Parallel Refinements for Lexically Constrained Text Generation with BART

Advisor :Jia-Ling, Koh

Speaker : Shu-Ming Yu

Source : EMNLP'21

Date : 2024/03/26

Outline

- **Introduction**
- Method
- Experiment
- Conclusion

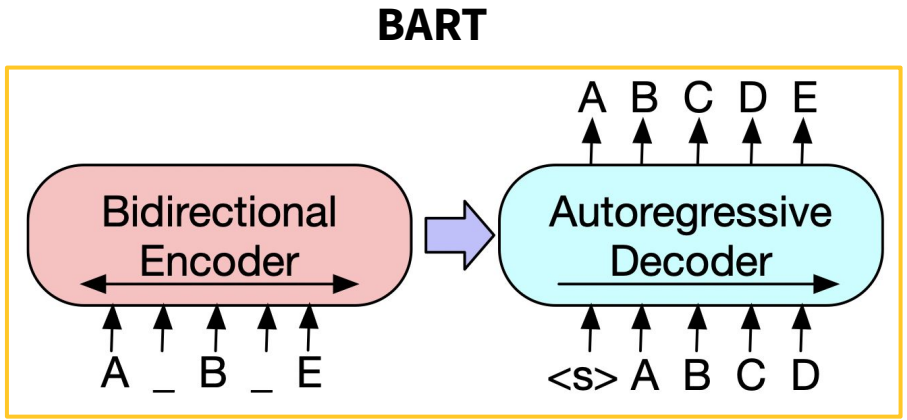
Introduction

- Lexically Constrained Text Generation

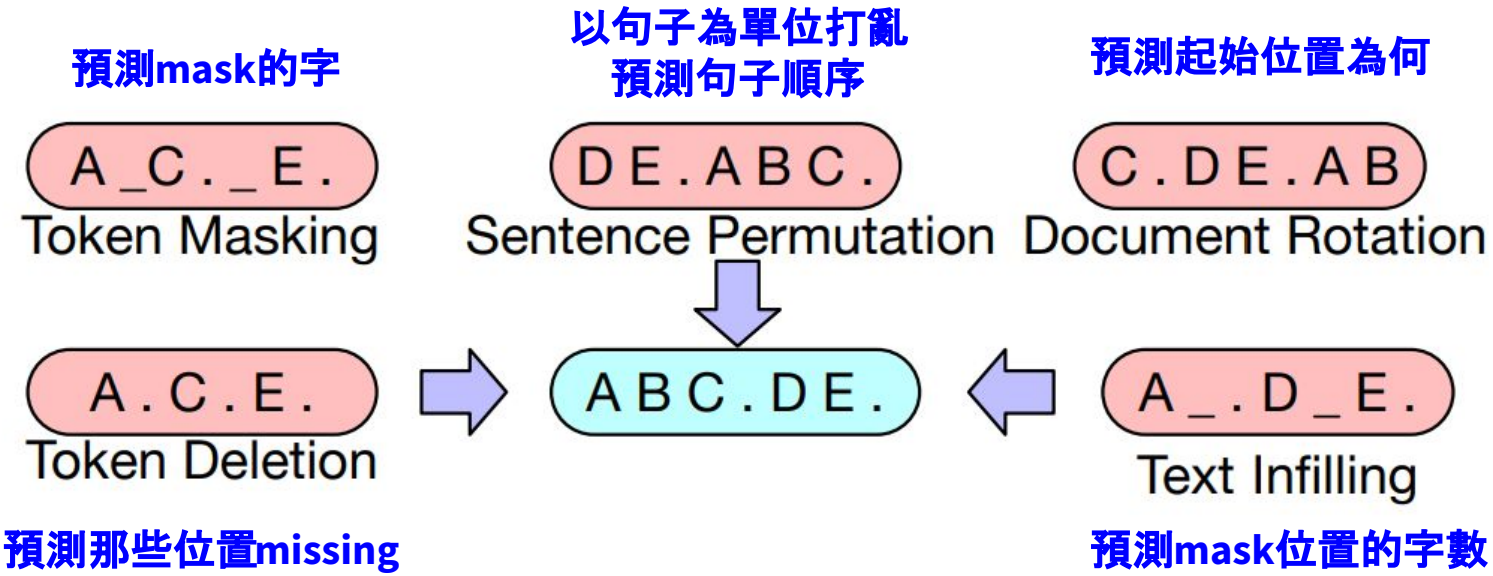
Cons	**lovely**, **time**, **forever**, **try**
Step 1	this **lovely** experience **time** and **forever** to **try** .
Step 2	this was **lovely** experience ! **time** and it **forever** to **try** this .
Step 3	this was a **lovely** experience ! first **time** here and it took **forever** to **try** this place .

Introduction

- Text generation



Introduction

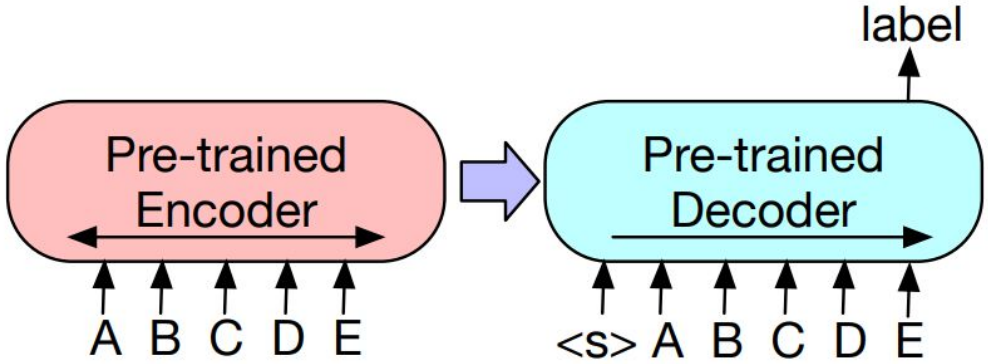


Introduction

- Sequence classification
- Token classification
- Sequence generation
- Machine translation

Introduction

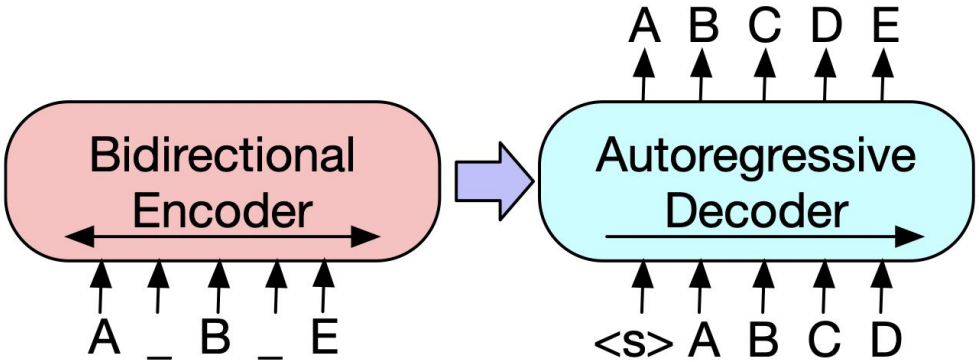
- Sequence classification
- Token classification



