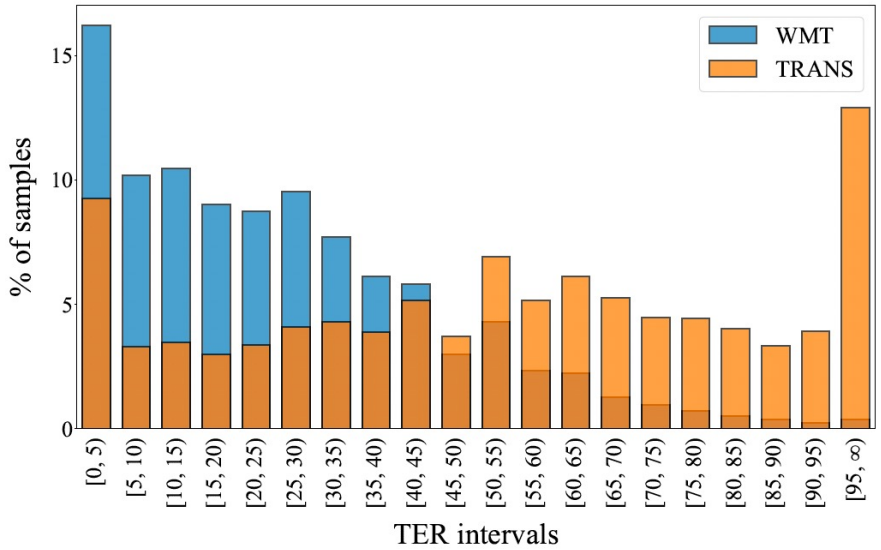
	‘’	These	is	an	examples
‘’	**0**	1	2	3	4
These	1	**0 (keep)**	1	2	3
are	2	1	**1 (sub)**	**2 (ins)**	3
examples	3	2	2	2	**2 (keep)**



Method: MLM Noising

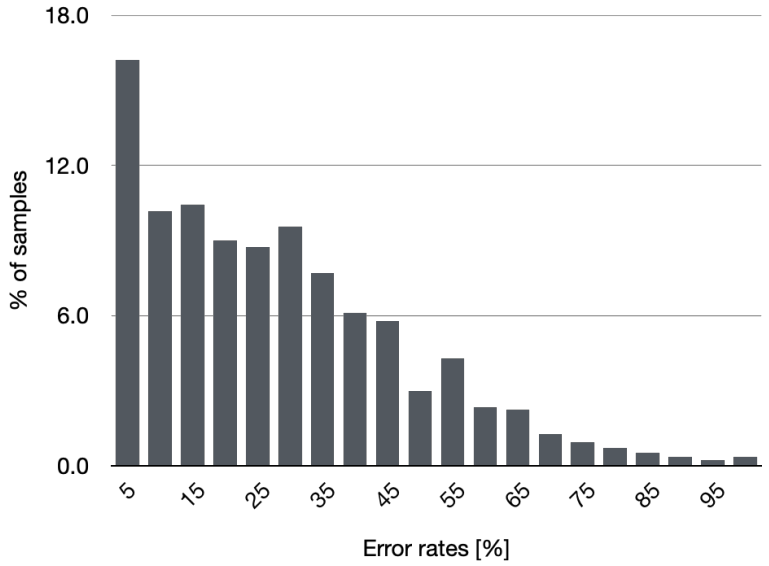
- Synthetic data made by translation (TRANS)
  - Significant discrepancy in error distribution between TRANS and gold data.
    - Edit(mt, ref) >> Edit(mt, pe)
  - Modify TRANS to resemble the error distribution of gold data, using it as additional training data for MLM noise.



