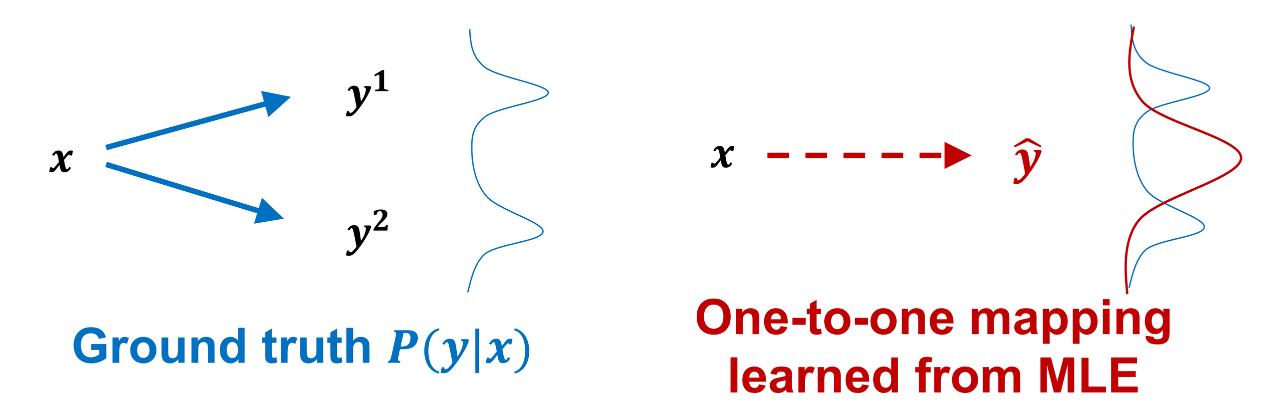
Mixture Content Selection for Diverse Sequence Generation

Jaemin Cho[1] Minjoon Seo[2,3] Hannaneh Hajishirzi[1,3]

Allen Institute of AI[1] Clova AI, NAVER[2] University of Washington[3]

Motivation

- Many NLP tasks require one-to-many mapping ($p(y|x)$ is multi-modal)
  - ex) Question Generation, Summarization
- RNN Encoder-Decoder (Seq2Seq) is ***not*** designed for one-to-many relationship
- Training with maximum likelihood estimation (MLE) can yield ***suboptimal mapping***