Same Input

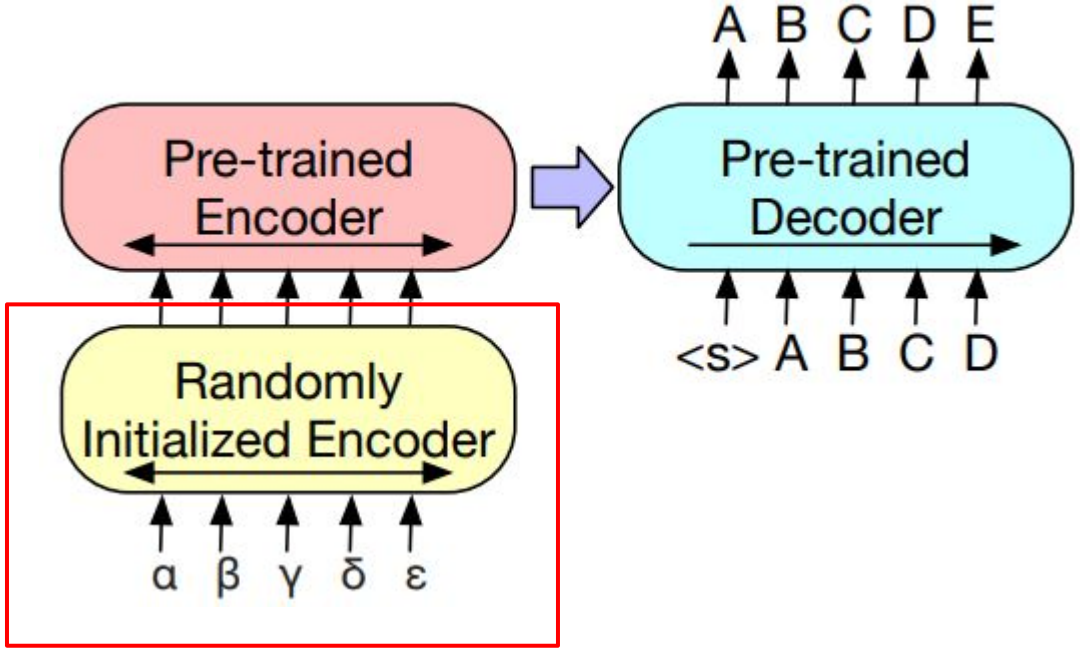
Introduction

- Sequence generation



Introduction

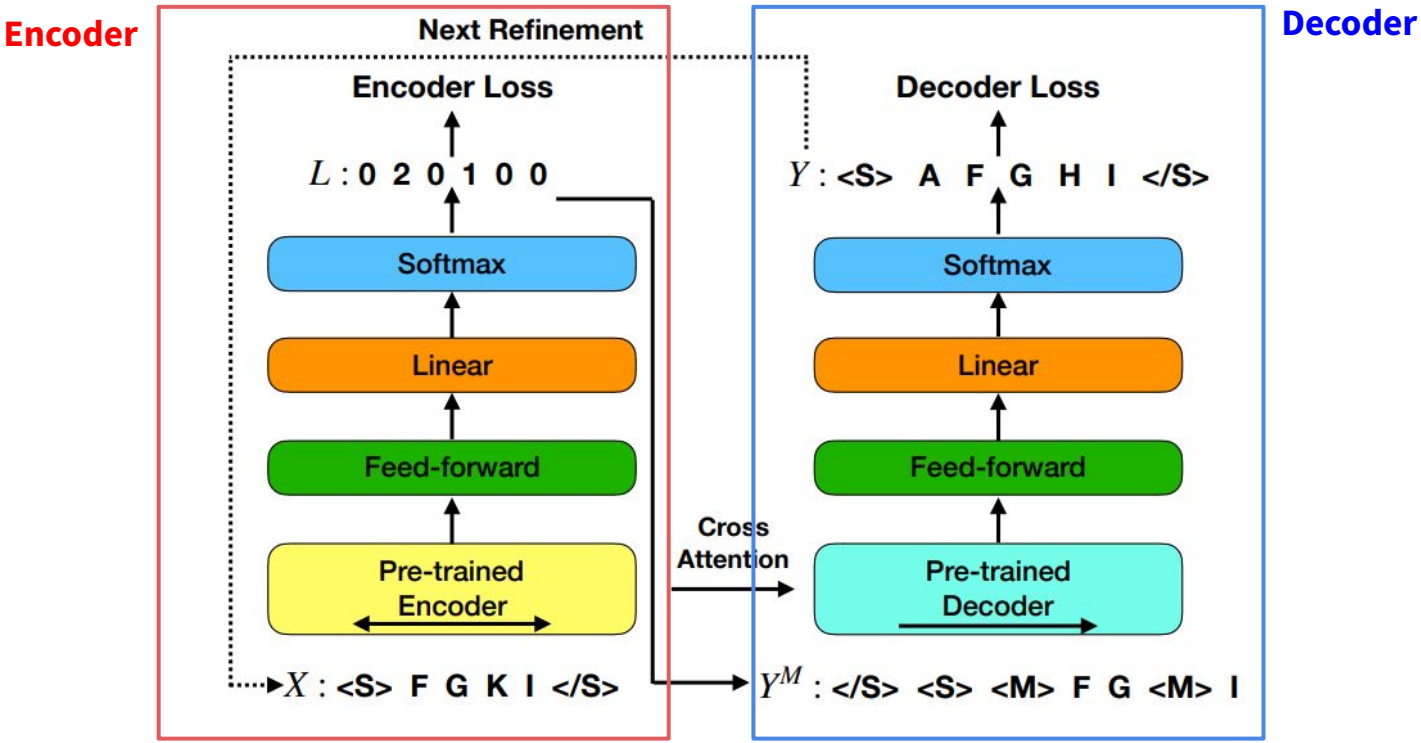
- Machine translation



Outline

- Introduction
- **Method**
- Experiment
- Conclusion

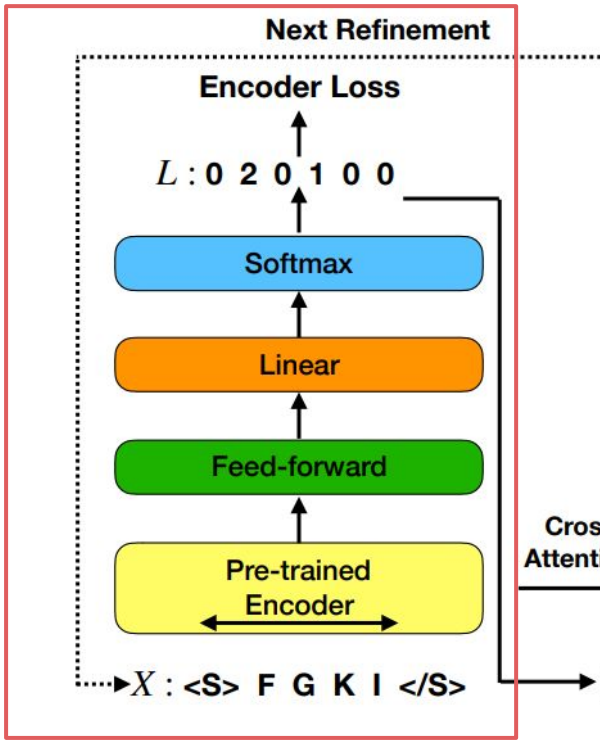
Method


Encoder
Decoder
Next Refinement
Encoder Loss
L : 0 2 0 1 0 0
Softmax
Linear
Feed-forward
Pre-trained Encoder
X : <S> F G K I </S>
Cross Attention
Decoder Loss
Y : <S> A F G H I </S>
Softmax
Linear
Feed-forward
Pre-trained Decoder
Y^M : </S> <S> <M> F G <M> I

