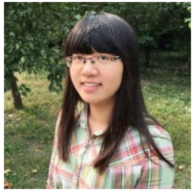
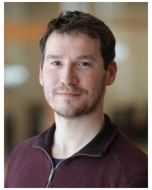
Mixture Models for Diverse Machine Translation: *Tricks of the Trade*






Tianxiao Shen*
tianxiao@mit.edu
(*: Equal contribution)

Myle Ott*

Michael Auli

Marc'Aurelio Ranzato

ICML 2019

MIT CSAIL

facebook
Artificial Intelligence Research

Translation Is One-To-Many

German	danke	sie brauchen zeit
English	thank you	you need time
	thanks	they need time
	thank you very much	it takes time

$p(y|x)$ is multi-modal, a sentence can have different translations

Translation Is One-To-Many

German	danke	sie brauchen zeit
English	thank you	you need time
	thanks	they need time
	thank you very much	it takes time

$p(y|x)$ is multi-modal, a sentence can have different translations

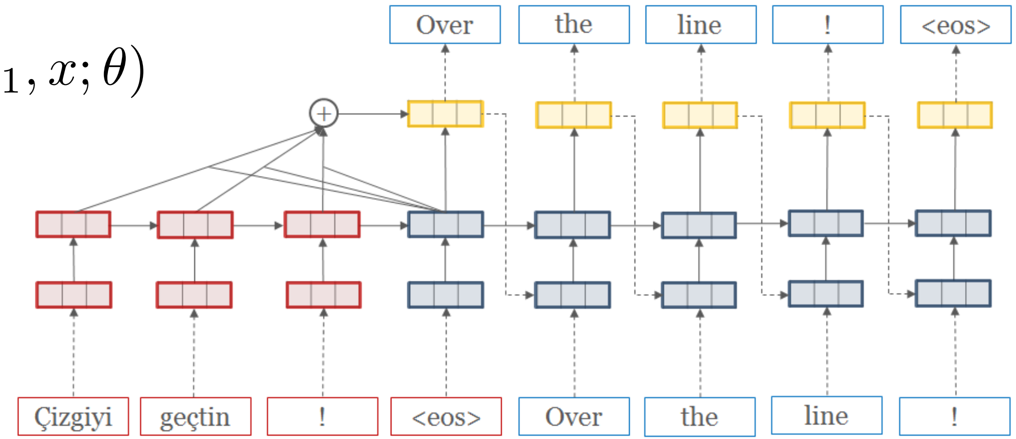
Goal: efficiently decode a diverse set of hypotheses

Neural Machine Translation

Input: source sentence $x = x_1, \cdots, x_L$

Output: target translation $y = y_1, \cdots, y_T$

$$p(y|x;\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}, x; \theta)$$



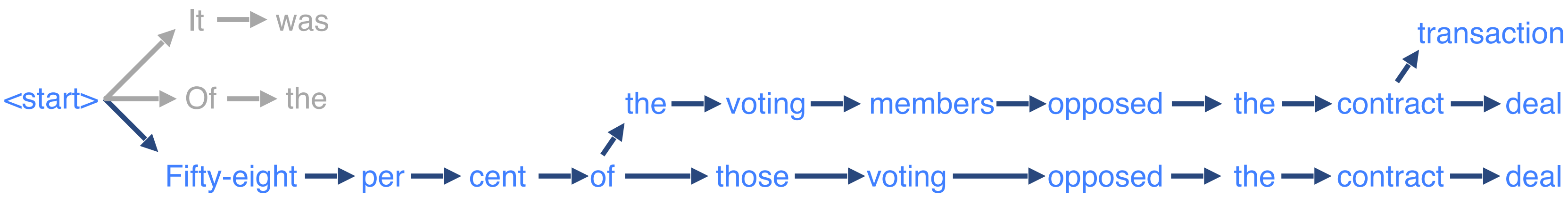
(opennmt.net)

Search for Multiple Modes Is Difficult...

$$\arg\max_{y_1,\cdots,y_T} \prod_{t=1}^{T} p(y_t|y_{1:t-1}, x; \theta)$$

Beam search can effectively find one likely y but cannot explore multiple modes

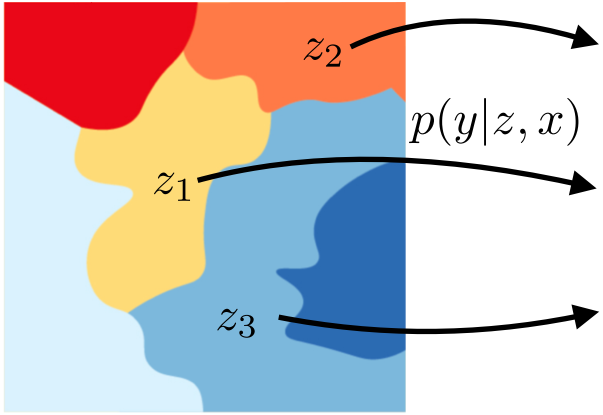
Source	参与投票的成员中，58% 反对该合同交易。
References	It was rejected by 58 % of its members who voted in the ballot .
	Of the members who voted , 58 % opposed the contract transaction .
	Of the members who participated in the vote , 58 % opposed the contract .



