Mixed Cross Entropy Loss for Neural Machine Translation

Haoran Li, Wei Lu

StatNLP Research

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

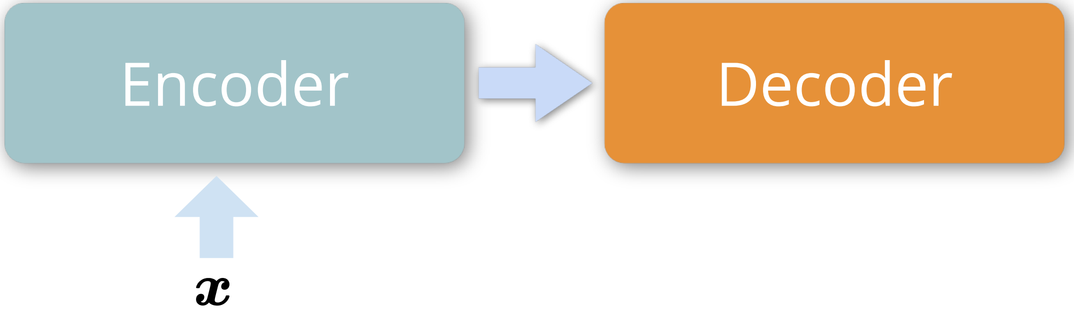
target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

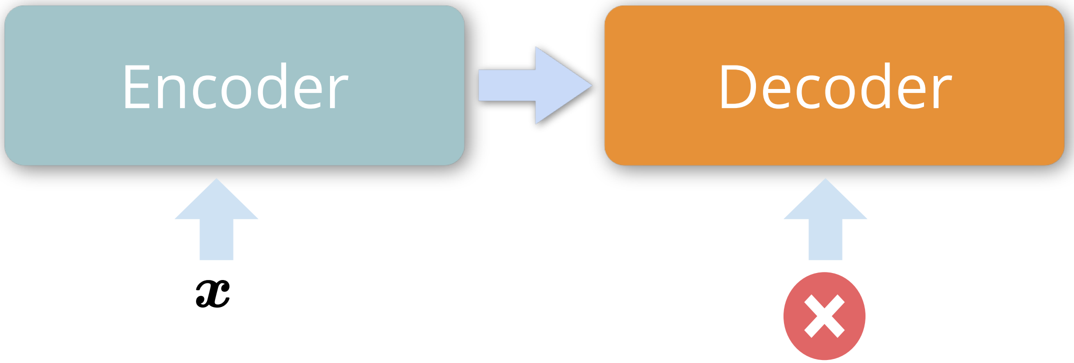


1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$



1. Background of NMT

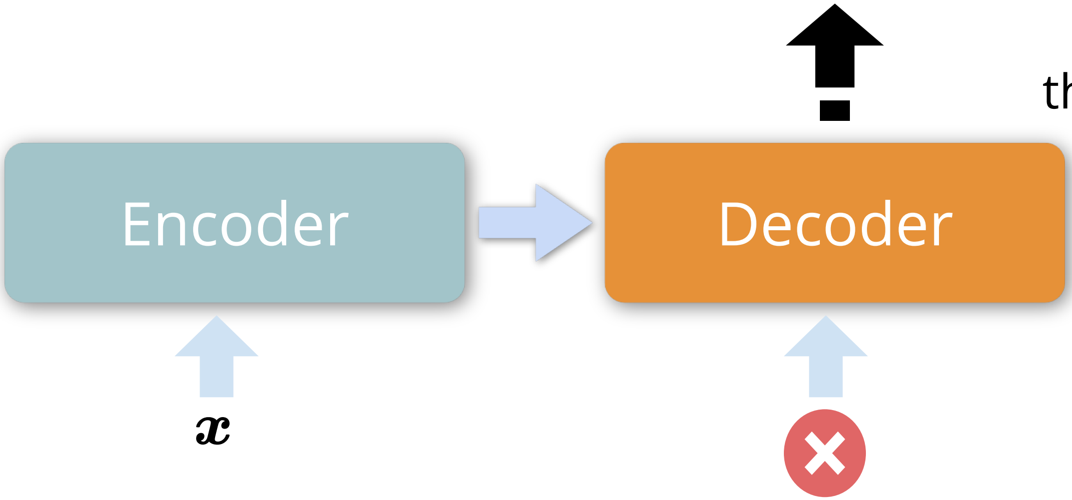
$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

Cross Entropy Loss: $-\sum_{t=1}^{n} \log p_\theta(y_t | \boldsymbol{x}, \hat{\boldsymbol{y}}_{<t})$

the model's own predictions



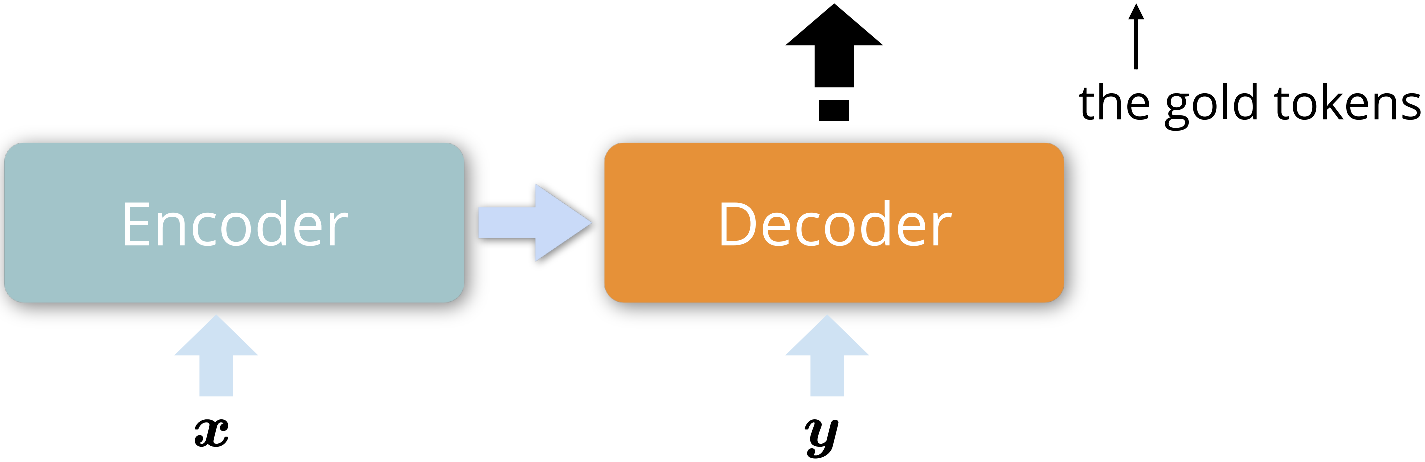
1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

Cross Entropy Loss: $-\sum_{t=1}^{n} \log p_\theta(y_t | \boldsymbol{x}, \boldsymbol{y}_{<t})$



Teacher Forcing (Williams & Zipser 1989)