Method

Input : $D = \{(X, L, Y^M, Y)\}$

X : Input sentence
L : Action label
(0: copy , 1: replace, 2: insert)



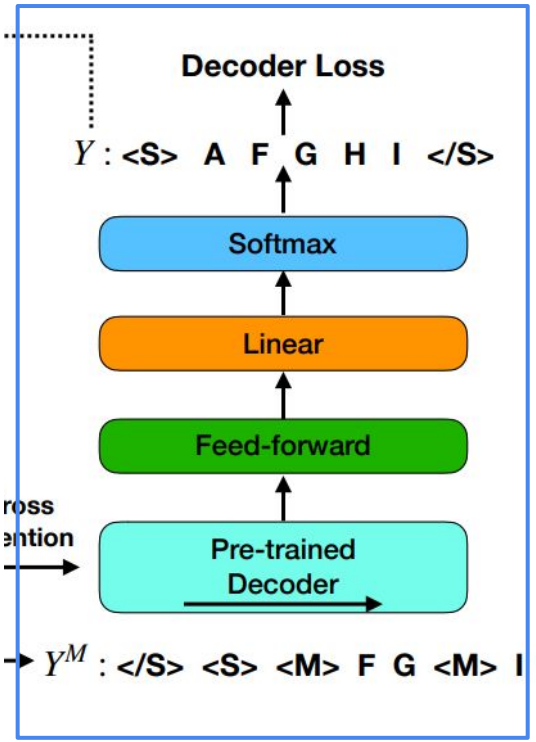
Method

Input : $D = \{(X, L, Y^M, Y)\}$

X : Input sentence
L : Action label
(0: copy , 1: replace, 2: insert)

Y^M : Masked sentence
Y : Output sentence

Decoder



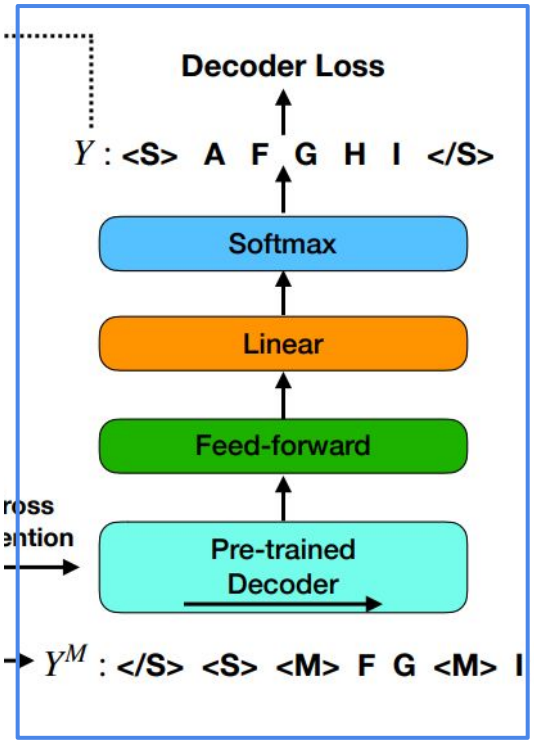
Method

Input : $D = \{(X, L, Y^M, Y)\}$

X **:** Input sentence
L **:** Action label
(0: copy , 1: replace, 2: insert)

Y^M **:** Masked sentence
Y **:** Output sentence

Decoder



Method

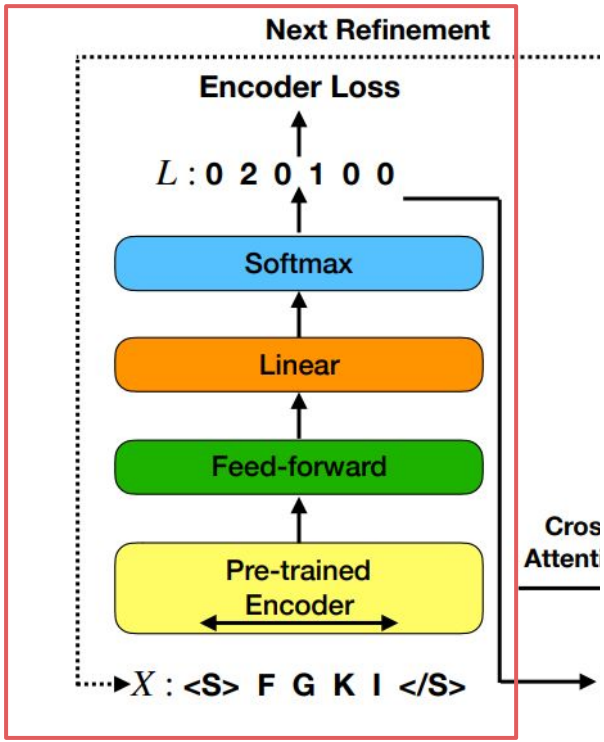
Input : $D = \{(X, L, Y^M, Y)\}$

Example:
Target sentence:
<S> A B C D E F G H I </S>

random select some token
=> <S> F G H I </S>
random replace 15%
=> <S> F G **K** I </S>

X

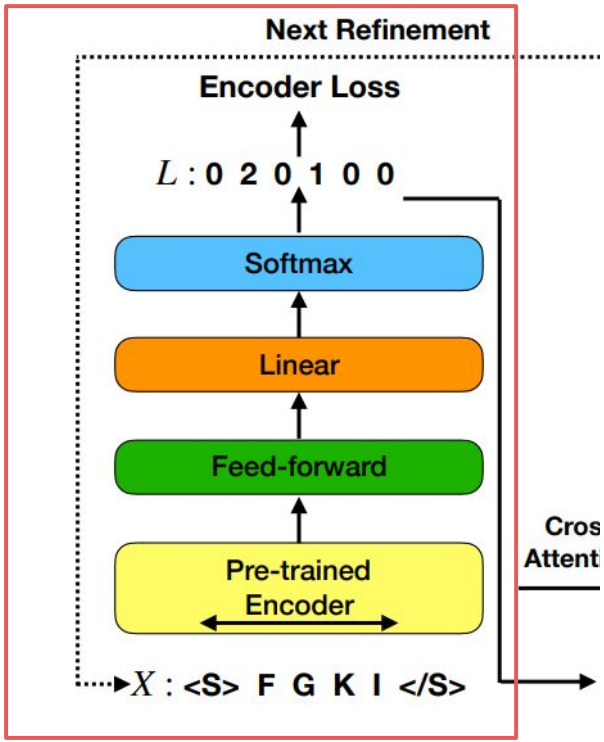
Encoder



Method

Input : $D = \{(X, L, Y^M, Y)\}$

Example:
Target sentence:
<S> A B C D E F G H I </S>



Method

Encoder

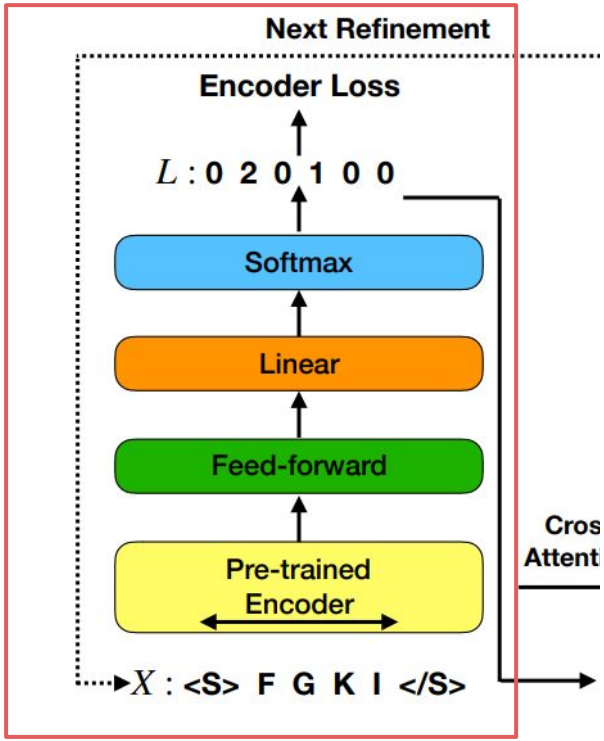
Input : $D = \{(X, L, Y^M, Y)\}$

Example:

<S>	F	G	K	I	</S>
0	2	0	1	0	0

L

$$L_{encoder} = -\frac{1}{n}\sum_{t=1}^{n} \log p(l_t | x_1, \ldots, x_n).$$



Method

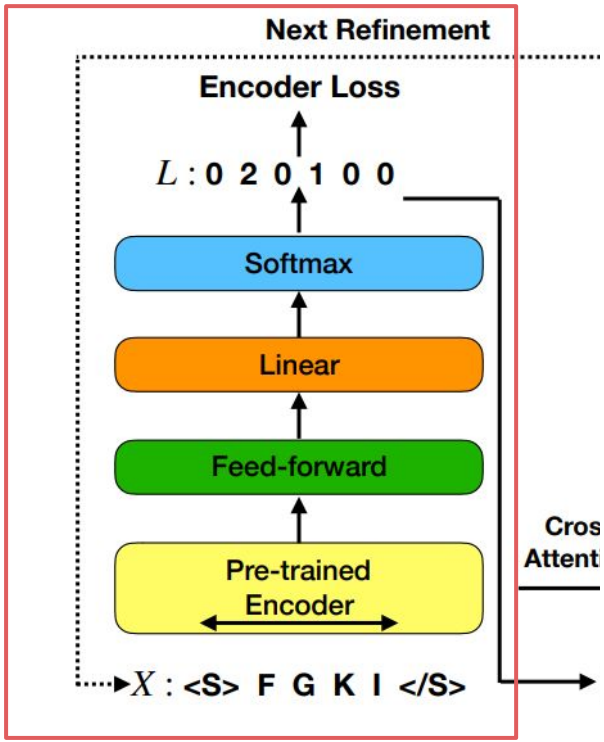
Input : $D = \{(X, L, Y^M, Y)\}$

Example:

<S>	F	G	K	I	</S>
0	2	0	1	0	0

L

=> <S> <mask> F G <mask> I </S>



Method

Input : $D = \{(X, L, Y^M, Y)\}$

Example:

<S>	F	G	K	I	</S>
0	2	0	1	0	0

L

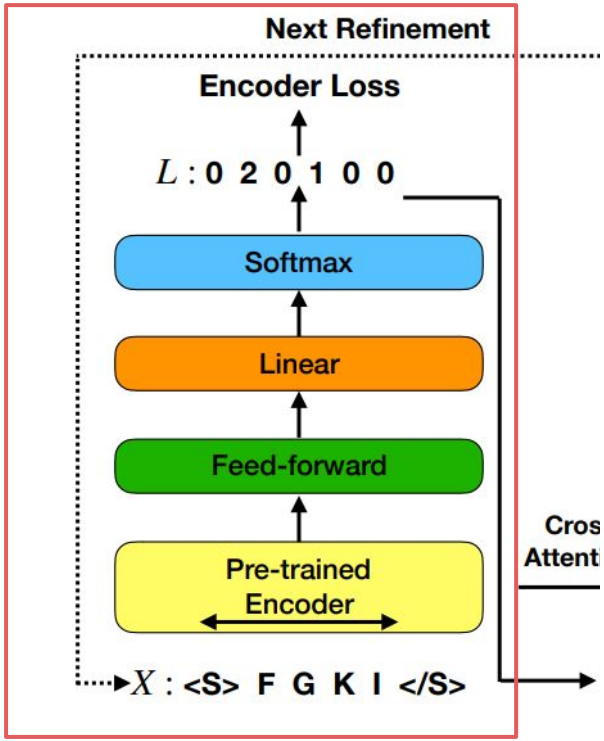
=> <S> <mask> F G <mask> I </S>

Shift right

=> </S> <S> <mask> F G <mask> I

Y^M

Encoder



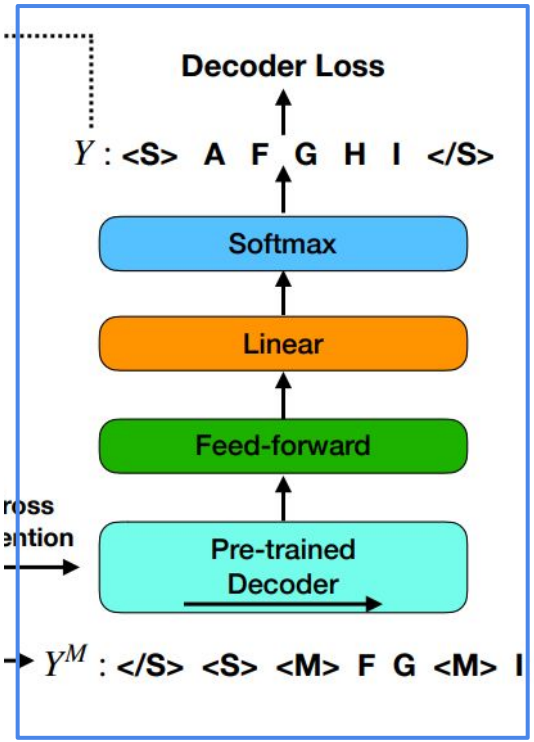
Method

Input : $D = \{(X, L, Y^M, Y)\}$

Example:
Y^M : </S> <S> <mask> F G <mask> I

$$L_{decoder} = -\frac{1}{m}\sum_{t=1}^{m}\log p(y_t|X, y_{<t}^M).$$

Decoder



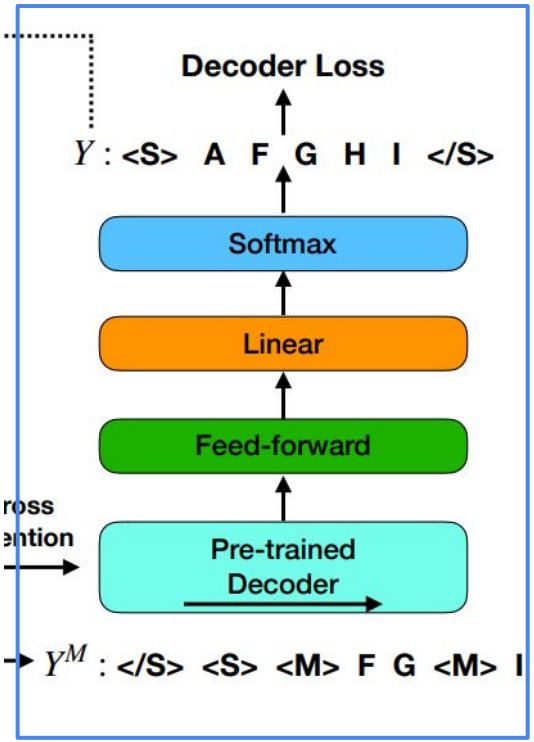
Method

Input : $D = \{(X, L, Y^M, Y)\}$

Example:
Y^M : </S> <S> <mask> F G <mask> I

$$L_{decoder} = -\frac{1}{m}\sum_{t=1}^{m}\log p(y_t|X, y_{<t}^M).$$

Decoder



Loss

1

$$L_{total} = L_{encoder} + \alpha L_{decoder},$$

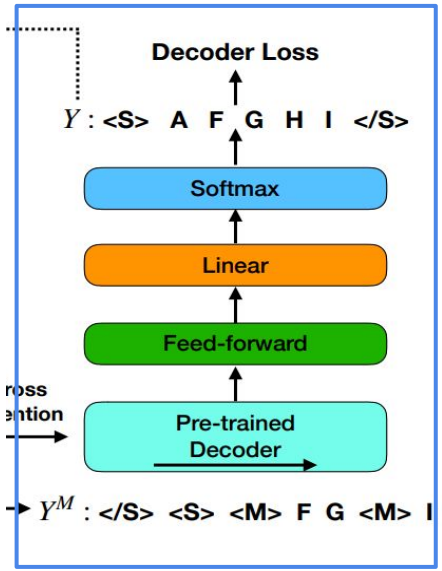
Method

- **Greedy Decoding**: Select the highest probability token

$$\hat{y}_t^r = \begin{cases} \arg\max p(y_t^r | X^r, y_{<t}^M), & y_{t+1}^M = <\mathrm{M}> \\ y_{t+1}^M, & y_{t+1}^M != <\mathrm{M}>. \end{cases}$$

- **Top-k Decoding**: Sample token from the k most probable

- **Top-p Decoding**: Sample token from the smallest possible set of probability exceed probability p

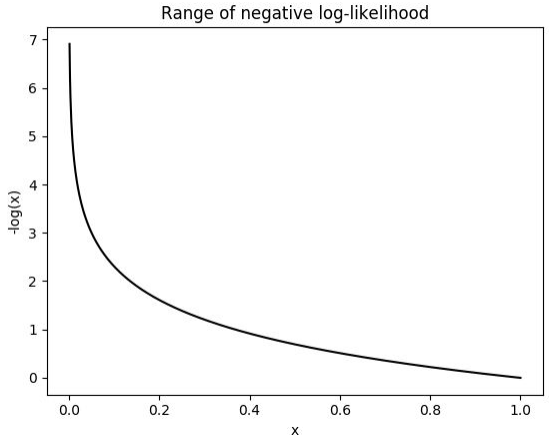


Method

- **Multiple-sequence Decoding:**

Generate multiple possible sentences during the generation process.
Then choose the lowest negative log-likelihood.

$$L(y) = -log(y)$$



Method

- **Repetition Penalty**:

$$p(\hat{y}_t^r = i) = \frac{exp(h_i / I(i \in Y^M))}{\sum_j exp(h_j / I(j \in Y^M))},$$

$I(c)$ True => θ > 1
False => 1

r round, t-th position

Outline

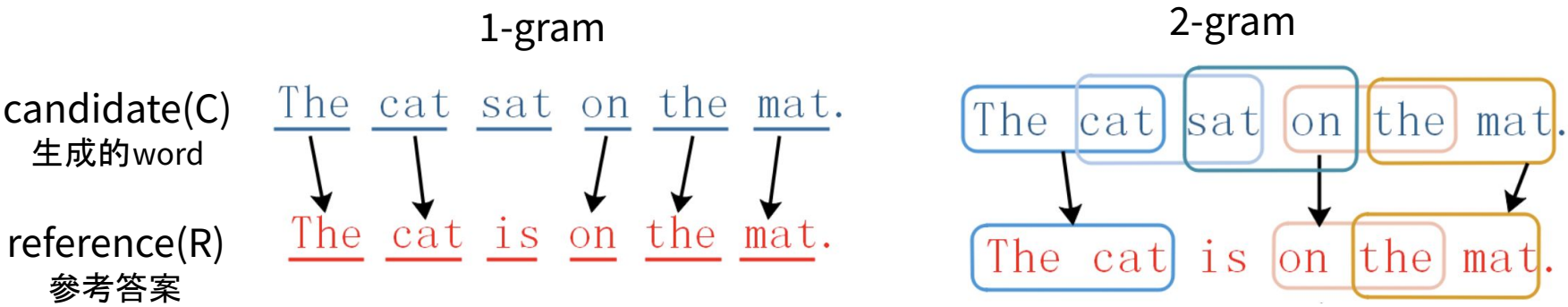
- Introduction
- Method
- **Experiment**
- Conclusion

Experiment

- **Automated Evaluation**
  - BLEU(B-2, B-4)
  - METEOR(M)
  - Self-BLEU(SB-4)
  - distinct(D-2, D-4)

Experiment

- **N-gram**



Experiment

- **Evaluation**
  - BLEU as **precision**

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log\, p_n)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Candidate中的uni-gram有幾個
也出現在reference中

candidate中的uni-gram有幾個

1-gram

candidate(C)
生成的word

The cat sat on the mat.

reference(R)
參考答案

The cat is on the mat.

$$p_n = \frac{5}{6}$$

Experiment

- **Evaluation**
  - BLEU

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log\, p_n)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

$$Count_{clip} = \min(Count, Max\_Ref\_Count)$$

candidate中這個uni-gram出現的次數

所有reference 中這個uni-gram出現最多的次數

1-gram

candidate(C)
生成的word

The the the

reference(R)
參考答案

The cat is on the mat.

$$p_n = \frac{\cancel{3}\ 2}{3}$$

$Count_{clip}$ = min(3, max(2))
= 2

Experiment

- **Evaluation**
  - BLEU

$$BLEU = \boxed{BP} \cdot exp(\sum_{n=1}^{N} w_n log\, p_n)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

1-gram

candidate(C)
生成的word

The cat sat on the mat.

reference(R)
參考答案

The cat is on the mat.

P1 = 5/6

$$BP = \begin{cases} 1 & lc > \text{lr} \\ \exp(1 - lr/lc) & lc \le lr \end{cases}$$

> 1

< 1

$lc =$ 生成的word的長度

$lr =$ 最短的參考答案的長度

Experiment

Candidate : 生成的word

Reference : 參考答案

- **Evaluation**
  - METEOR

$$METEOR = (1 - pen) \times F_{means}$$

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha)R}$$

$\alpha = 0.5$

=> F_{means} as F-1

P : $\frac{\text{Candidate的uni-gram也出現在reference中的數量}}{\text{Candidate的長度}}$ as **precision**

R : $\frac{\text{Candidate的uni-gram也出現在reference中的數量}}{\text{reference的長度}}$ as **recall**