Explicitly Model Uncertainty

Introduce a latent variable z to capture different translation modes

Better explore the search space, decode different y from different z



It was rejected by 58 % of its members who voted in the ballot .

Of the members who voted , 58 % opposed the contract transaction .

Of the members who participated in the vote , 58 % opposed the contract .

Previous Attempt: Conditional VAE

Gaussian z, $p(y|x;\theta) = \int_z p(z|x;\theta)p(y|z,x;\theta)$

$$\log p(y|x;\theta) \geq \mathbb{E}_{q(z|x,y;\phi)}[\log p(y|z,x;\theta)] - D_{\mathrm{KL}}(q(z|x,y;\phi)\|p(z|x;\theta))$$

(Kingma & Welling, 2014; Zhang et al., 2016)

"Posterior collapse" in language modeling, the latent variable is ignored

(Bowman et al., 2016)

Our Approach: Mixture Model

Multinomial z, taking values in $\{1, \cdots, K\}$

$$p(y|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta)$$

Simplest, enumerable, exact marginal

MIT CSAIL

facebook Artificial Intelligence Research

Our Approach: Mixture Model

Multinomial z, taking values in $\{1, \cdots, K\}$

$$p(y|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta) = \frac{1}{K}\sum_{z=1}^{K} p(y|z,x;\theta)$$

Even simpler:

—set $p(z|x;\theta) = 1/K$, each component is equally likely a priori

Our Approach: Mixture Model

Multinomial z, taking values in $\{1, \cdots, K\}$

$$p(y|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta) = \frac{1}{K}\sum_{z=1}^{K} p(y|z,x;\theta) \geq \frac{1}{K}\max_{z} p(y|z,x;\theta)$$

Even simpler:

—assume $p(y|z,x;\theta)$ is large for one z, but nearly zero for others

a particular translation is only explained by a particular component

Training Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim \text{data}}\left[\min_z - \log p(y|z,x;\theta)\right]$$

$$p(y|x;\theta) = \sum_{z=1}^{K} p(z|x;\theta)p(y|z,x;\theta) = \frac{1}{K}\sum_{z=1}^{K} p(y|z,x;\theta) \geq \frac{1}{K}\max_z p(y|z,x;\theta)$$

facebook Artificial Intelligence Research

MIT CSAIL

EM Training

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\text{data}}\left[\min_z - \log p(y|z,x;\theta)\right]$$

Take a mini-batch $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

E-step (hard): estimate the responsibility of each component

$$r_z^{(i)} \leftarrow \mathbb{1}[z = \arg\max_{z'} p(y^{(i)}|z', x^{(i)};\theta)]$$

M-step: update θ through each component with gradients

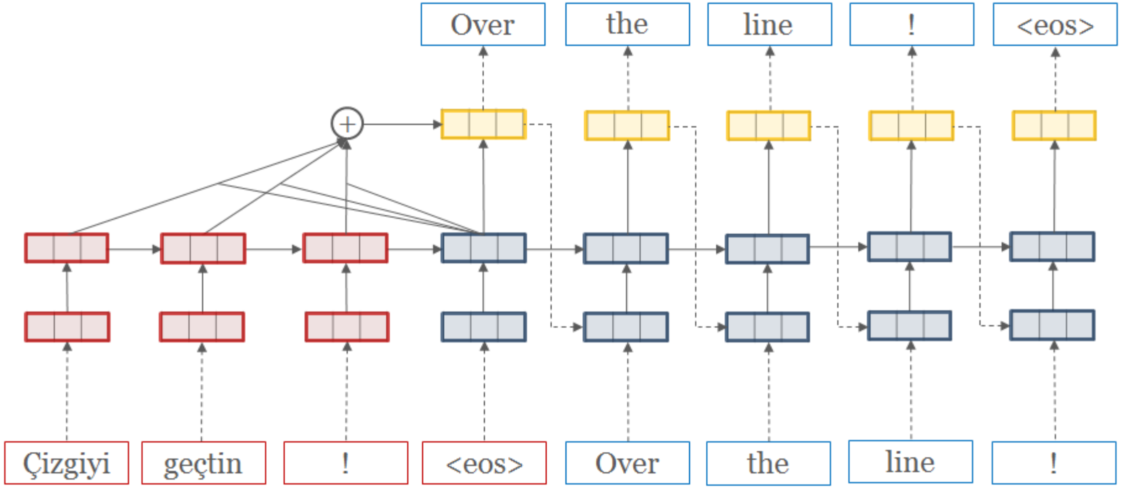
$$r_z^{(i)} \cdot \nabla_\theta \log p(y^{(i)}|z, x^{(i)};\theta)$$

Just like training mixture of Gaussians but in text space and conditioned on source