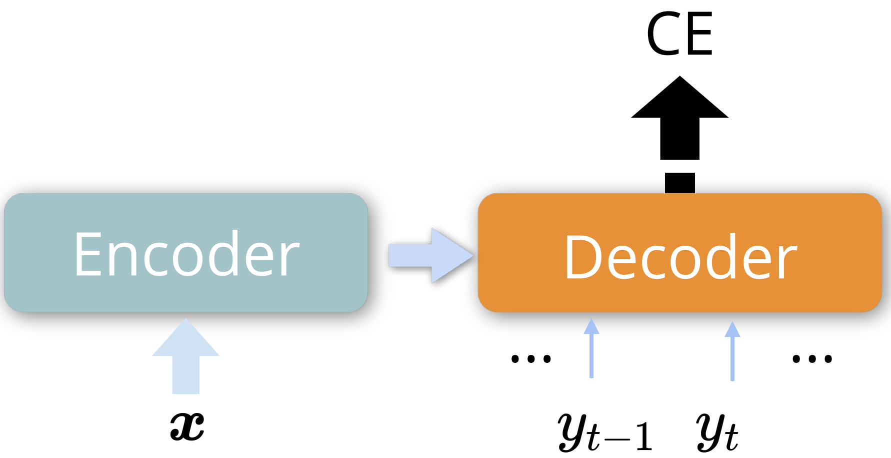
1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

Exposure Bias (M. Ranzato et al. 2016)



Training

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$

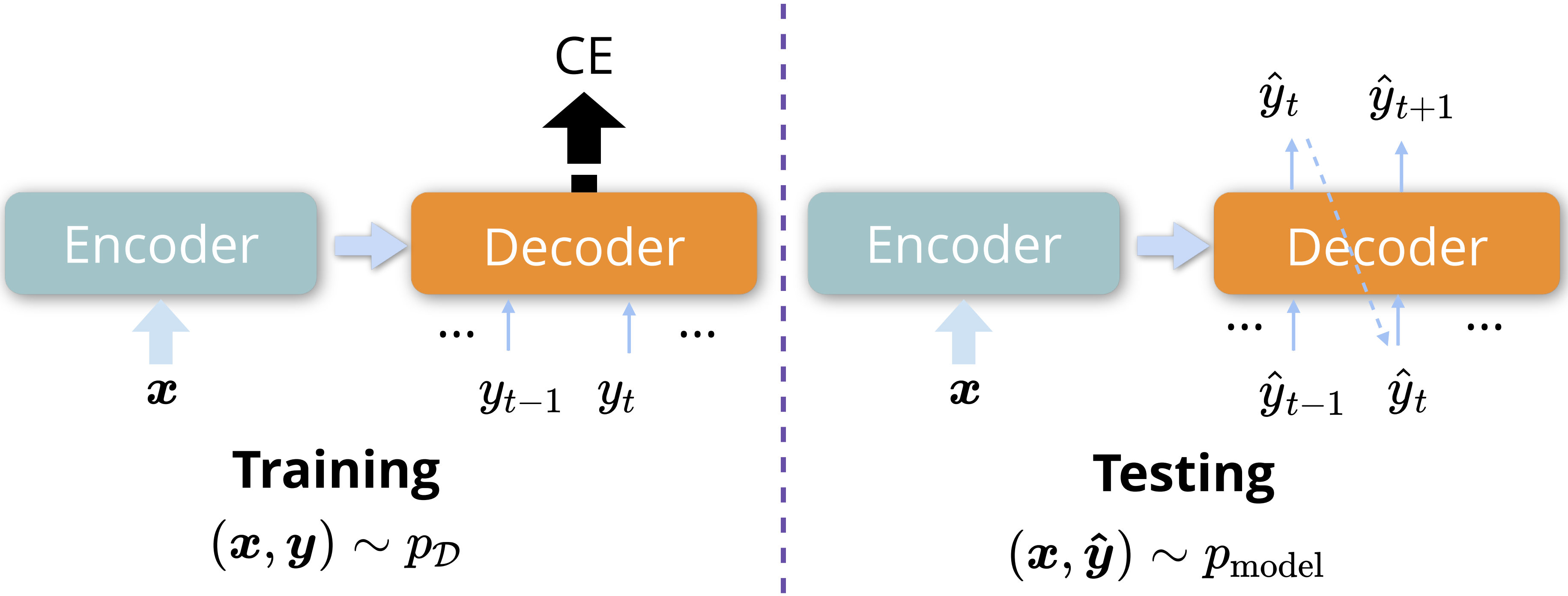
1. Background of NMT

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, \ldots, y_n)$

Exposure Bias (M. Ranzato et al. 2016)



1. Background of NMT

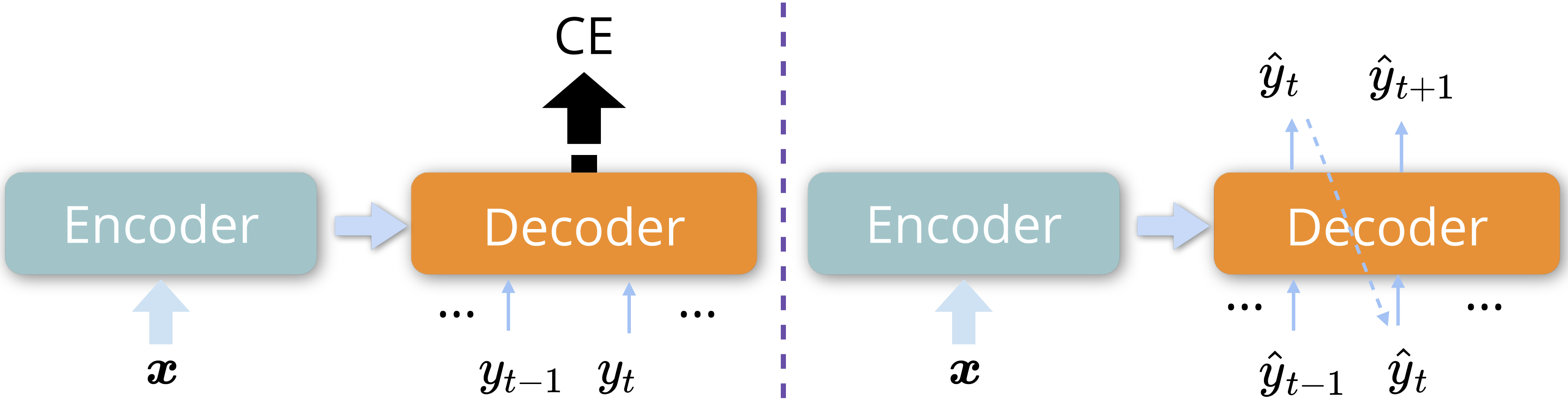
$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, empirical data distribution

source sentence: $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

target sentence: $\boldsymbol{y} = (y_0, y_1, ..., y_n)$

Exposure Bias (M. Ranzato et al. 2016)

$p_{\mathcal{D}} \neq p_{\text{model}}$



Training

$(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$

Testing

$(\boldsymbol{x}, \hat{\boldsymbol{y}}) \sim p_{\text{model}}$

1. Background of NMT

How to mitigate exposure bias?

1. Background of NMT

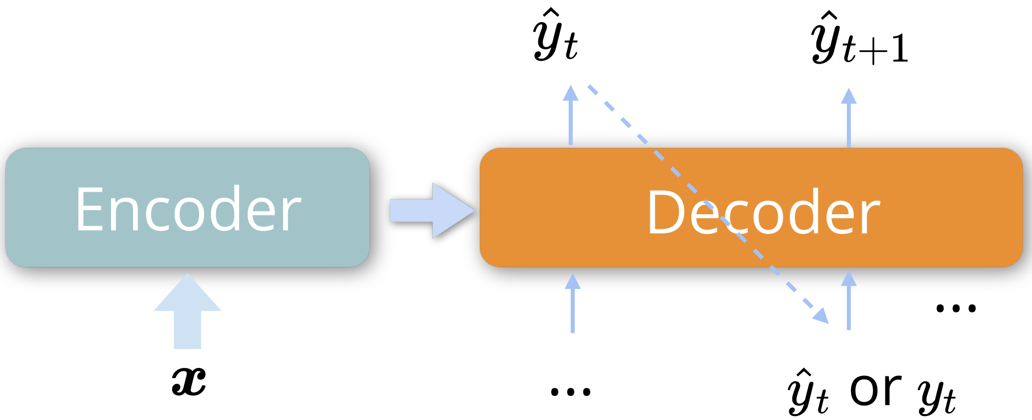
How to mitigate exposure bias?