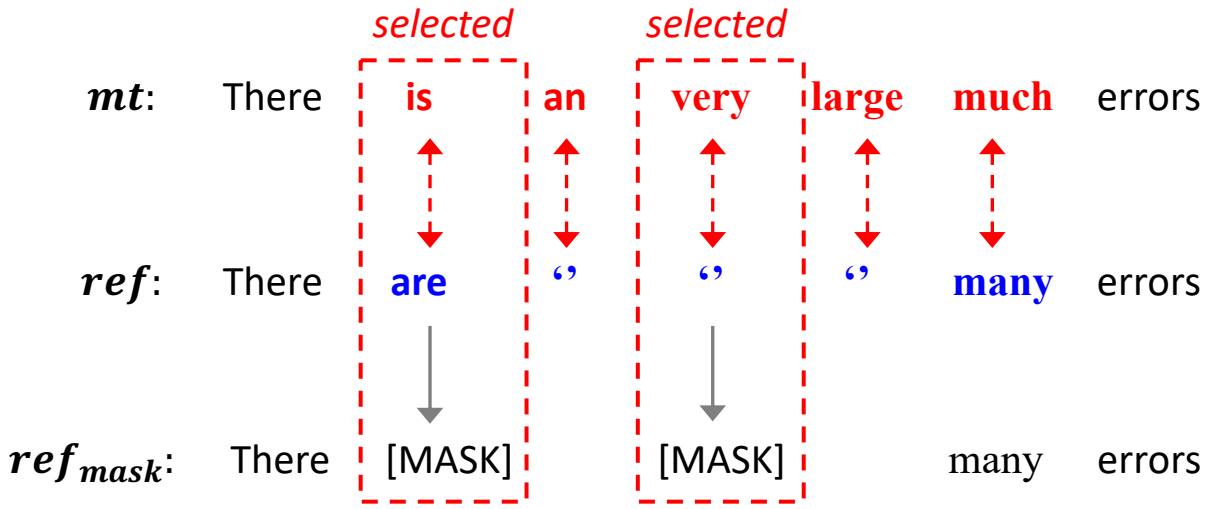


Method: MLM Noising

- Determine the error quantity according to the error-rate distribution of gold data
  - Determine **num errors** (k) by sampled error rate from gold distribution.
  - Randomly select ***k* tokens** from all n ref tokens aligned with erroneous mt tokens at every iteration (thereby covering all possible errors in expectation)