Parameterization

$$\log p(y|x;\theta)$$

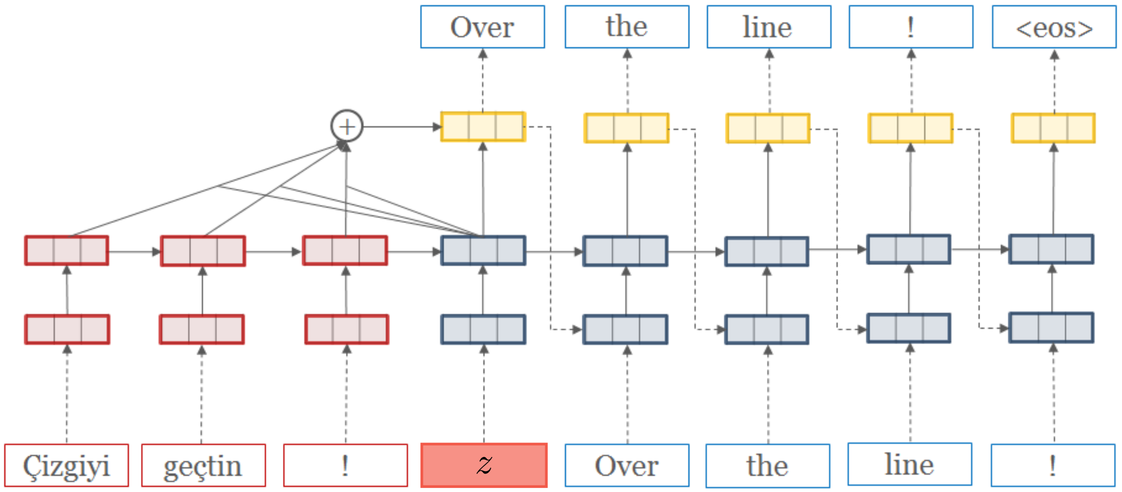
Before:



Parameterization

$$\log p(y|z, x; \theta)$$

After:



Testing

Generate K hypotheses by greedily decoding $p(y|z, x; \theta), \;\; z = 1, \cdots, K$

✓ Computationally efficient and parallelizable

Solely depend on the latent variable to produce different hypotheses

✓ No heuristic diverse decoding methods

Try It Out

Source	参与投票的成员中，58% 反对该合同交易。
Hypotheses	`Fifty-eight per cent of those voting opposed the contract deal .` $z = 1$
	`Fifty-eight per cent of those voting opposed the contract deal .` $z = 2$
	`Fifty-eight per cent of those voting opposed the contract deal .` $z = 3$

The latent variable is ignored (as in VAE) $p(y|z, x; \theta) \rightarrow p(y|x; \theta)$

Sharing too many parameters that $p(y|z, x; \theta)$ does not differentiate?

—use independently parameterized decoders

Try Again with Independent Decoders

Source	参与投票的成员中，58% 反对该合同交易。	
Hypotheses	`Fifty-eight per cent of those voting opposed the contract deal .`	$z = 1$
	.	$z = 2$
	.	$z = 3$

Only one component gets trained $p(y|z, x; \theta)$ is poor except for one z

"Rich gets richer"—once a component is better than others, it receives more gradients while others starve and eventually die (Teh, 2010)

Mixture Models Are Prone to Degeneracies

D1: all components behave the same, the latent variable is ignored

D2: only one component gets trained, other components are poor

Turns out how to train mixture models is not obvious...

Let's take a closer look

EM Training

Shared params, latent variable is ignored

E-step (hard): estimate the responsibility of each component

$$r_z^{(i)} \leftarrow \mathbb{1}[z = \arg\max_{z'} p(y^{(i)}|z', x^{(i)}; \theta)]$$

M-step: update θ through each component with gradients

$$r_z^{(i)} \cdot \nabla_\theta \log p(y^{(i)}|z, x^{(i)}; \theta)$$

Effect of Dropout

Shared params, latent variable is ignored

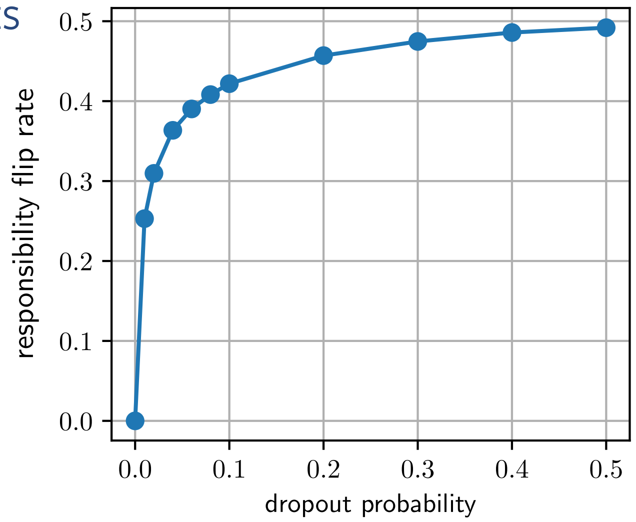
E-step (hard): estimate the responsibility of each component

$$r_z^{(i)} \leftarrow \mathbb{1}[z = \arg\max_{z'} p(y^{(i)}|z', x^{(i)}; \theta)]$$