- We expose the model to its own predictions during training.

1. Background of NMT

How to mitigate exposure bias?

- We expose the model to its own predictions during training.
- Scheduled Sampling for RNNs (S. Bengio et al. 2015)

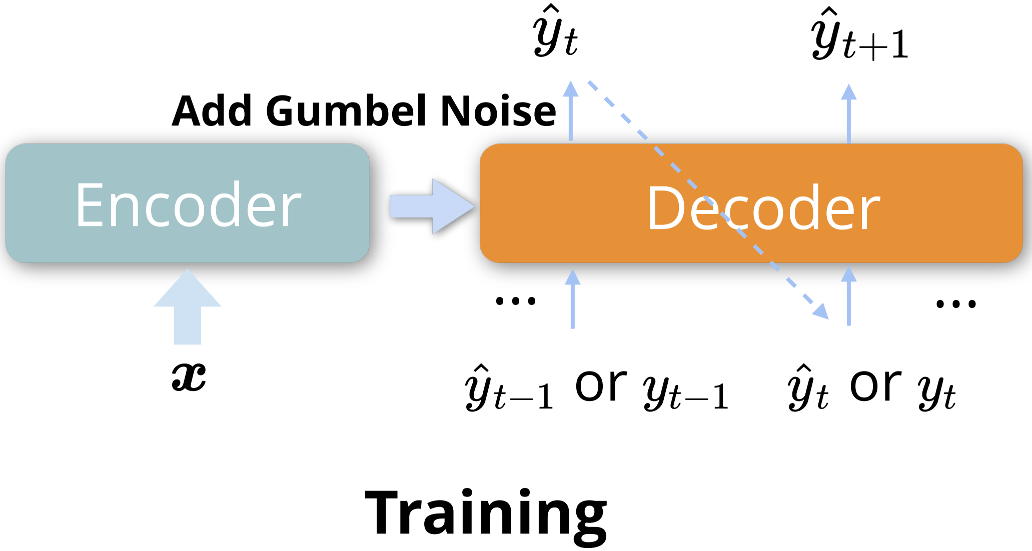


Training

1. Background of NMT

How to mitigate exposure bias?

- We expose the model to its own predictions during training.
- Scheduled Sampling for RNNs (S. Bengio et al. 2015)
- Word Oracle: add Gumbel Noise (K. Goyal et al. 2017, W. Zhang et al. 2019)



1. Background of NMT

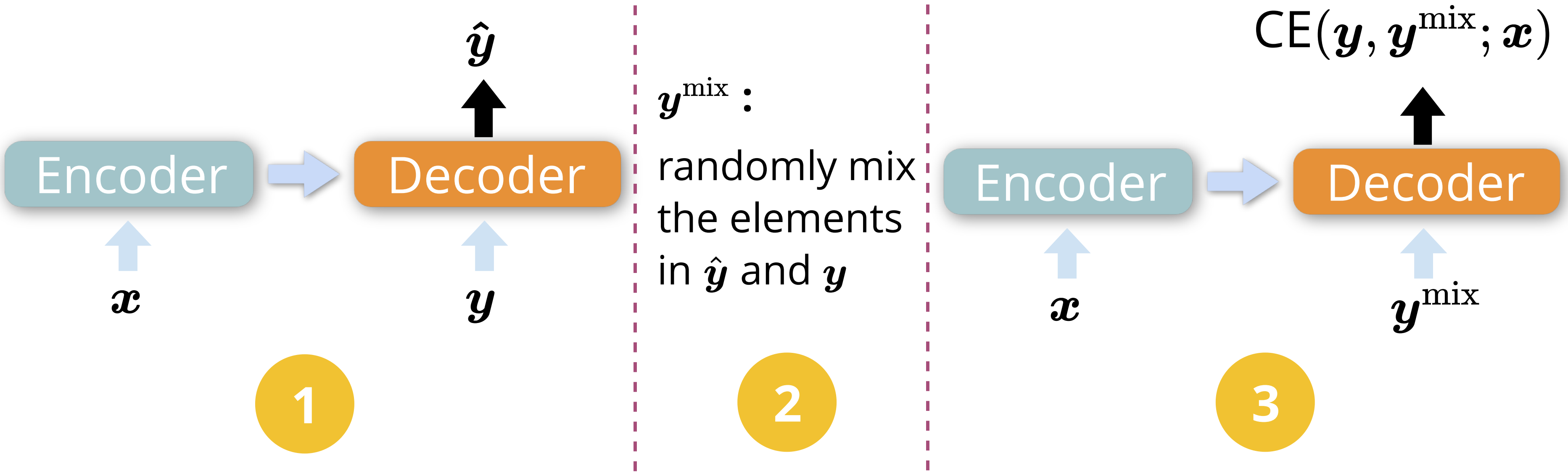
How to mitigate exposure bias?

- We expose the model to its own predictions during training.
- Scheduled Sampling for RNNs (S. Bengio et al. 2015)
- Word Oracle: add Gumbel Noise (K. Goyal et al. 2017, W. Zhang et al. 2019)
- Scheduled Sampling for Transformers (T. Mihaylova et al. 2019, D. Duckworth et al. 2019, Wen Zhang et al. 2019)

1. Background of NMT

How to mitigate exposure bias?

- We expose the model to its own predictions during training.
- Scheduled Sampling for RNNs (S. Bengio et al. 2015)
- Word Oracle: add Gumbel Noise (K. Goyal et al. 2017, W. Zhang et al. 2019)
- Scheduled Sampling for Transformers (T. Mihaylova et al. 2019, D. Duckworth et al. 2019, Wen Zhang et al. 2019)