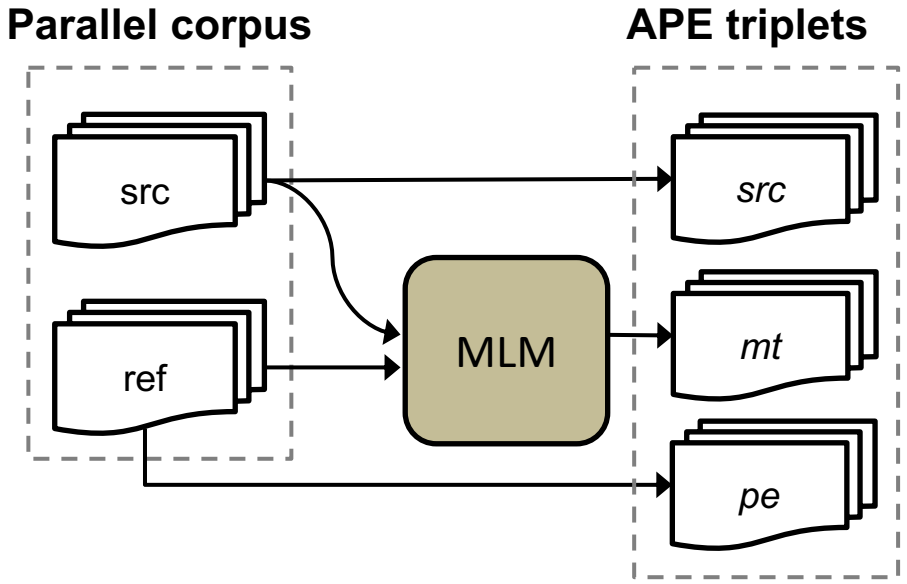


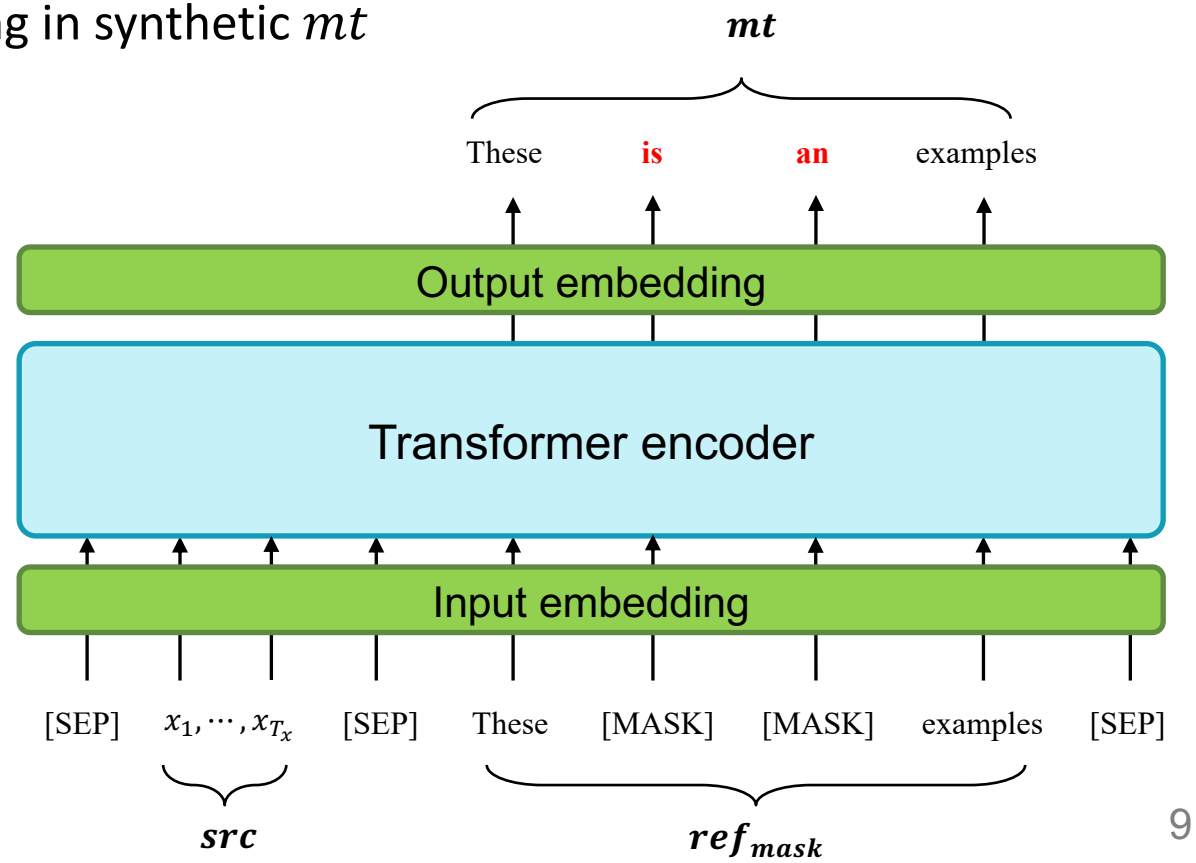
$k = 2$
$(\#\ errors)$



Method: MLM Noising

- MLM Inference
  1. Determine the error frequency n by sampling from the error distribution of the gold data.
  2. Select random n ref tokens to be masked.
  3. Replace masked token by mask-infilling, resulting in synthetic mt