Dropout noise here can confuse latent variable assignments

M-step: update θ through each component with gradients

$$r_z^{(i)} \cdot \nabla_\theta \log p(y^{(i)}|z, x^{(i)}; \theta)$$



Fix Dropout

Shared params, latent variable is ignored

E-step (hard): estimate the responsibility of each component

$$r_z^{(i)} \leftarrow \mathbb{1}[z = \arg\max_{z'} p(y^{(i)}|z', x^{(i)}; \theta)]$$

no dropout

M-step: update θ through each component with gradients

$$r_z^{(i)} \cdot \nabla_\theta \log p(y^{(i)}|z, x^{(i)}; \theta)$$

dropout

MIT CSAIL

facebook Artificial Intelligence Research

Try Our Modified Dropout Strategy

Source	参与投票的成员中，58% 反对该合同交易。	
Hypotheses	Fifty-eight per cent of the members who voted opposed the contract deal .	$z = 1$
	Of the members who voted , 58 % opposed the deal .	$z = 2$
	Fifty-eight per cent of the voting members opposed the contract deal .	$z = 3$

It works! :)

Design Space

- Model variants
  - hard mixture
    - uniform prior $\max_z (1/K) \cdot p(y|z, x; \theta)$
- Training schedule
  - online
- Parameterization
  - shared
- Regularization
  - no dropout at E-step, dropout at M-step

Design Space

- Model variants
  - hard mixture
    - uniform prior $\max_z (1/K) \cdot p(y|z, x; \theta)$
- Training schedule
  - online interleave E-step and M-step for each mini-batch
  - offline perform E-step for all training examples before M-step
- Parameterization
  - shared
  - independent D2: only one component gets trained
- Regularization
  - no dropout at E-step, dropout at M-step

Design Space

- Model variants
  - hard mixture
    - uniform prior $\max_z (1/K) \cdot p(y|z, x; \theta)$
    - learned prior $\max_z p(z|x; \theta) \cdot p(y|z, x; \theta)$
  - soft mixture
    - uniform prior $\sum_z (1/K) \cdot p(y|z, x; \theta)$
    - learned prior $\sum_z p(z|x; \theta) \cdot p(y|z, x; \theta)$
- Training schedule
  - online
  - offline
- Parameterization
  - shared
  - independent
- Regularization — no dropout at E-step, dropout at M-step

Metrics

BLEU (Papineni et al., 2002): modified n-gram precision metric for sentence similarity from 0 (no overlap) to 100 (same)

Metrics

- **BLEU (quality):** average BLEU of each hypothesis against the references

Source: Thanks a lot!		
Hypo1: Merci!	→	**Ref1:** Merci beaucoup!
Hypo2: Merci merci!		**Ref2:** Merci beaucoup.
Hypo3: Merci beaucoup!		**Ref3:** Merci!

Metrics

- **BLEU (quality)**: average BLEU of each hypothesis against the references
- **Pairwise-BLEU (diversity)**: average BLEU over each pair of hypotheses

Source: Thanks a lot!

Hypo1: Merci!
Hypo2: Merci merci!
Hypo3: Merci beaucoup!

Metrics

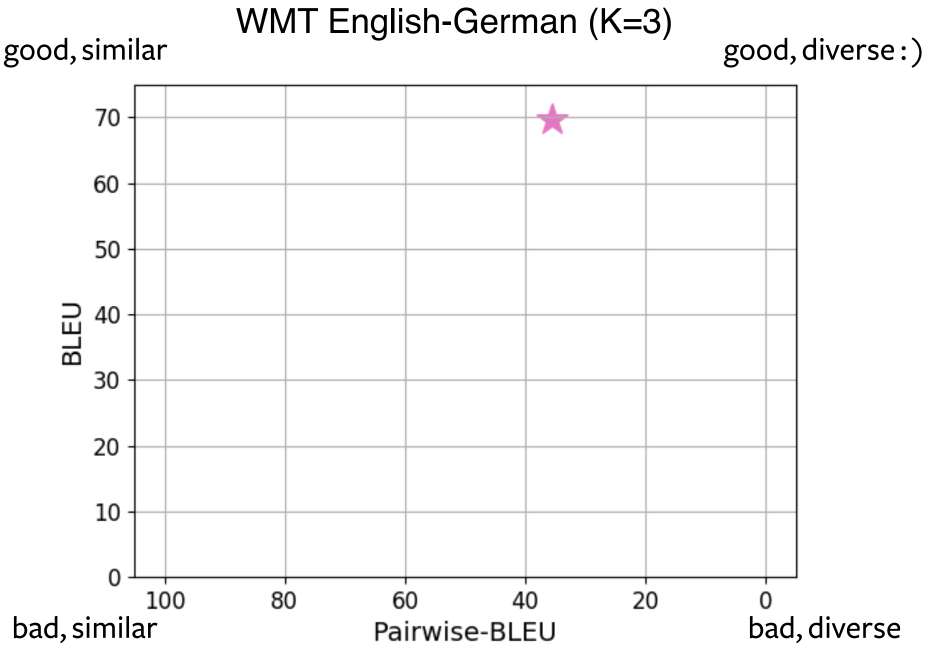
- **BLEU (quality)**: average BLEU of each hypothesis against the references
- **Pairwise-BLEU (diversity)**: average BLEU over each pair of hypotheses