2. Motivation

- Teacher Forcing
- Scheduled Sampling

2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$ **one-to-one** mapping

2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$ **one-to-one** mapping

Machine Translation is inherently a **one-to-many** mapping problem

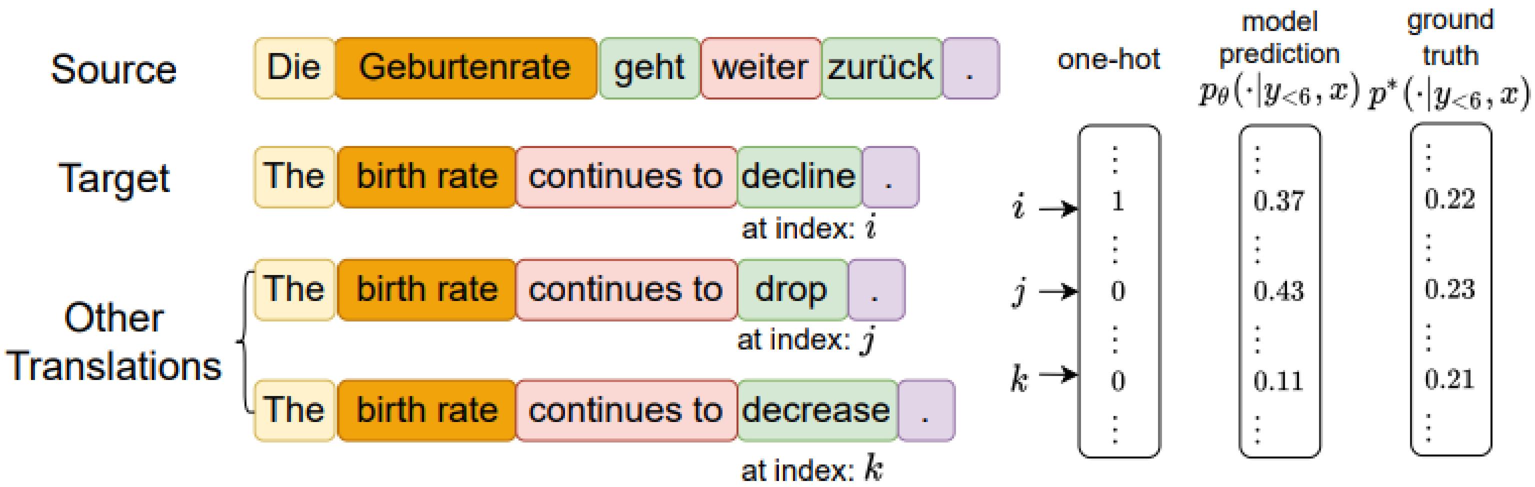
2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$$ **one-to-one** mapping

Machine Translation is inherently a **one-to-many** mapping problem



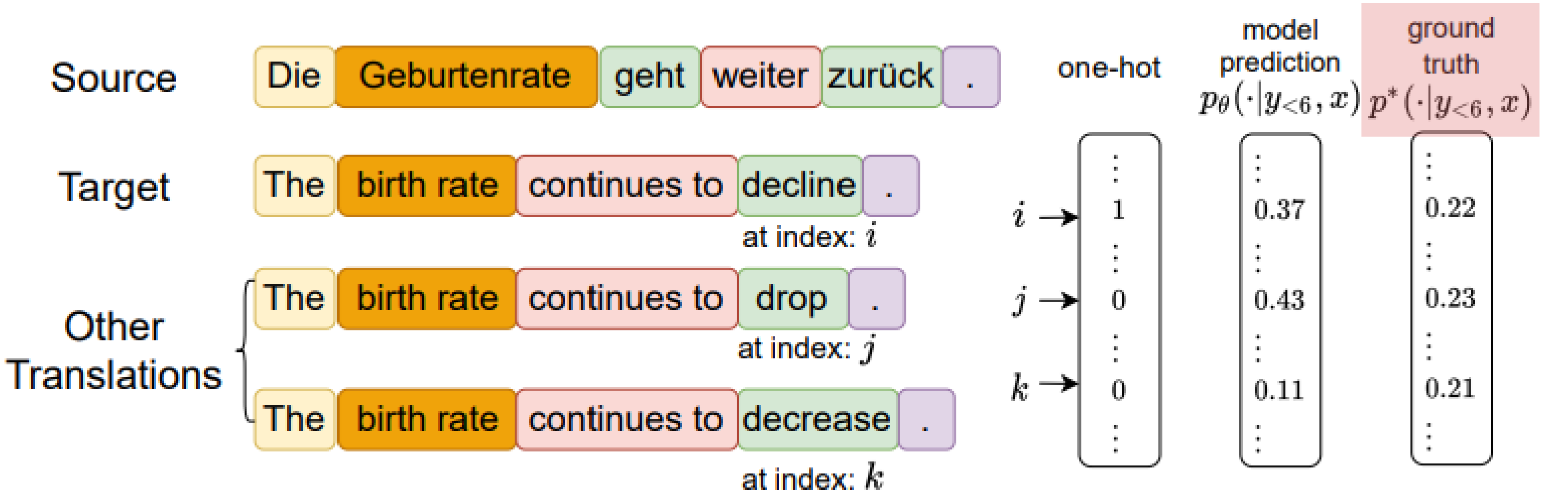
2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$ **one-to-one** mapping

Machine Translation is inherently a **one-to-many** mapping problem



Ideally, the target should be $p^*(\cdot|y_{<t}, x)$, instead of the one-hot encoding.

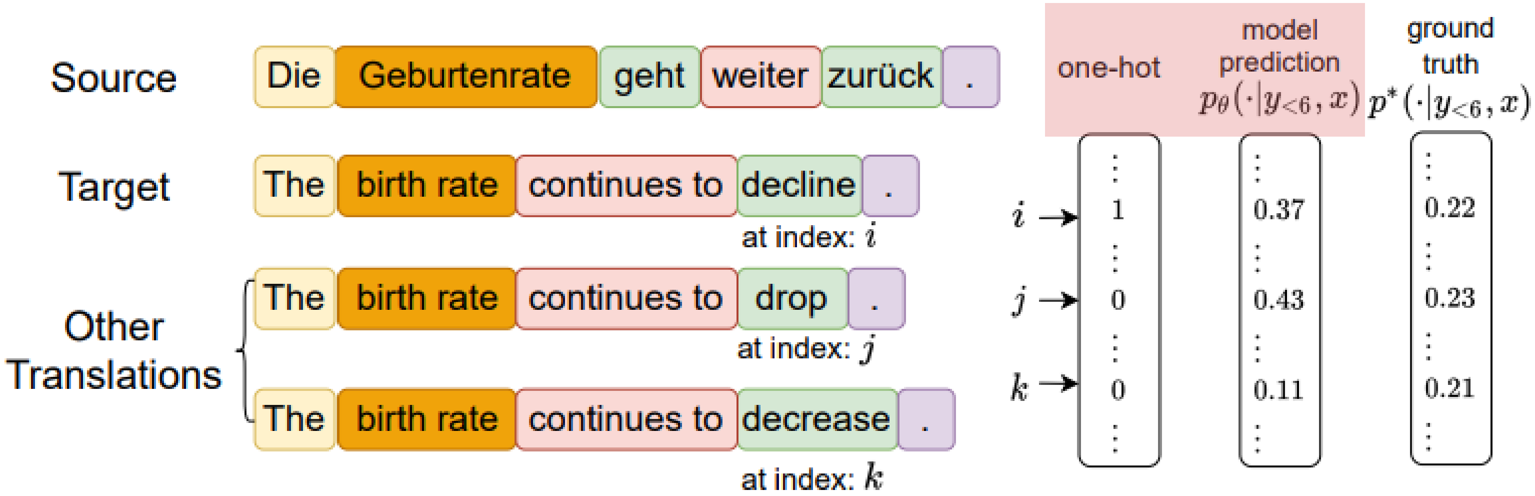
2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$ **one-to-one** mapping

Machine Translation is inherently a **one-to-many** mapping problem



Ideally, the target should be $p^*(\cdot|y_{<t}, x)$, instead of the one-hot encoding.

But only one-hot encoding and model predictions are available.

2. Motivation

- Teacher Forcing
- Scheduled Sampling

Teacher Forcing: $\boldsymbol{x}, \boldsymbol{y}_{<t} \Rightarrow y_t$ one-hot encoding

$p_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t}) \Rightarrow [0, ..., 1, ...0]$ **one-to-one** mapping

Machine Translation is inherently a **one-to-many** mapping problem

- How to exploit the ground truth information $p^*(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<t})$?