Experiment

Candidate : 生成的word

Reference : 參考答案

- **Evaluation**
  - METEOR

$$METEOR = (1 - pen) \times F_{means}$$

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha)R}$$

P : precision
R : recall

$$Pen = \frac{\#chunks}{m}$$

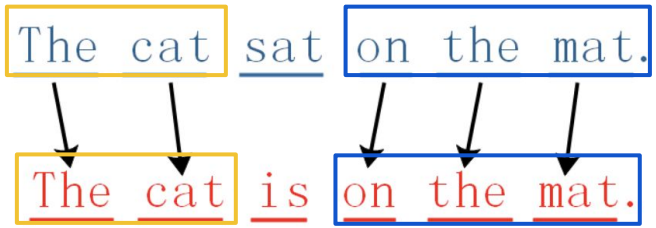
m : number of match

$$Pen = \frac{2}{5}$$

1-gram

candidate(C)
生成的句子

reference(R)
參考答案



Experiment

- **Automated Evaluation**
  - Self-BLEU(SB-4)

    Average BLEU score for every generated sentences
    (one sentence as hypothesis and the others as reference)

  - distinct(D-2, D-4)

$$\frac{\textbf{The number of distinct bigrams and four-grams}}{\textbf{total number of generated words}}$$

Experiment

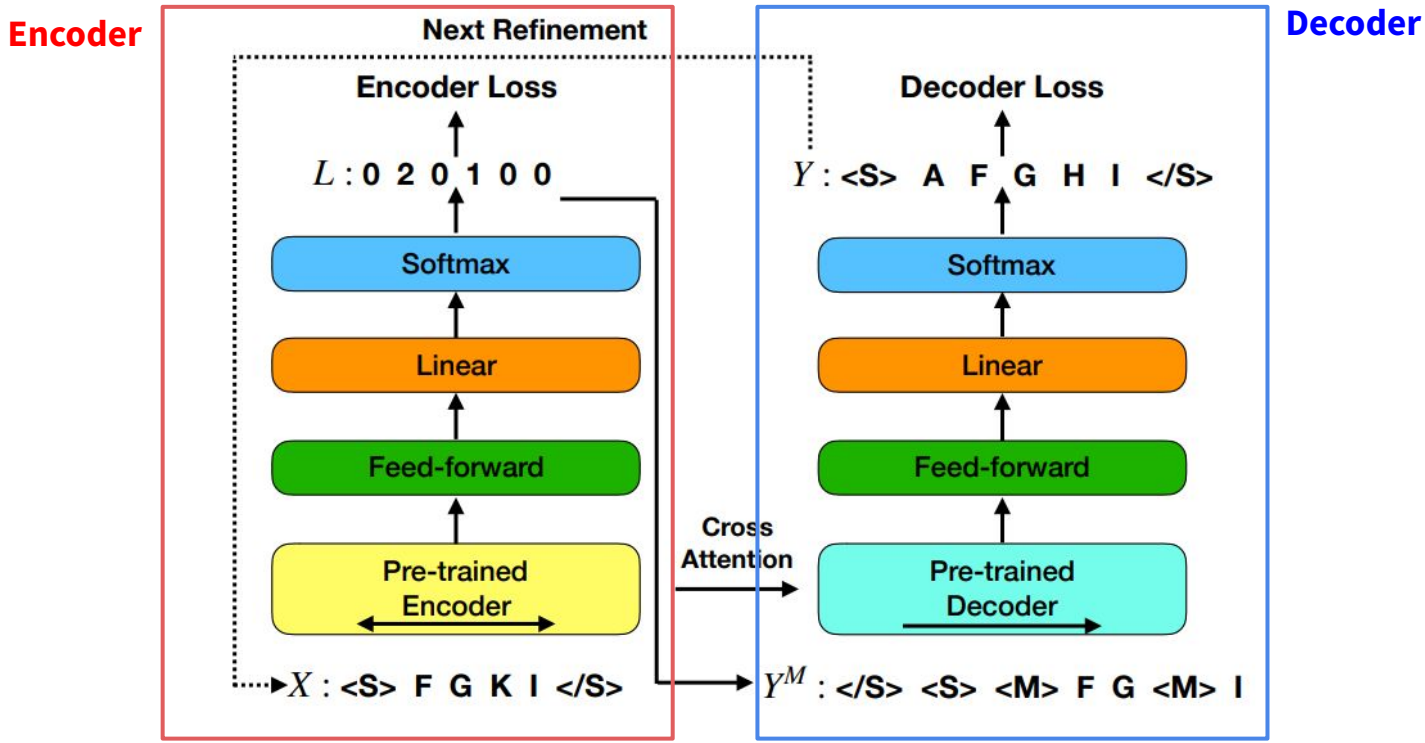
- **Dataset**
  - One-Billion-word

    A public dataset for language modeling produced
  - Yelp

    Business review on Yelp

句子長度限制 10~40
Training: 1M
Validation: 0.1M
Test: 1K
Number of keyword: 1~6

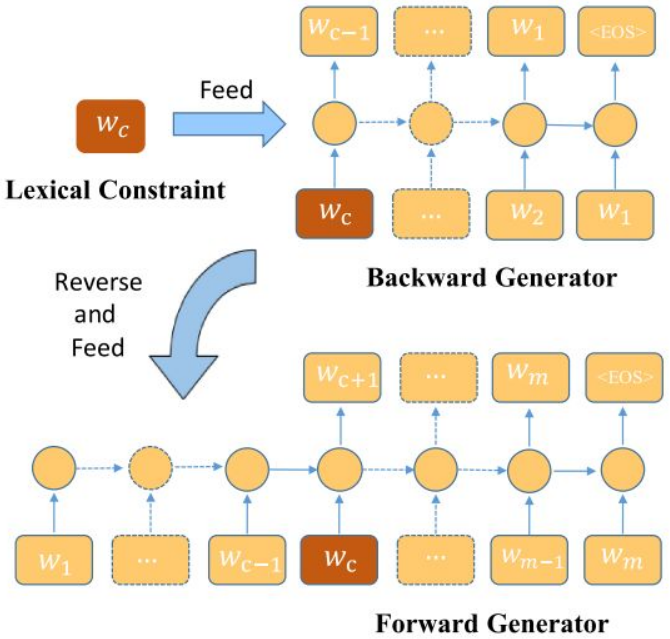
Method


Encoder
Decoder
Next Refinement
Encoder Loss
L : 0 2 0 1 0 0
Softmax
Linear
Feed-forward
Pre-trained Encoder
X : <S> F G K I </S>
Cross Attention
Decoder Loss
Y : <S> A F G H I </S>
Softmax
Linear
Feed-forward
Pre-trained Decoder
Y^M : </S> <S> <M> F G <M> I