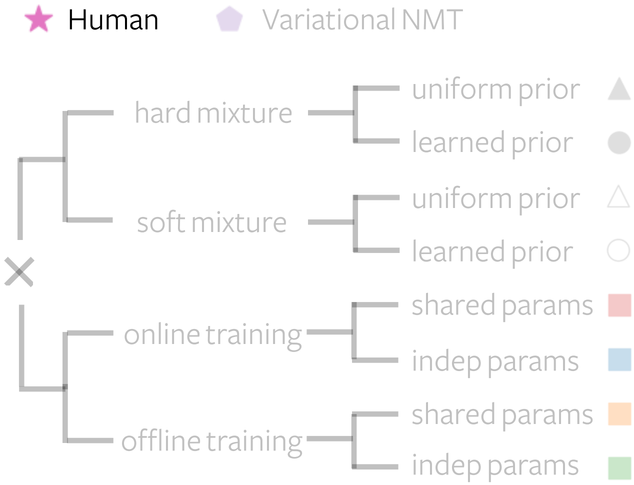
Also compute human BLEU and Pairwise-BLEU

Datasets

	#train, #ref	#test, #ref
WMT'17 English-German:	4.5M, 1	500, 10
WMT'14 English-French:	36M, 1	500, 10
WMT'17 Chinese-English:	20M, 1	2001, 3