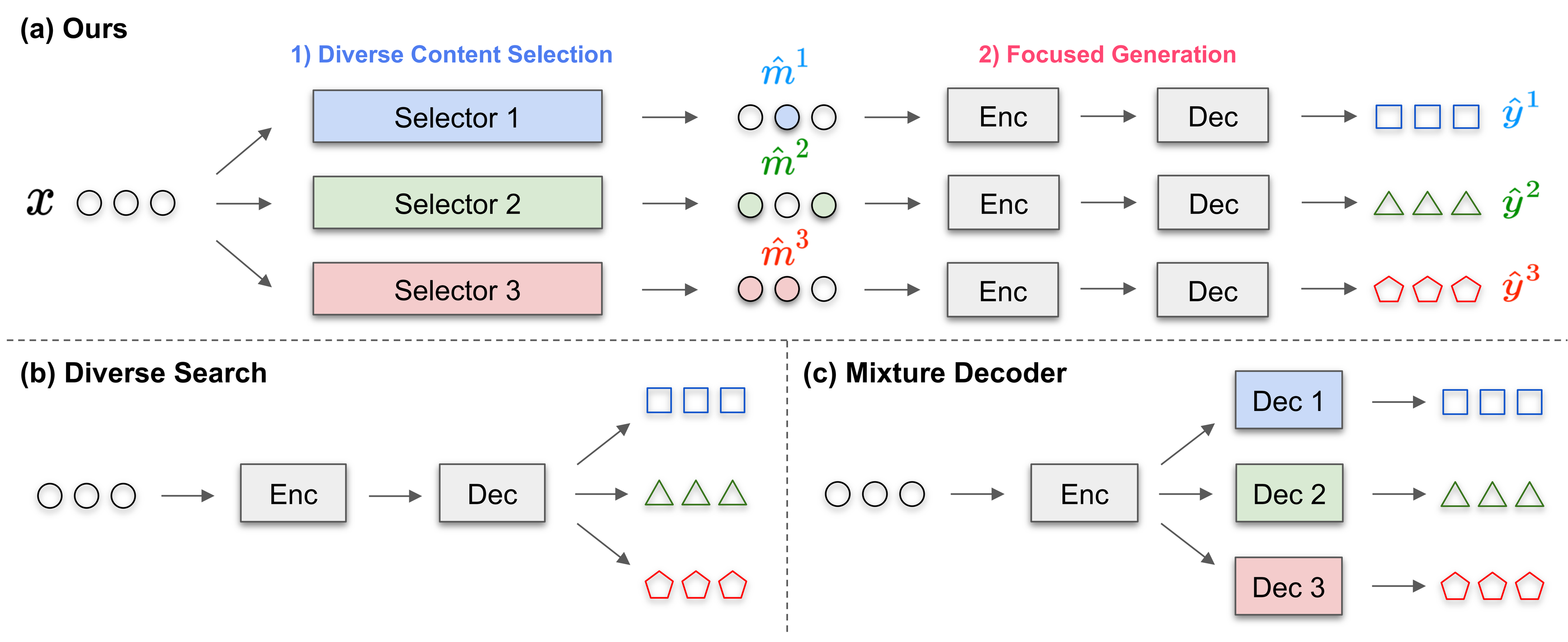


Our Contributions

- Promoting diversity with **Two-stage Generation**
  - 1) Diverse Content Selection => One-to-Many Relationship
  - 2) Focused Generation => One-to-One Relationship
- Significant improvements on Accuracy & Diversity
- Can be added to any sequence generation models

Overview

Current approaches	Ours
Diverse Decoding Manipulate ***what's already encoded*** Similar semantics in different expressions (paraphrasing)	Diverse Encoding Manipulate ***where to focus*** High semantic diversity



Two-stage Generation

- **Factorizing $p(y|x)$** with latent variable Focus m
  - $p(y|x) = \sum_m \underbrace{p(m|x)}_{\text{1) Diverse Content Selection}} * \underbrace{p(y|x,m)}_{\text{2) Focused Generation}}$
- Diverse Content Selection with **Mixture-of-Experts Selector**
  - $p(m|x) = \frac{1}{K} \sum_z^{1\ldots K} p(m|x,z)$
- Shape of Focus m
  - **Binary masks** on the source sequence ($|x| = |m|$)
  - t-th focus m_t: whether t-th token x_t is focused during generation

Ex)
x: Jaemin attended EMNLP

m^1	1	0	0	y^1: Who is Jaemin?
m^2	0	1	0	y^2: Where did he attend?
m^3	0	0	1	y^3: How was EMNLP?

Selector $p(m|x)$

- Sample binary masks m
- Mixture of experts => Diverse masks
- $p(m|x,z) = \sigma(FC([GRU(x); \underbrace{emb_z}_{\text{z-th expert embedding}}]))$

Generator $p(y|x,m)$

- Standard encoder-decoder (Seq2Seq)
- Generate sequences guided by masks
- Input: $[\underbrace{emb_x}_{\text{Source token embedding}}; \underbrace{emb_{mask}}_{\text{Mask embedding}}]$

1) Diverse Content Selection	2) Focused Generation
In December 1878, Tesla left Graz and severed all relations with his family to hide the fact that he dropped out of school.	What did Tesla do?
In December 1878, Tesla left Graz and severed all relations with his family to hide the fact that he dropped out of school.	What did Tesla do in December 1878?
In December 1878, Tesla left Graz and severed all relations with his family to hide the fact that he dropped out of school.	What did Tesla do to hide he dropped out of school?