Experiment

- **Baseline**
  - Sep-B/F
  - Asyn-B/F

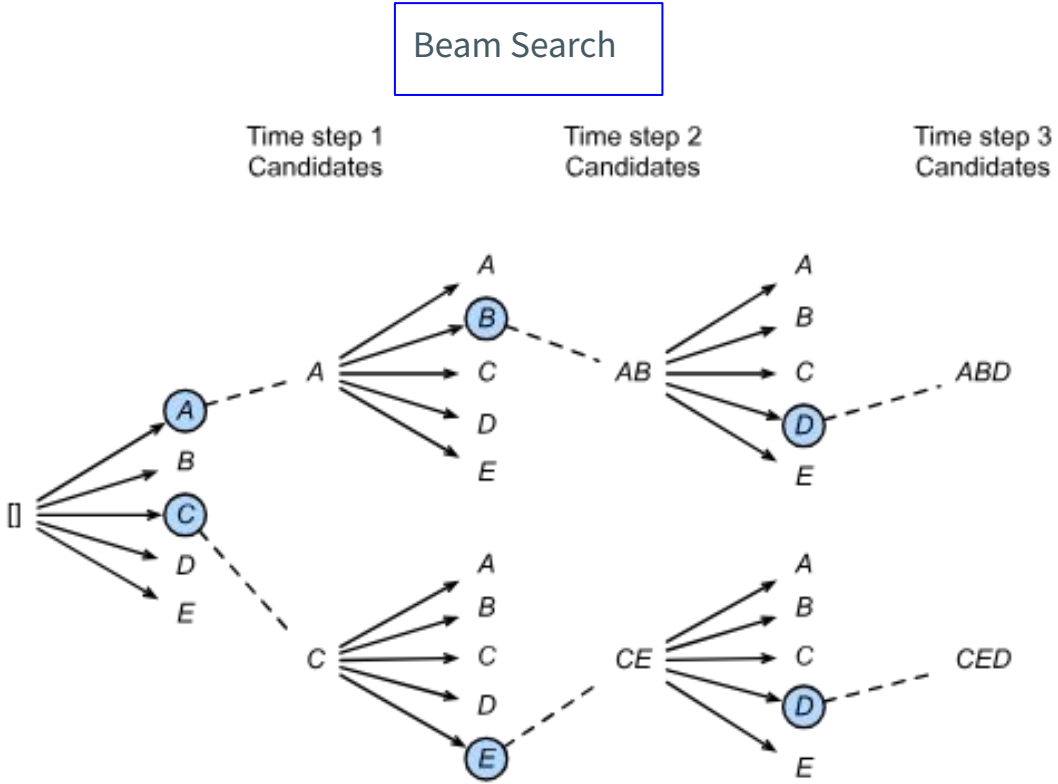
差在forward、backward是否同時訓練



只能限制1個word

Experiment

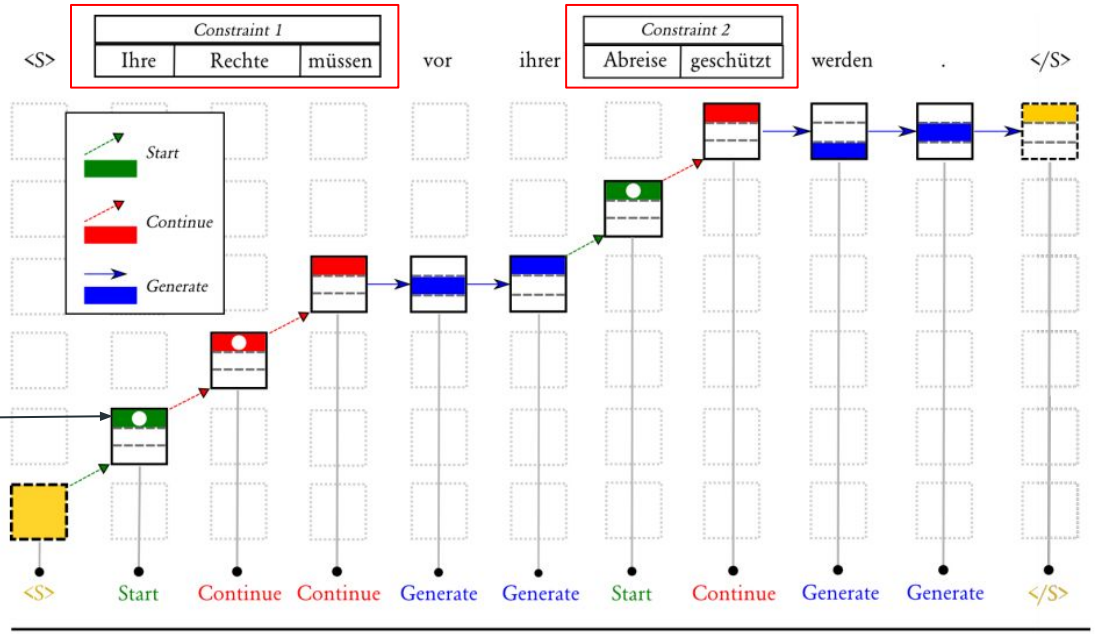
- **Baseline**
  - GBS
    (decoder based)



Experiment

- **Baseline**
  - GBS
    (decoder based)

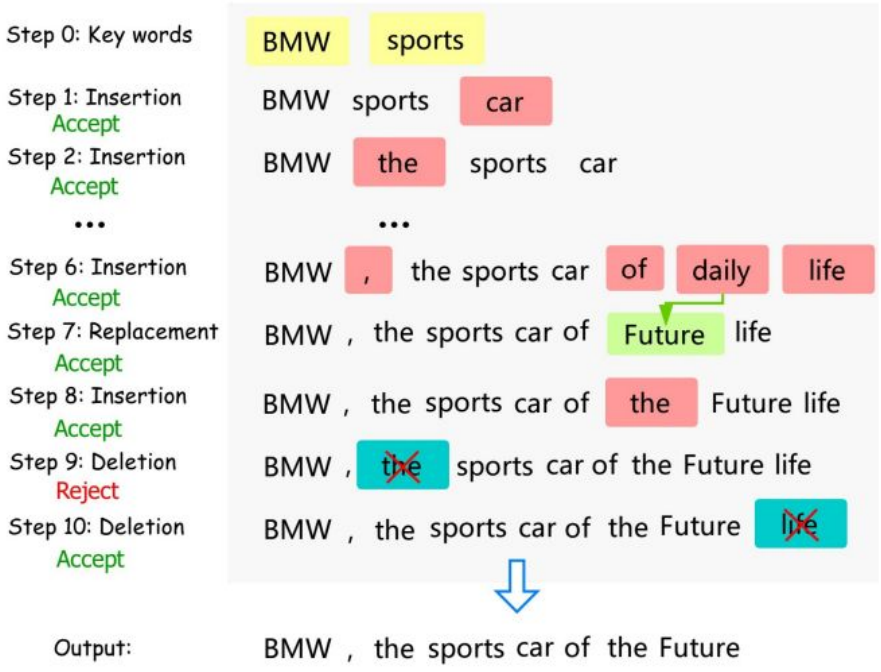
下一個字
只能生成指定的word



Experiment

- **Baseline**
  - CGMH

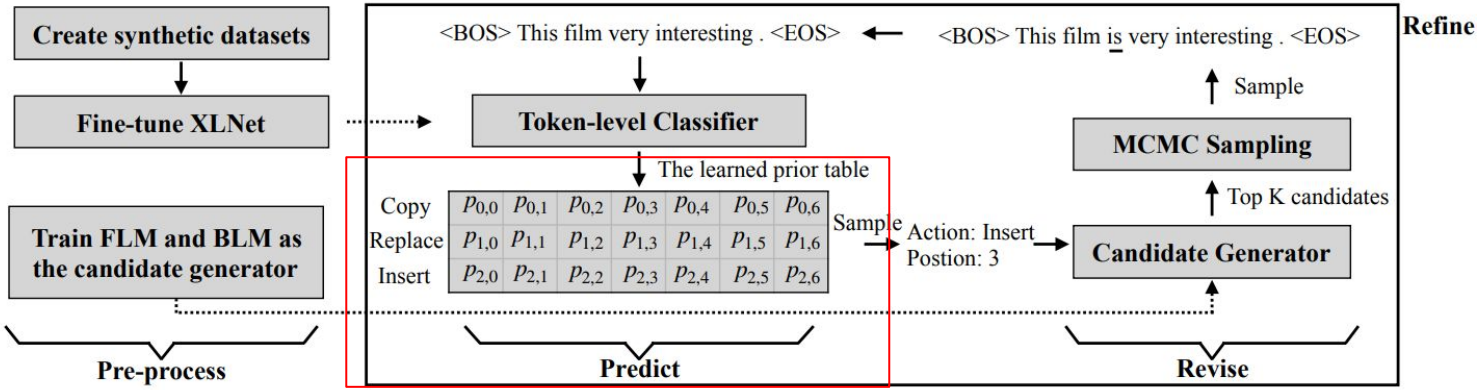
一次一個字
隨機選字
隨機位置
隨機行為(3種)