Goal: High Quality and Diversity

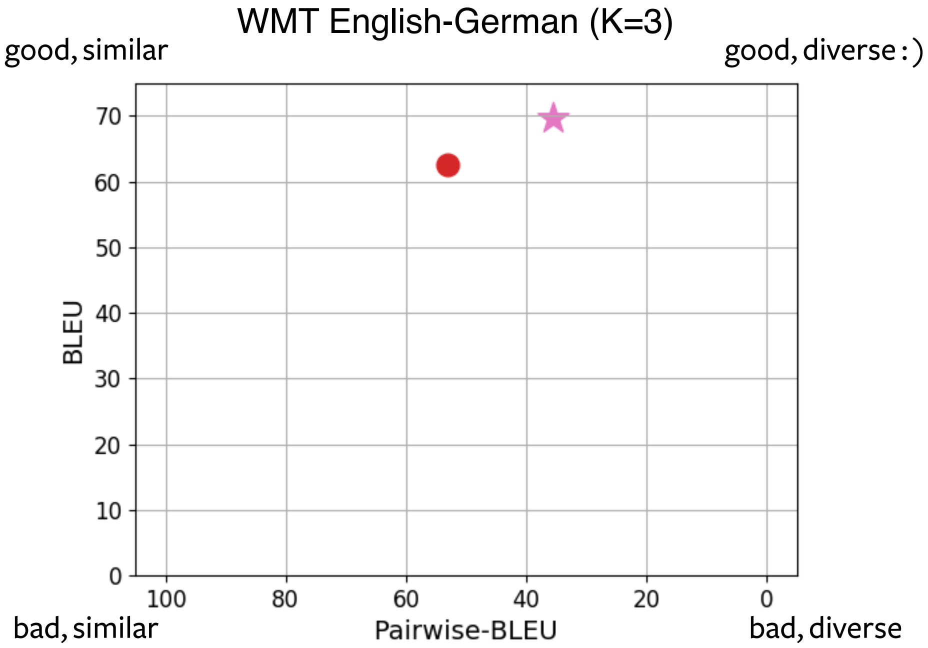


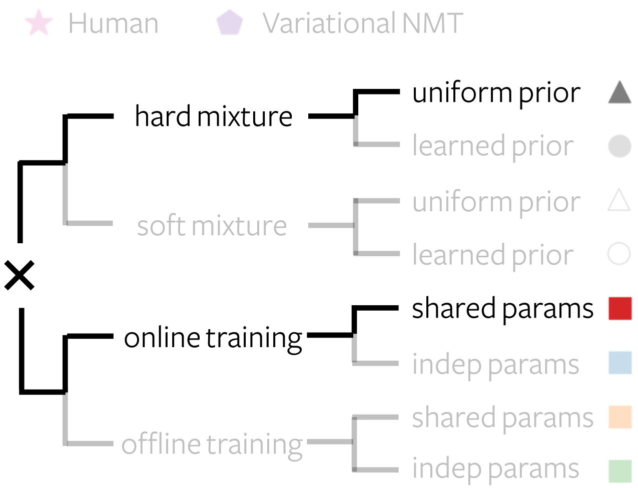


MIT CSAIL

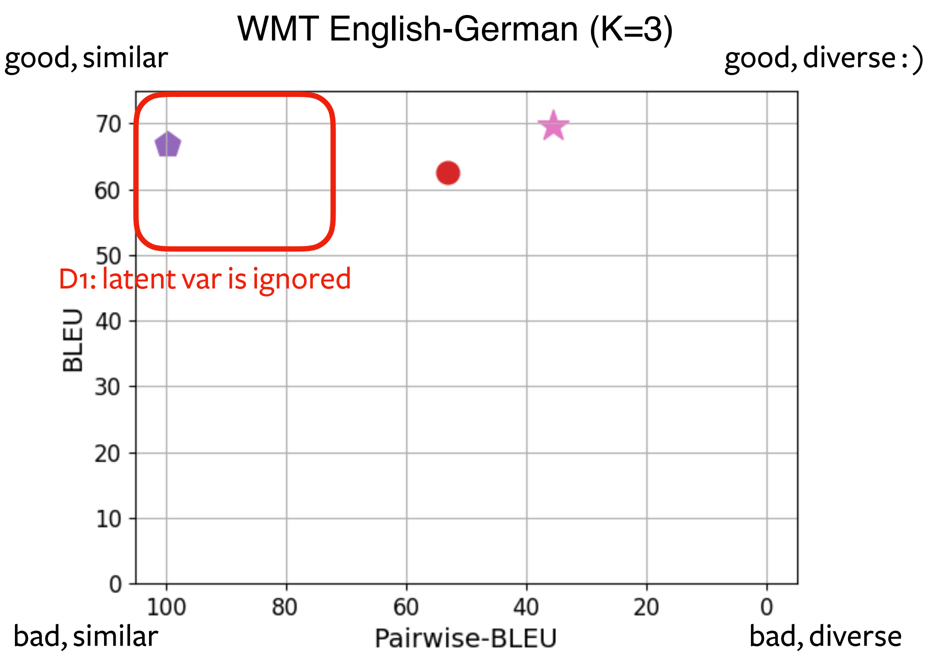
facebook
Artificial Intelligence Research