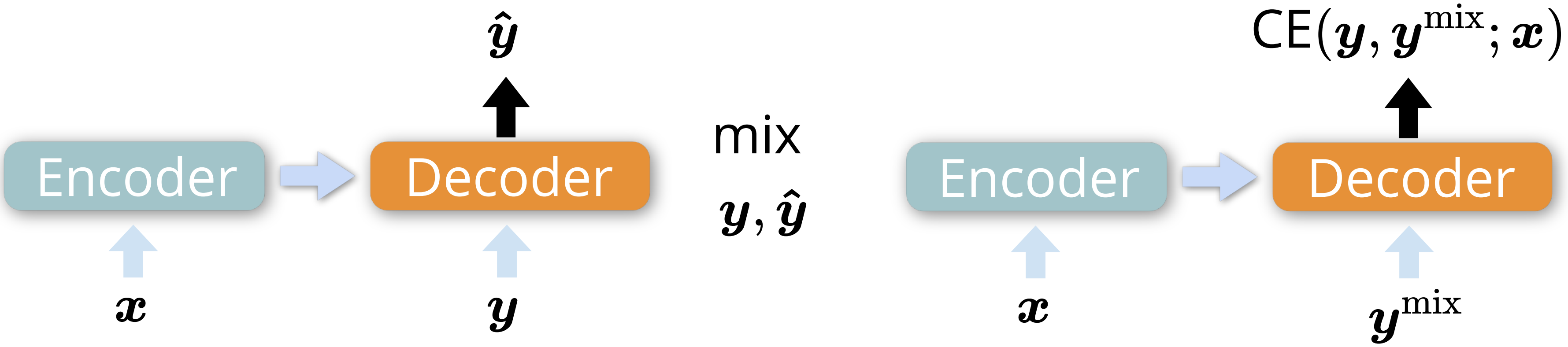
  - *Assumption:*
    Givened a well-trained model with parameters θ, if $\hat{y}_t = \arg\max p_\theta(\cdot|x, y_{<t}) \neq y_t$, then $\hat{y}_t$ is very likely to be a synonym or part of a synonym of the gold token y_t.

  - Use y_t and $\hat{y}_t$ in **mixed CE**.

2. Motivation

- Teacher Forcing
- Scheduled Sampling

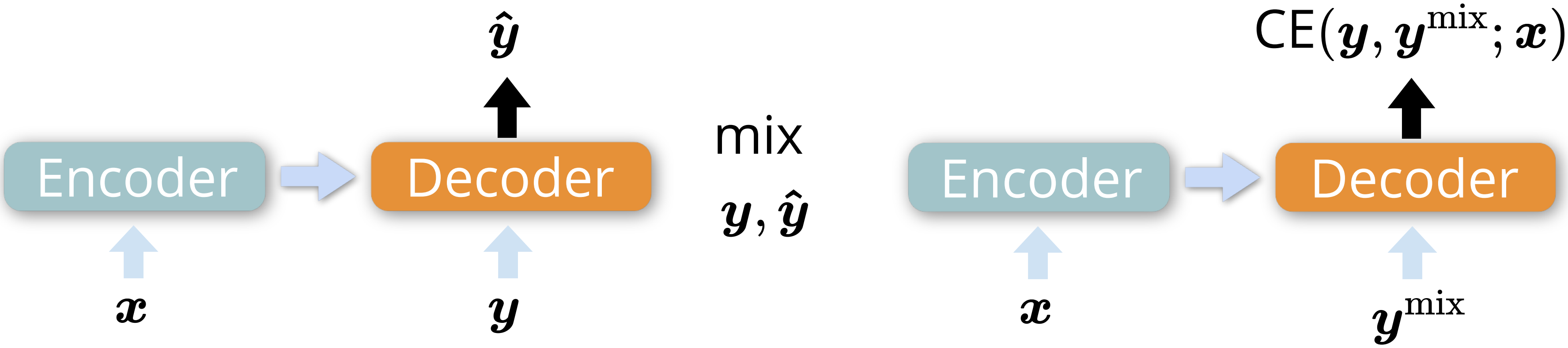
Scheduled Sampling:



2. Motivation

- Teacher Forcing
- Scheduled Sampling

Scheduled Sampling:

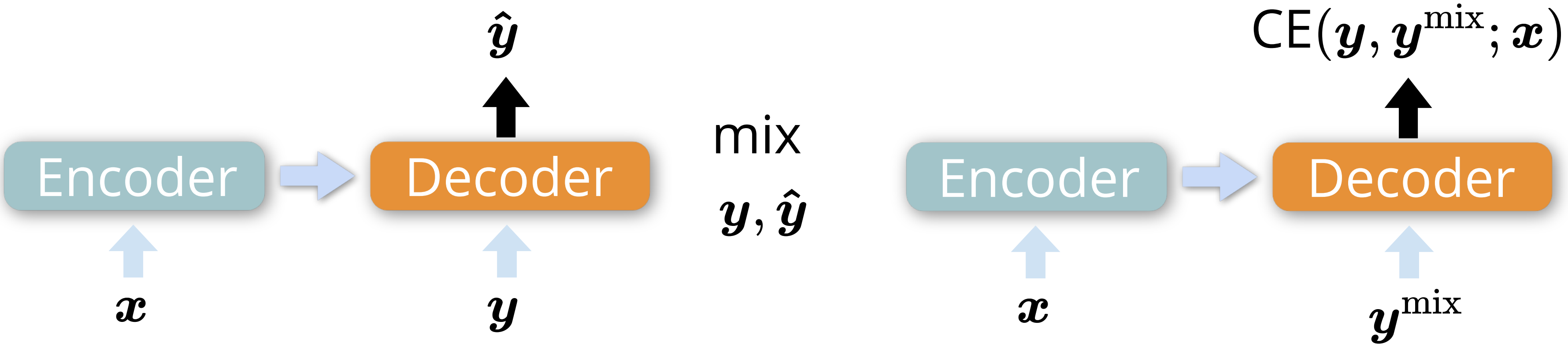


- We force input distribution $p_{\mathcal{D}}$ to approximate p_{model} by $(\boldsymbol{x}, \boldsymbol{y}) \rightarrow (\boldsymbol{x}, \boldsymbol{y}^{\mathrm{mix}})$

2. Motivation

- Teacher Forcing
- Scheduled Sampling

Scheduled Sampling:



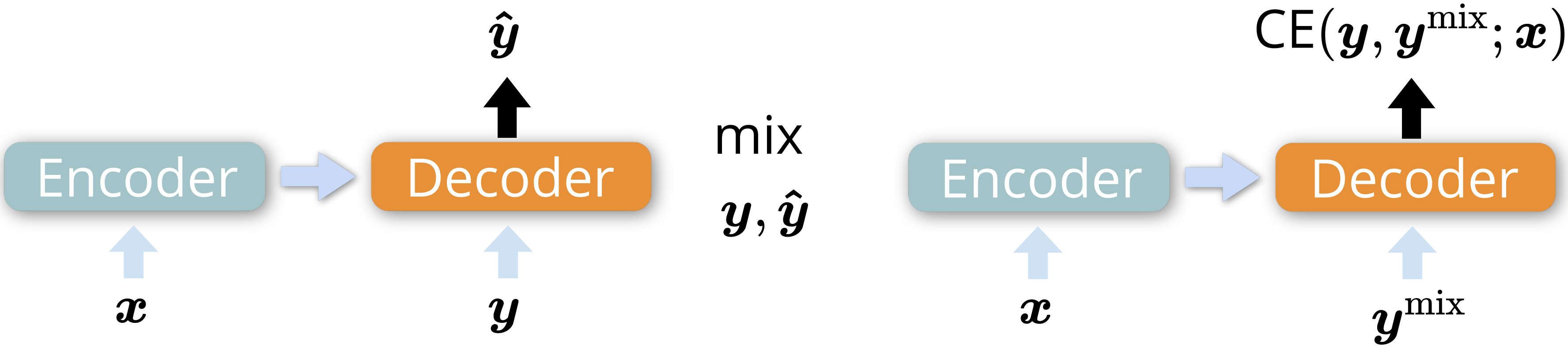
- We force input distribution $p_{\mathcal{D}}$ to approximate p_{model} by $(\boldsymbol{x}, \boldsymbol{y}) \rightarrow (\boldsymbol{x}, \boldsymbol{y}^{\text{mix}})$

- But can we make the model insensitive to the inputs from different distributions?