Experiment

- **Baseline**
  - X-MCMC-C

    可一次多個字



Experiment

- **Baseline**
  - POINTER

  Encoder-based

  可一次多個字

Stage	Generated text sequence
0 (X^0)	sources sees structure perfectly
1 (X^1)	sources **company** sees **change** structure perfectly **legal**
2 (X^2)	sources **suggested** company sees **reason** change **tax** structure **which** perfectly legal **.**
3 (X^3)	**my** sources **have** suggested **the** company sees **no** reason **to** change **its** tax structure **,** which **are** perfectly legal .
4 (X^4)	my sources have suggested the company sees no reason to change its tax structure , which are perfectly legal .

Experiment

One-Billion-word

Yelp

Generation quality **Generation diversity**

		Metrics	BLEU ↑ B-2	BLEU ↑ B-4	↑ M	↓ SB-4	Distinct ↑ D-2	Distinct ↑ D-4
llion-Word		Human	-	-	-	10.3%	78.1%	99.5%
	Baselines	sep-B/F	4.4%	0.7%	7.0%	52.1%	46.3%	78.8%
		asyn-B/F	4.3%	0.7%	6.8%	50.3%	47.8%	80.9%
		GBS	10.1%	2.8%	13.5%	37.0%	59.3%	87.2%
		CGMH	9.9%	3.5%	13.1%	10.2%	**78.9%**	99.3%
		X-MCMC-C	12.5%	4.1%	13.8%	16.9%	69.7%	98.8%
		POINTER	2.5%	0.1%	10.2%	-	-	-
		POINTER-2	8.7%	1.6%	14.3%	37.3%	46.5%	90.9%
		Greedy	15.6%	**6.6%**	**15.2%**	22.1%	66.6%	97.2%
elp		Human	-	-	-	26.1%	57.7%	97.0%
	Baselines	sep-B/F	6.9%	2.1%	8.7%	67.1%	31.9%	64.6%
		asyn-B/F	7.5%	2.3%	9.0%	68.0%	31.0%	64.3%
		GBS	13.6%	4.5%	15.3%	59.3%	37.5%	70.2%
		CGMH	12.3%	4.6%	14.6%	23.6%	**60.7%**	97.7%
		X-MCMC-C	15.3%	5.4%	15.5%	38.5%	47.4%	92.4%
		POINTER	4.0%	0.3%	13.0%	-	-	-
		POINTER-2	10.6%	2.4%	16.8%	49.1%	35.2%	86.3%
	RT	**Greedy**	19.4%	**9.0%**	**17.4%**	45.1%	44.4%	88.1%

Experiment

Generation quality **Generation diversity**

One-Billion-word

		Metrics	BLEU ↑ B-2	BLEU ↑ B-4	↑ M	↓ SB-4	Distinct ↑ D-2	Distinct ↑ D-4
llion-Word		Human	-	-	-	10.3%	78.1%	99.5%
	Baselines	sep-B/F	4.4%	0.7%	7.0%	52.1%	46.3%	78.8%
		asyn-B/F	4.3%	0.7%	6.8%	50.3%	47.8%	80.9%
		GBS	10.1%	2.8%	13.5%	37.0%	59.3%	87.2%
		CGMH	9.9%	3.5%	13.1%	10.2%	**78.9%**	99.3%
		X-MCMC-C	12.5%	4.1%	13.8%	16.9%	69.7%	98.8%
		POINTER	2.5%	0.1%	10.2%	-	-	-
		POINTER-2	8.7%	1.6%	14.3%	37.3%	46.5%	90.9%
		Greedy	15.6%	**6.6%**	**15.2%**	22.1%	66.6%	97.2%

Yelp

elp		Human	-	-	-	26.1%	57.7%	97.0%
	Baselines	sep-B/F	6.9%	2.1%	8.7%	67.1%	31.9%	64.6%
		asyn-B/F	7.5%	2.3%	9.0%	68.0%	31.0%	64.3%
		GBS	13.6%	4.5%	15.3%	59.3%	37.5%	70.2%
		CGMH	12.3%	4.6%	14.6%	23.6%	**60.7%**	97.7%
		X-MCMC-C	15.3%	5.4%	15.5%	38.5%	47.4%	92.4%
		POINTER	4.0%	0.3%	13.0%	-	-	-
		POINTER-2	10.6%	2.4%	16.8%	49.1%	35.2%	86.3%
	RT	**Greedy**	19.4%	**9.0%**	**17.4%**	45.1%	44.4%	88.1%

Experiment

One-Billion-word

Yelp

		Metrics	↓ Ref	↓ La
One-Billion-Word		Human	-	-
	Baselines	sep-B/F	-	1.900
		asyn-B/F	-	1.865
		GBS	-	9.234
		CGMH	200	9.871
		X-MCMC-C	200	31.41
		POINTER	6	-
		POINTER-2	6	0.727
		Greedy	**4.8**	**0.351**
Yelp		Human	-	-
	Baselines	sep-B/F	-	1.807
		asyn-B/F	-	1.771
		GBS	-	8.634
		CGMH	200	11.09
		X-MCMC-C	200	31.68
		POINTER	6	-
		POINTER-2	6	0.741
		Greedy	**4.9**	0.357

Ref: 生成回合數
La ：時間
Rep: 重複n-gram比例
Len: 平均長度

Experiment

- Human evaluation: Fluency、Informative

Compare two sentences generated by different models

Fluency: which sentence is more fluent?				
Model A won		Tied	Model B won	
CBART	32.7%	24.0%	**43.3%**	Human
CBART	**70.0%**	14.0%	16.0%	GBS
CBART	**56.0%**	25.3%	18.7%	CGMH
CBART	**44.7%**	33.3%	22.0%	X-MCMC-C
CBART	**77.3%**	18.0%	4.7%	POINTER-2
Informativeness: which sentence is more informative?				
Model A won		Tied	Model B won	
CBART	8.7%	9.3%	**82.0%**	Human
CBART	**65.3%**	10.7%	24.0%	GBS
CBART	**52.7%**	18.0%	29.3%	CGMH
CBART	**33.3%**	35.4%	31.3%	X-MCMC-C
CBART	**48.0%**	12.0%	40.0%	POINTER-2

Experiment

- More number of constraints boots the generation quality

Metrics CBART	↑ B-2	↑ M	↓ SB-4	Rep
$N = 1$	5.7%	8.3%	55.2%	0.3%
$N = 2$	9.7%	11.8%	52.3%	0.7%
$N = 3$	16.0%	15.7%	45.5%	1.8%
$N = 4$	22.4%	19.4%	42.2%	2.8%
$N = 5$	28.0%	22.7%	38.7%	3.6%
$N = 6$	**34.5%**	**26.3%**	**36.5%**	4.7%

Conclusion

- CBART achieves lexical constraint based on BART
- CBART can generate fluent and diverse text and dramatically reduce the inference time