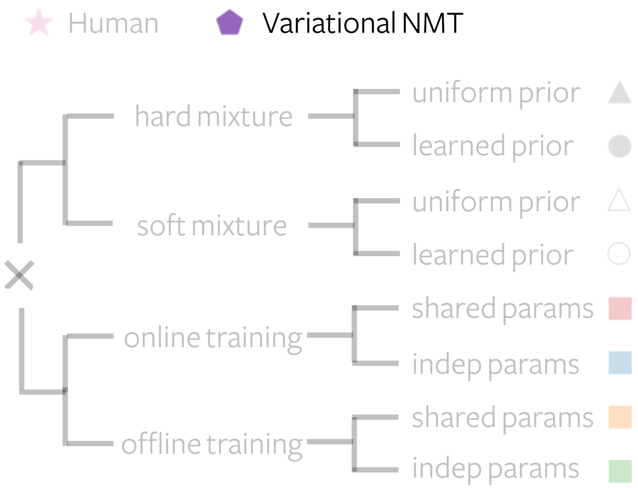
Model Exploration





Model Exploration

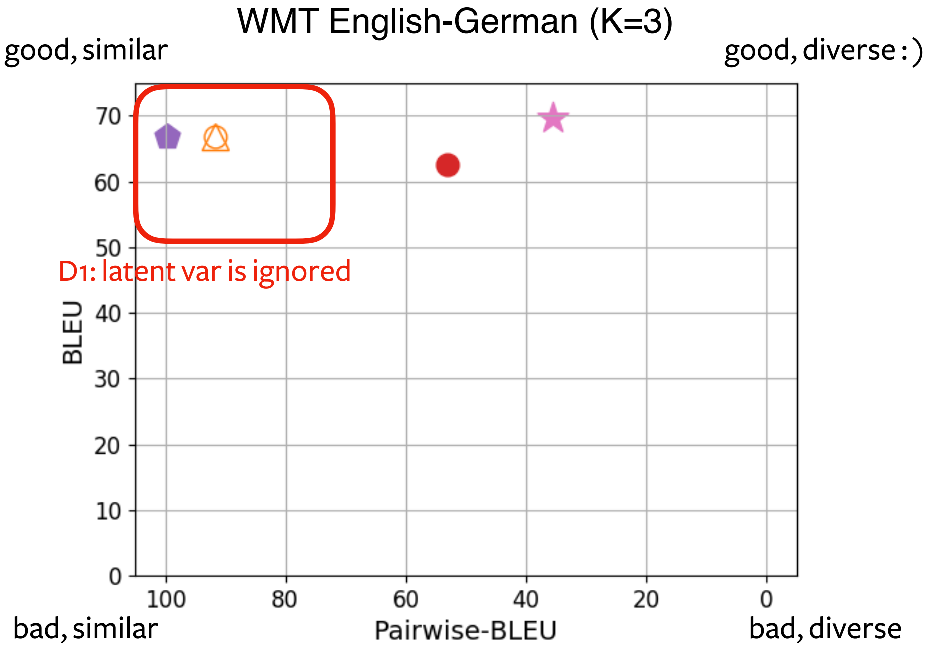


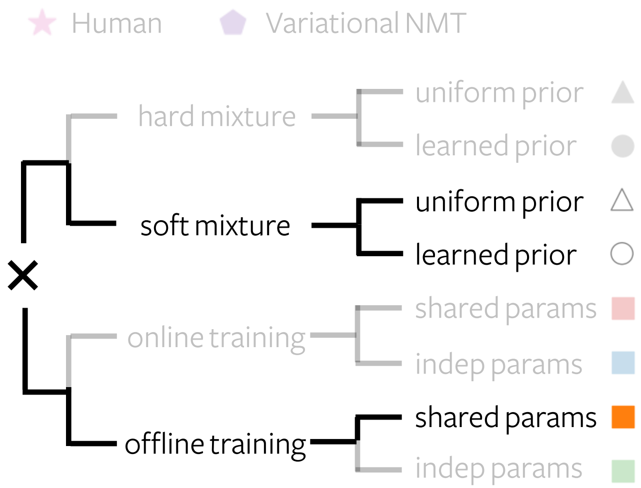


facebook
Artificial Intelligence Research

MIT CSAIL

Model Exploration