Training

- No ground truth supervision for Focus m
  => **Focus Guide: overlap between source & target**
- Selector Training with **Hard-EM**
  Sample K masks from K-mixture Selector
  (E-step) Choose a mask closest to the target
  (M-step) Backprop with the chosen mixture

Algorithm 1: Training
(N: Dataset size, K: Number of mixtures)

Data: $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{m}^{\text{guide (i)}})\}_{i=1}^N$
for $i \in \{1 \ldots N\}$ **do**
  /* Selector $p_\phi(\mathbf{m}|\mathbf{x}, z)$ E-step */
  for $z \in \{1 \ldots K\}$ **do**
    $L_{\text{select}}^{(i)\ z} = -\log p_\phi(\mathbf{m}^{\text{guide}\ (i)}|\mathbf{x}^{(i)}, z)$
  end
  $z^{\text{best}\ (i)} = \operatorname{argmin}_z L_{\text{select}}^{(i)\ z}$
  /* Selector $p_\phi(\mathbf{m}|\mathbf{x}, z)$ M-step */
  $\phi_{z^{\text{best}\ (i)}} = \phi_{z^{\text{best}\ (i)}} - \alpha \nabla_{\phi_{z^{\text{best}\ (i)}}} L_{\text{select}}^{(i)\ z^{\text{best}\ (i)}}$
  /* Generator $p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{m})$ Update */
  $L^{(i)}{}_{\text{gen}} = -\log p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \mathbf{m}^{\text{guide (i)}})$
  $\theta = \theta - \alpha \nabla_\theta L_{\text{gen}}^{(i)}$
end

Experiments

Tasks

	Question Generation	Abstractive Summarization
Backbone	NQG++ (Zhou et al. 2017)	Pointer Generator (See et al. 2017)
Dataset	SQuAD	CNN-DM

Baselines

Diverse Search / Sampling	Mixture of Experts + Greedy Decoding
Diverse Beam Search (Vijayakumar et al. 2018) Top-k Sampling (Fan et al. 2018)	Mixture Decoder (Shen et al. 2019) **Ours**

Significant improvements on Accuracy / Diversity / Human preference

- SQuAD Question Generation

Method	BLEU-4 (Top-1)	Oracle (Top-K)	Pairwise (Self-sim)
NQG++	13.27	-	-
Search-based Methods			
3-Beam	13.590	16.848	67.277
5-Beam	13.526	18.809	74.674
3-D. Beam	13.696	16.989	68.018
5-D. Beam	13.379	18.298	74.795
3-T. Sampling	11.890	15.447	**37.372**
5-T. Sampling	11.530	17.651	45.990
Mixture of Experts + Greedy Decoding			
3-M. Decoder	14.720	19.324	51.360
5-M. Decoder	15.166	21.965	58.727
3-M. SELECTOR (Ours)	**15.874**	20.437	47.493
5-M. SELECTOR (Ours)	15.672	**22.451**	59.815
Focus Guide during Test Time			
5-Beam + Focus Guide	24.580		-

- CNN-DM Abstractive Summarization

Method	ROUGE-2 (Top-1)	Oracle (Top-K)	Pairwise (Self-sim)
PG	17.28	-	-
Search-based Methods			
3-Beam	16.533	18.509	85.598
5-Beam	16.634	19.442	84.765
3-D. Beam	16.667	18.722	85.496
5-D. Beam	16.632	19.659	84.043
3-T. Sampling	12.914	17.068	17.306
5-T. Sampling	13.049	19.161	**16.720**
Mixture of Experts + Greedy Decoding			
3-M. Decoder	15.854	21.214	43.168
5-M. Decoder	16.104	21.801	67.196
3-M. SELECTOR (Ours)	17.930	21.316	51.092
5-M. SELECTOR (Ours)	**18.309**	**22.511**	47.280
Focus Guide during Test Time			
5-Beam + Focus Guide	42.757		-