2. Motivation

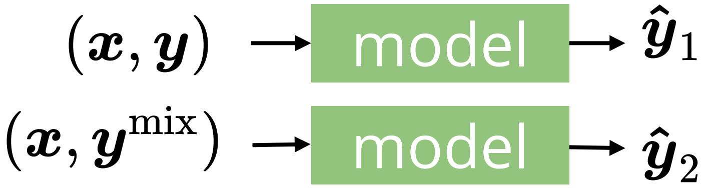
- Teacher Forcing
- Scheduled Sampling

Scheduled Sampling:



- We force input distribution $p_{\mathcal{D}}$ to approximate p_{model} by $(\boldsymbol{x}, \boldsymbol{y}) \rightarrow (\boldsymbol{x}, \boldsymbol{y}^{\mathrm{mix}})$

- But can we make the model insensitive to the inputs from different distributions?

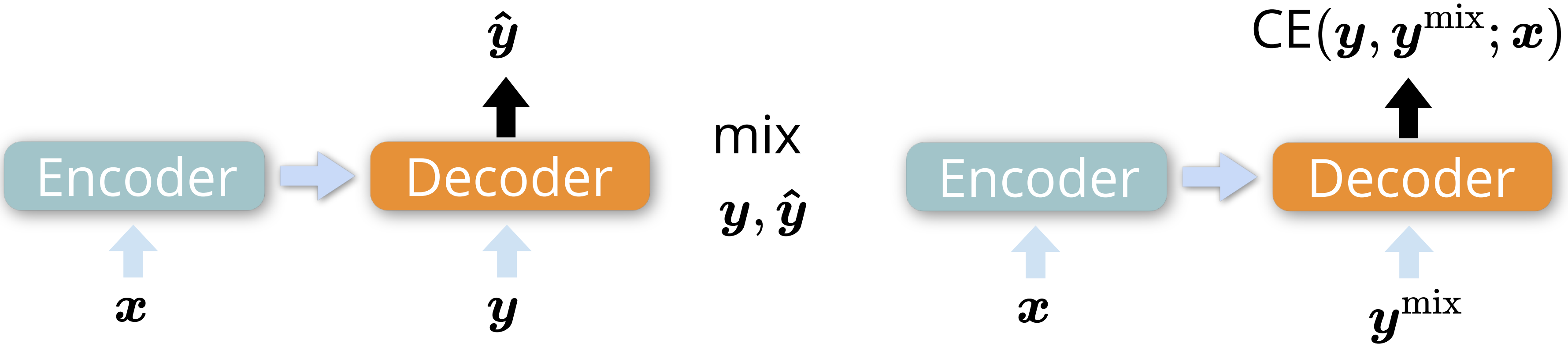


$$\hat{\boldsymbol{y}}_1 \approx \hat{\boldsymbol{y}}_2 \approx \boldsymbol{y}$$

2. Motivation

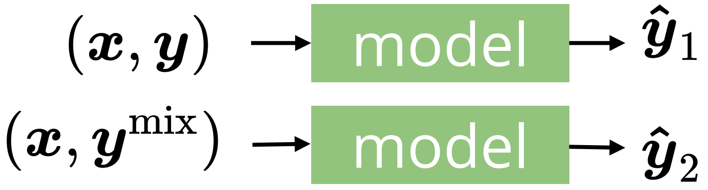
- Teacher Forcing
- Scheduled Sampling

Scheduled Sampling:



- We force input distribution $p_{\mathcal{D}}$ to approximate p_{model} by $(\boldsymbol{x}, \boldsymbol{y}) \rightarrow (\boldsymbol{x}, \boldsymbol{y}^{\text{mix}})$

- But can we make the model insensitive to the inputs from different distributions?



Ignore the discrepancy in the decoder inputs of which the source inputs are the same.

3. Approach

- Mixed CE in teacher forcing:

$$\mathcal{L}_{mix} = -\left[(1-\alpha_i) \cdot \underline{\sum_{t=1}^{n} \log p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x})} + \alpha_i \cdot \sum_{t=1}^{n} \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}, \boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1 \leq k \leq |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t}, \boldsymbol{x})$$

$$\alpha_i = m \cdot \frac{i}{\text{total\_iter}}, \quad m = 0.5$$

3. Approach

- Mixed CE in teacher forcing:

$$\mathcal{L}_{mix} = -\left[(1-\alpha_i)\cdot \underline{\sum_{t=1}^{n} \log p_\theta(y_t|\boldsymbol{y}_{<t},\boldsymbol{x})} + \alpha_i \cdot \sum_{t=1}^{n} \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t},\boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1\le k\le |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t},\boldsymbol{x})$$

$$\alpha_i = m\cdot \frac{i}{\text{total\_iter}}, \quad m = 0.5$$

- Mixed CE in scheduled sampling:

$$\mathcal{L}_{mix} = -\left[(1-\alpha_i)\cdot \underline{\sum_{t=1}^{n} \log p_\theta(y_t|\boldsymbol{y}^{\text{mix}}_{<t},\boldsymbol{x})} + \alpha_i \cdot \sum_{t=1}^{n} \log p_\theta(\hat{y}_t|\boldsymbol{y}^{\text{mix}}_{<t},\boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1\le k\le |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t},\boldsymbol{x})$$

$$\alpha_i = m\cdot \frac{i}{\text{total\_iter}}, \quad m = 0.5$$

3. Approach

- Mixed CE in teacher forcing:

$$\mathcal{L}_{mix} = -\left[(1-\alpha_i) \cdot \underline{\sum_{t=1}^{n} \log p_\theta(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x})} + \alpha_i \cdot \sum_{t=1}^{n} \log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}, \boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

$$\alpha_i = m \cdot \frac{i}{\text{total\_iter}}, \quad m = 0.5$$

- How to understand mixed CE ?
  - **When $y_t = \hat{y}_t$, it degenerates to standard CE.**
  - **When $y_t \neq \hat{y}_t$, $\hat{y}_t$ is very likely to be a synonym of y_t**