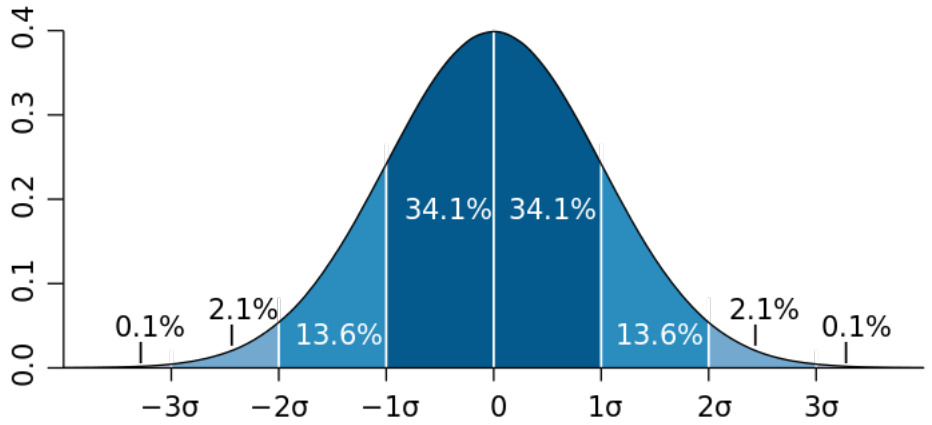




Method: Selective Corpus Interleaving

- Motivation
  - TRANS samples with similar error frequency to gold data may better resemble gold data than those generated by MLM noising.
  - Given parallel texts (src, ref), select mt from TRANS if it satisfy Edit$(mt, ref) \approx$ Edit(mt, pe); otherwise, select synthetic mt (i.e., $\widetilde{mt}$) made by MLM noising

- Method (3-sigma rule)
  - For every (src, ref), select mt by
    - $mt = \begin{cases} mt, & \text{if } |\text{Edit}(mt, ref) - \mu| \leq \lambda\sigma \\ \widetilde{mt}, & \text{otherwise} \end{cases}$
    - where μ and σ are the mean and stdv of Edit(mt, pe) from gold data; and $\lambda \in [1, 3]$ is a hyperparameter



Experiments

Approach	Test16		Test17		Test18		Test Avg.	
	TER(↓)	BLEU(↑)	TER(↓)	BLEU(↑)	TER(↓)	BLEU(↑)	TER(↓)	BLEU(↑)
TRANS	16.87	73.95	17.30	73.08	17.80	72.41	17.32	73.15
BT-FG	17.26	73.56	17.56	72.78	17.89	72.14	17.57	72.82
BT-BG	17.61	73.04	17.60	72.49	18.01	71.89	17.74	72.47
RAND	17.23	73.59	17.61	72.69	17.81	72.38	17.55	72.88
MLM Noising (w/o interleave)	16.90	74.03	17.31	72.90	17.62	72.43	17.28	73.12
MLM Noising (w/ interleave)	**16.71**	**74.58**	**16.74**	**73.79**	**17.43**	**72.88**	**16.96**	**73.75**

Table 1: Comparative result of different synthetic data used for training

	Test Avg.		Sample Ratio	
	TER(↓)	BLEU(↑)	MLM	TRANS
$\lambda = 0$ (MLM Noising)	17.32	73.15	100.0%	0.0%
$\lambda = 1$	17.25	73.26	71.5%	28.5%
$\lambda = 2$	**16.96****	**73.75****	41.5%	58.5%
$\lambda = 3$	17.03**	73.48**	20.8%	79.2%
$\lambda = \infty$ (TRANS)	17.28	73.12	0.0%	100.0%

Table 2: Results of varying hyperparameters for corpus interleaving

- Comparative analysis: our synthetic data vs. existing datasets
  - The model trained on synthetic data generated through our method exhibited the most superior performance compared to the existing datasets.

- Variations of hyperparameters for selective corpus interleaving
  - The model performed optimally when combining approximately half of TRANS and MLM noising datasets, indicating that both approaches their own strengths.

Summary & Conclusion

- The scarcity of training data underscores the importance of semi-supervised learning, which utilizes synthetic data to augment available training resources.

- We introduce **MLM noising** method, creating synthetic data that accurately mimics the essential characteristics of gold data.

- We suggest **selective corpus interleaving** method to combines two synthetic datasets, TRANS and our MLM, effectively leveraging their strengths and addressing their limitations.

End of the presentation