- Human Evaluation

	Diversity (%)			Accuracy (%)		
Baselines	Win	Lose	Tie	Win	Lose	Tie
vs. 3-D. Beam	**49.7**	31.3	19.0	**43.9**	36.9	19.2
vs. 3-T. Sampling	**46.7**	35.1	18.2	**45.3**	36.1	18.6
vs. 3-M. Decoder	**47.6**	32.5	19.9	**41.8**	36.0	22.2

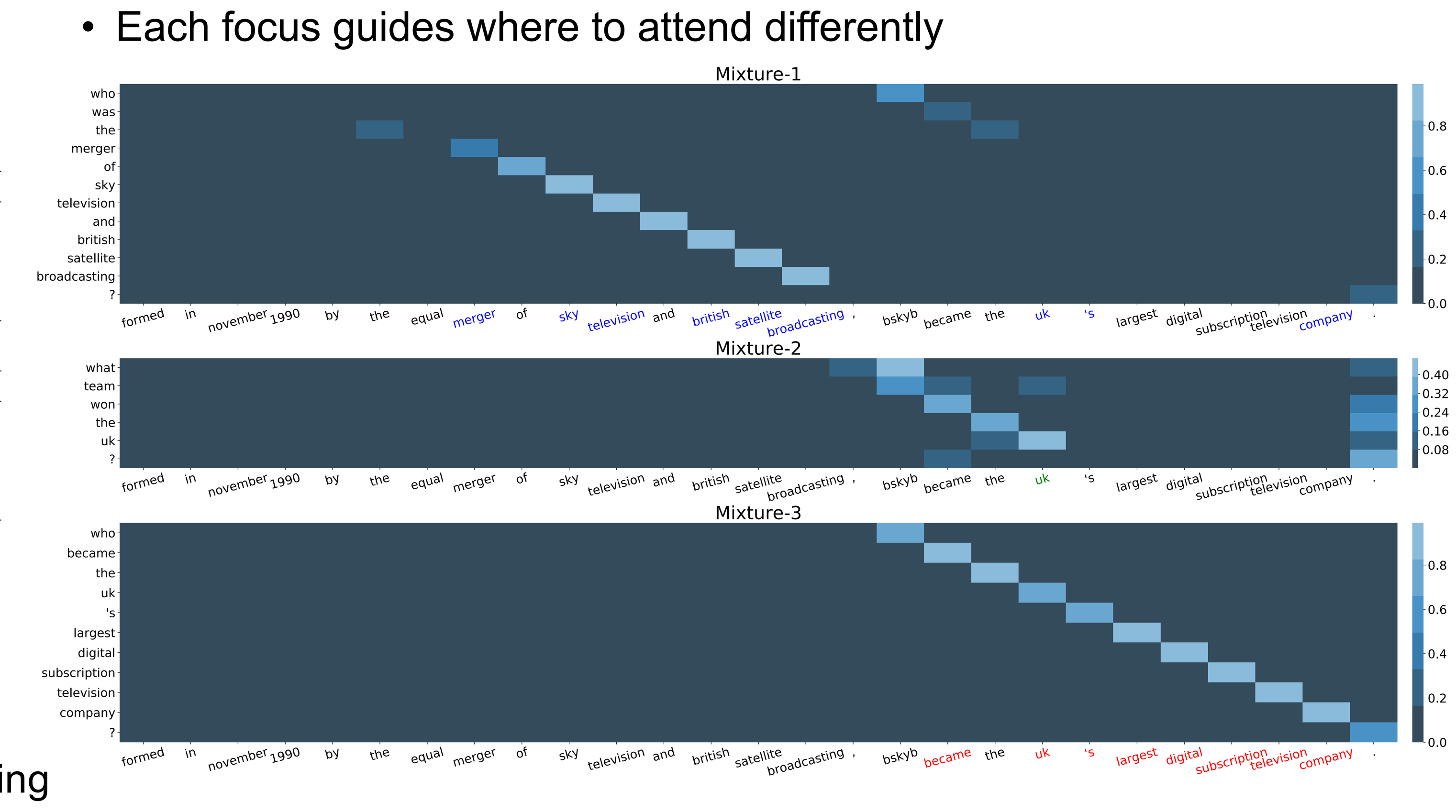
(a) SQuAD question generation

	Diversity (%)			Accuracy (%)		
Baselines	Win	Lose	Tie	Win	Lose	Tie
vs. 3-D. Beam	**50.4**	40.9	8.7	**46.2**	38.5	15.3
vs. 3-T. Sampling	**48.7**	42.0	9.3	**50.3**	41.2	8.5
vs. 3-M. Decoder	**49.7**	39.6	10.7	**46.5**	37.5	16.0

(b) CNN-DM abstractive summarization

Attention during generation

- Each focus guides where to attend differently

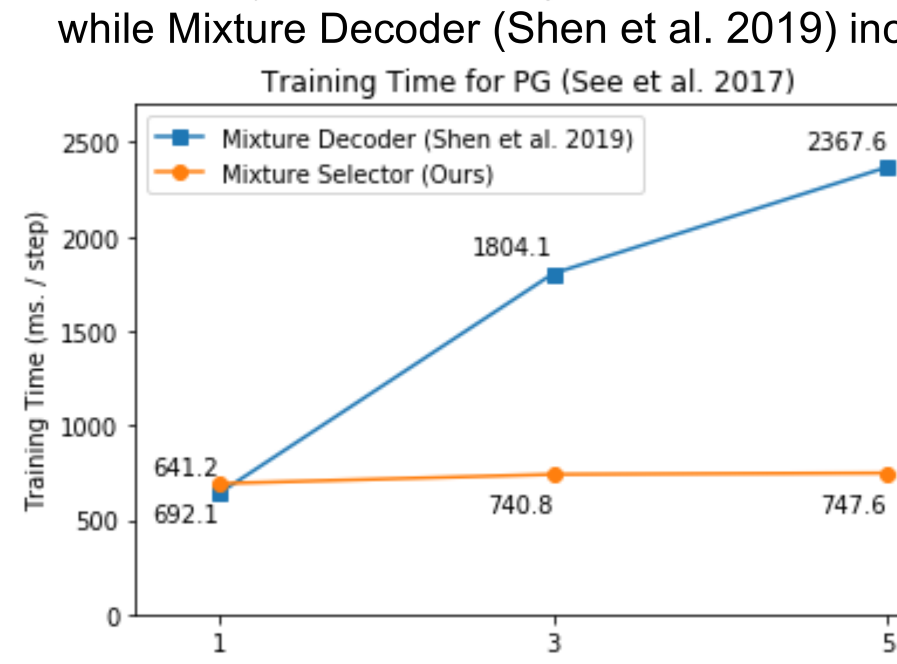


Evaluation Metrics

- Extensions of BLEU-4 / ROUGE-2
- **Top-1 Accuracy (⇑)**
  - BLUE/ROUGE of the output with the best log-likelihood
- **Oracle Accuracy (⇑)**
  - Calculate BLUE/ROUGE with K outputs, then pick the best one
  - Assumes optimal ranking method (oracle)
  - High oracle accuracy => broad coverage of target distribution
- **Pairwise Similarity (⇓)**
  - Average of pairwise BLEU/ROUGE between generated outputs
  - High similarity => mode collapse
  - Low similarity => highly diverse

No Training Overhead

- Decoder is bottleneck in Seq2Seq training
- Ours hardly affects training time,
  while Mixture Decoder (Shen et al. 2019) increases training time linearly with the number of mixtures



Seq2Seq of all trades? Master of none.
Let it simply learn one-to-one mapping!

Paper

Code (PyTorch)

- **Paper:** https://arxiv.org/abs/1909.01953
- **Code:** https://github.com/clovaai/FocusSeq2Seq
- **Twitter:** @jmin__cho