3. Approach

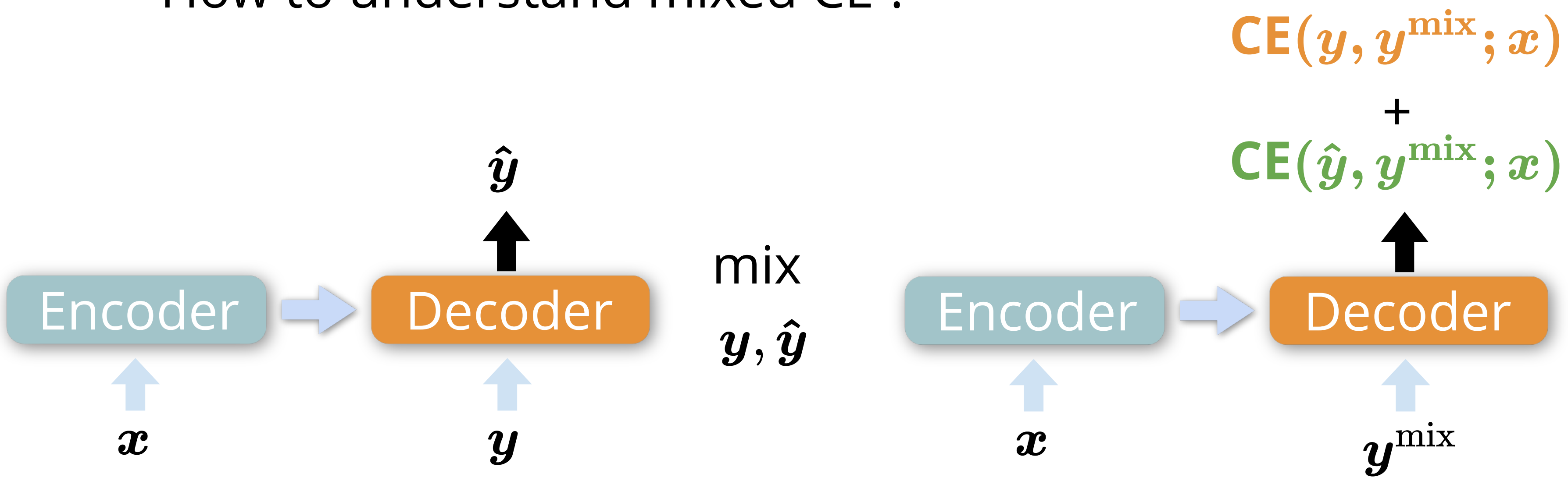
- Mixed CE in scheduled sampling:

$$\mathcal{L}_{mix} = -\Big[(1-\alpha_i)\cdot \underline{\textstyle\sum_{t=1}^{n} \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})} + \alpha_i \cdot \underline{\textstyle\sum_{t=1}^{n} \log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})}\Big]$$

$$\hat{y}_t = \arg\max_{1\le k\le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

$$\alpha_i = m \cdot \frac{i}{\text{total\_iter}}, \quad m = 0.5$$

- How to understand mixed CE ?



4. Experiments

- Teacher Forcing: single reference test set

Table 1. BLEU scores on test sets of Transformers trained with CE and mixed CE. The results of beam search decoding with beam size 1/5 are presented. All results are averaged over 3 runs.

DATA SET	LOSS	SINGLE	AVERAGE
RO-EN	CE	30.63/31.42	32.07/32.59
	DSD	**31.17**/31.80	32.03/32.74
	SELF-DIST	28.65/31.45	31.66/32.61
	MIXED CE	**31.17/32.02**	**32.63/33.25**
RU-EN	CE	28.87/30.24	29.48/30.79
	DSD	28.89/30.30	29.69/30.90
	SELF-DIST	28.76/30.34	29.32/30.63
	MIXED CE	**29.59/30.74**	**30.14/31.05**
EN-DE	CE	26.23/26.91	26.67/27.41
	DSD	26.10/26.84	26.66/27.30
	SELF-DIST	24.15/25.98	24.23/25.91
	MIXED CE	**26.32/27.28**	**26.72/27.61**

4. Experiments

- Teacher Forcing: multi-reference set (M. Ott et al. 2018)
  - 10 references for each of the 500 test sentences taken from the original test set
  - We generate 10 hypotheses for each source sentence using beam search

Table 2. BLEU improvement of mixed CE over CE on 10 additional references of WMT'14 En-De test set. All results are averaged over 3 runs.

REF	AVG		TOP	
	CE	MIXED CE	CE	MIXED CE
REF 1	36.73	**37.32 (+0.59)**	38.61	**39.13 (+0.52)**
REF 2	47.48	**48.50 (+1.02)**	50.08	**51.36 (+1.28)**
REF 3	42.59	**43.25 (+0.66)**	44.89	**45.89 (+1.00)**
REF 4	28.93	**29.78 (+0.85)**	30.29	**30.98 (+0.69)**
REF 5	31.75	**32.53 (+0.78)**	33.48	**34.18 (+0.70)**
REF 6	26.41	**26.83 (+0.42)**	27.60	**27.96 (+0.36)**
REF 7	42.18	**42.89 (+0.71)**	44.37	**44.90 (+0.53)**
REF 8	32.36	**33.05 (+0.69)**	33.77	**34.55 (+0.78)**
REF 9	28.51	**29.03 (+0.52)**	29.65	**30.27 (+0.62)**
REF 10	33.75	**33.94 (+0.19)**	35.23	**35.68 (+0.45)**
MEAN	35.07	**35.71 (+0.64)**	36.80	**37.49 (+0.69)**

4. Experiments

- Teacher Forcing: WMT'19 En-De paraphrased reference set (M. Freitag et al. 2020)
    - Each reference is paraphrased from the original reference by human experts and differs significantly from the original one in word choices and sentence structures

Table 3. BLEU scores of beam search/sampling results on WMT'19 En-De paraphrased test set. As a reference, Freitag et al. (2020) reported that the BLEU score improvement of the machine translation system augmented with Automatic-Post-Editing/Back-Translation (Freitag et al., 2019; Sennrich et al., 2016a) on this paraphrased set was 0.2/0.4 BLEU.

LOSS	BEAM 1	BEAM 10	SAMPLING
CE	11.26	11.67	8.89
MIXED CE	**11.60**	**11.94**	**9.90**

4. Experiments

- Teacher Forcing: comparison with Label Smoothing (LS) (Szegedy et al. 2016)
  - Pairwise BLEU (PB): measuring the diversity of the hypothesis translations (Shen et al. 2019)
  - High PB, more similar; Low PB, less similar.

Table 4. PB, BLEU on WMT'14 En-De validation set. Pairwise-BLEU is obtained using sampling decoding while the BLEU score is obtained using beam search. LS is short for label smoothing.

Loss	PB (↓)	BLEU (↑)
No LS, No Mixed CE	17.52	25.81
+ LS	5.22	26.48
+ Mixed CE	25.99	26.26
+ LS, mixed CE	7.79	**26.75**

4. Experiments

- Teacher Forcing: comparison with Label Smoothing (LS) (Szegedy et al. 2016)

  - Cumulative Sequence Probability: cumulative probability of the hypotheses generated using beam search (M. Ott et al. 2018)

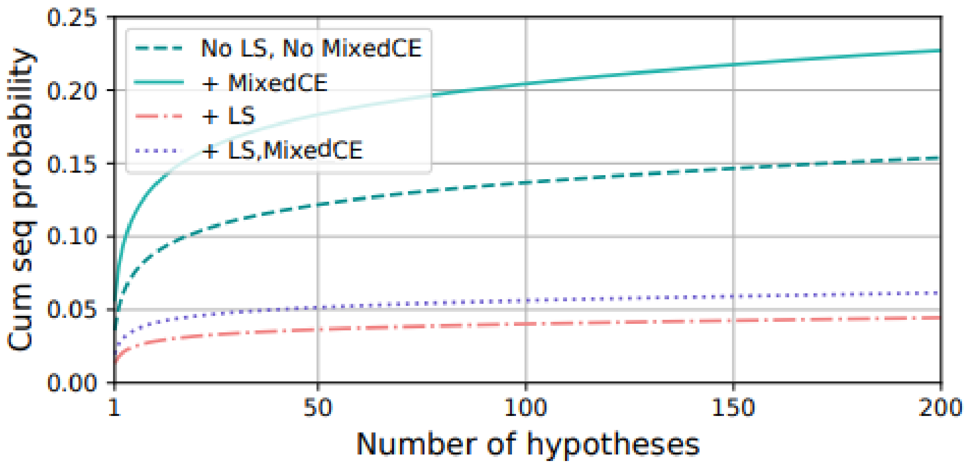


Figure 4. Cumulative sequence probability of generated hypotheses using beam search with beam size 200 on WMT'14 En-De validation set.

4. Experiments

- Scheduled Sampling

Table 5. BLEU scores on test sets of Transformers trained with CE and mixed CE. The results of beam search decoding with beam size 1/5 are presented. All results are averaged over 3 runs.

Data set	Loss	Scheduield Sampling	
		Single	Average
Ro-En	CE	30.71/31.72	32.29/33.05
	Mixed CE	**31.71/32.53**	**32.88/33.45**
Ru-En	CE	29.28/30.63	29.62/30.83
	Mixed CE	**30.19/31.23**	**30.47/31.39**
En-De	CE	26.36/27.29	26.84/27.56
	Mixed CE	**26.75/27.57**	**26.99/27.71**

Data set	Loss	Word Oracle	
		Single	Average
Ro-En	CE	31.71/32.37	33.05/33.76
	Mixed CE	**32.43/33.06**	**33.66/34.14**
Ru-En	CE	29.40/30.61	29.87/31.00
	Mixed CE	**30.24/31.09**	**30.72/31.50**
En-De	CE	26.66/27.45	**26.94**/27.71
	Mixed CE	**26.81/27.80**	**26.94/27.88**

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i) \cdot \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \underline{\log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})} \right]$$

$$\hat{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Is it really important?

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i) \cdot \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \underline{\log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})}\right]$$

$$\hat{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Is it really important?

Top-2 Mixed CE: replace the above $\hat{y}_t$ with

$$\hat{y}_t = \text{Rand}\Big(\text{Top-2}_{1 \le k \le |V|}\big(\log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})\big)\Big)$$

Table 6. BLEU scores of Transformers trained with different loss functions on the WMT'16 Ro-En validations sets.

LOSS	SS	WORD ORACLE
CE	32.66	33.82
MIXED CE	**33.64**	**34.51**
TOP-2 MIXED CE	32.17	32.76
RANDOM MIXED CE	33.26	34.18
SOFT MIXED CE	32.03	33.08

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i) \cdot \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \underline{\log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})} \right]$$

$$\hat{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Is it really important?

Random Mixed CE: replace the above $\hat{y}_t$ with

$$\hat{y}_t = \begin{cases} y_t, \text{if } y_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x}) \\ \text{Rand}(V), \text{otherwise} \end{cases}$$

Table 6. BLEU scores of Transformers trained with different loss functions on the WMT'16 Ro-En validations sets.

LOSS	SS	WORD ORACLE
CE	32.66	33.82
MIXED CE	**33.64**	**34.51**
TOP-2 MIXED CE	32.17	32.76
RANDOM MIXED CE	33.26	34.18
SOFT MIXED CE	32.03	33.08

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n}\left[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1\leq k\leq |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t}, \boldsymbol{x})$$

Is it really important?

Soft Mixed CE: replace the above $\hat{y}_t$ with

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n}\Bigg[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \sum_{k=1}^{|V|} q_\theta(w_k|\boldsymbol{y}_{<t}, \boldsymbol{x}) \cdot \log p_\theta(w_k|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\Bigg]$$

Table 6. BLEU scores of Transformers trained with different loss functions on the WMT'16 Ro-En validations sets.

Loss	SS	Word Oracle
CE	32.66	33.82
Mixed CE	**33.64**	**34.51**
Top-2 Mixed CE	32.17	32.76
Random Mixed CE	33.26	34.18
Soft Mixed CE	32.03	33.08

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\right]$$

$$\hat{y}_t = \arg\max_{1\leq k\leq |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t}, \boldsymbol{x})$$

Double Mixed CE: we also apply mixed CE to output in 2nd pass in schduled sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \Big[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \frac{\alpha_i}{2} \cdot \left(\log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \log p_\theta(\tilde{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\right)\Big]$$

$$\tilde{y}_t = \arg\max_{1\leq k\leq |V|} \log p_\theta(w_k|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})$$

Table 7. BLEU scores of Transformers trained with *double mixed CE* and *mixed CE-2nd pass* on validations sets.

LOSS	RO-EN	RU-EN	EN-DE
CE	33.82	29.83	26.51
MIXED CE	**34.51**	**30.46**	26.88
DOUBLE MIXED CE	34.23	**30.46**	**27.06**
MIXED CE-2ND PASS	33.84	30.16	26.83

4. Experiments

- Scheduled Sampling

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i) \cdot \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\hat{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) \right]$$

$$\hat{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Mixed CE 2nd pass:

$$\mathcal{L}_{mix} = -\sum_{t=1}^{n} \left[(1-\alpha_i) \cdot \log p_\theta(y_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\tilde{y}_t | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) \right]$$

$$\tilde{y}_t = \arg\max_{1 \le k \le |V|} \log p_\theta(w_k | \boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})$$

Table 7. BLEU scores of Transformers trained with *double mixed CE* and *mixed CE-2nd pass* on validations sets.

LOSS	RO-EN	RU-EN	EN-DE
CE	33.82	29.83	26.51
MIXED CE	**34.51**	**30.46**	26.88
DOUBLE MIXED CE	34.23	**30.46**	**27.06**
MIXED CE-2ND PASS	33.84	30.16	26.83

4. Experiments

- The effect of α_i

TF: $\mathcal{L}_{mix} = -\sum_{t=1}^{n}\left[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\right]$

SS: $\mathcal{L}_{mix} = -\sum_{t=1}^{n}\left[(1-\alpha_i)\cdot \log p_\theta(y_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x}) + \alpha_i \cdot \log p_\theta(\hat{y}_t|\boldsymbol{y}_{<t}^{\text{mix}}, \boldsymbol{x})\right]$

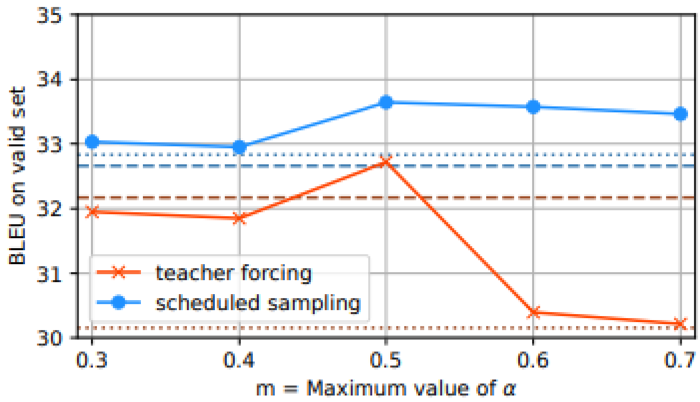


Figure 5. BLEU scores on the WMT'16 Ro-En validation set with different m values. The blue and orange dotted lines denote the BLEU scores of the model with $\alpha_i = 0.5$ while the dashed lines denote the result of training with CE loss.

5. Conclusion

- Introducing **mixed cross entropy (mixed CE)** loss in teacher forcing and scheduled sampling training

- In teacher forcing, mixed CE exploits the model's greedy predictions during training to learn a one-to-many mapping.
    - Superior performance in single reference set, multi-reference set, paraphrased reference set.

- In scheduled sampling, mixed CE can mitigate exposure bias more effectively by encouraging the model to produce similar outputs given different inputs from different distributions.

Thanks!

haoran2_li@mymail.sutd.edu.sg