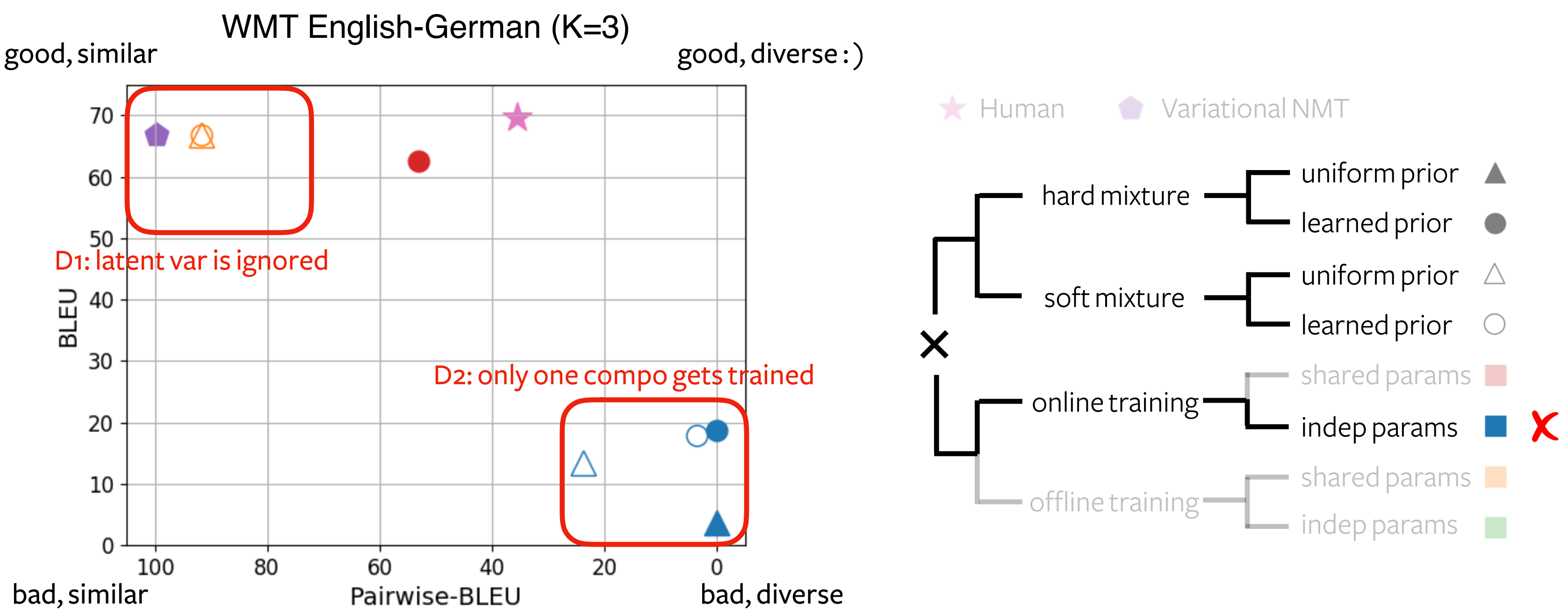




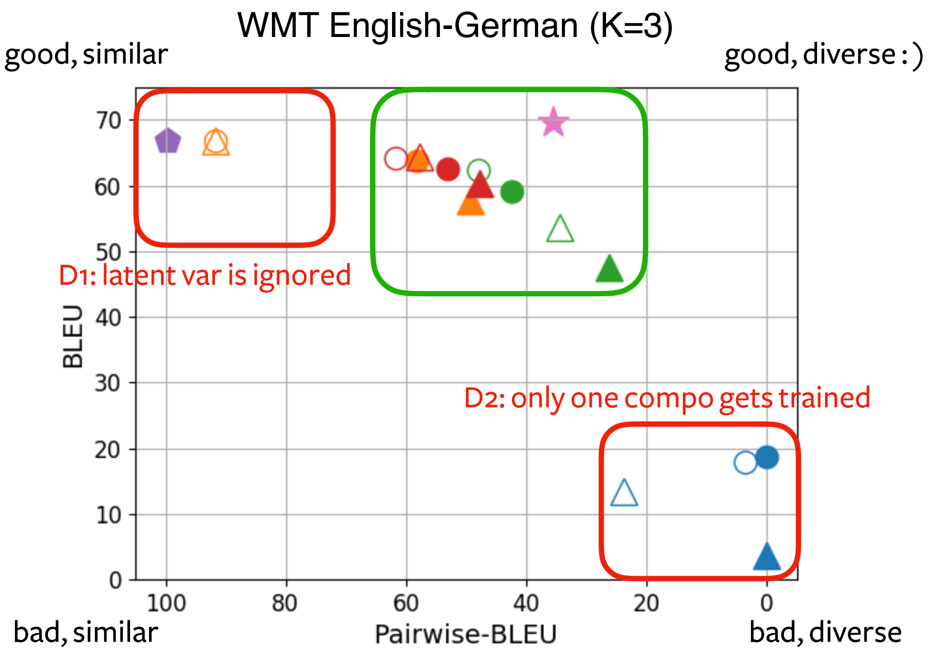
MIT CSAIL

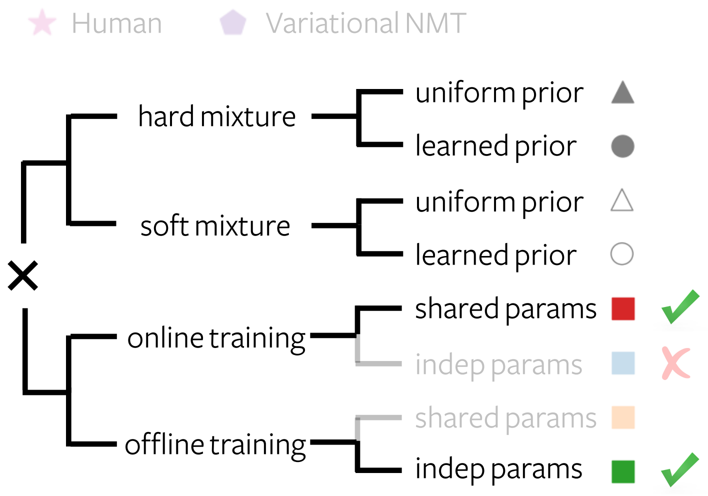
facebook
Artificial Intelligence Research

Model Exploration



Model Exploration



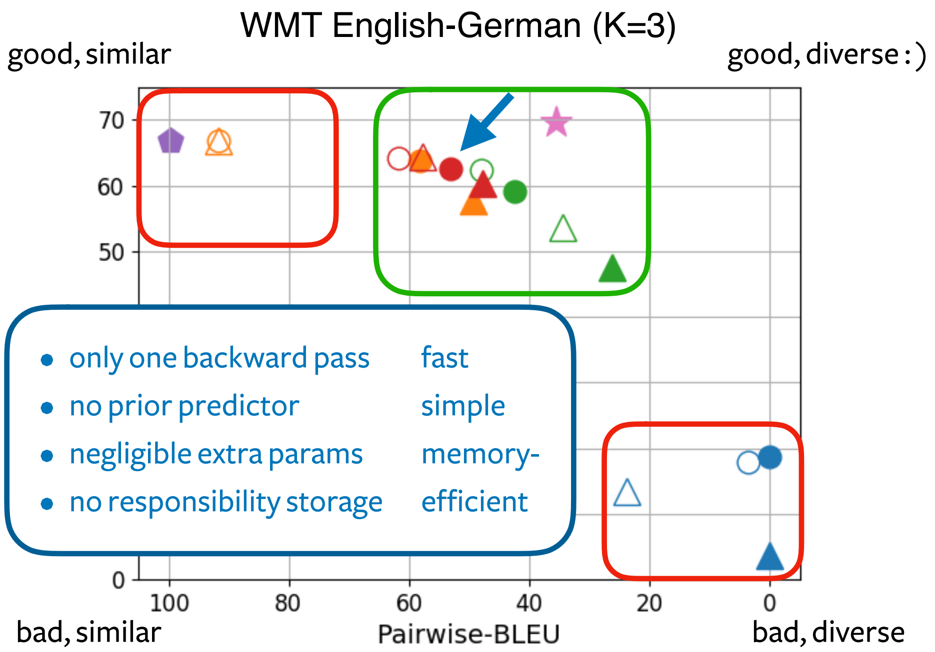


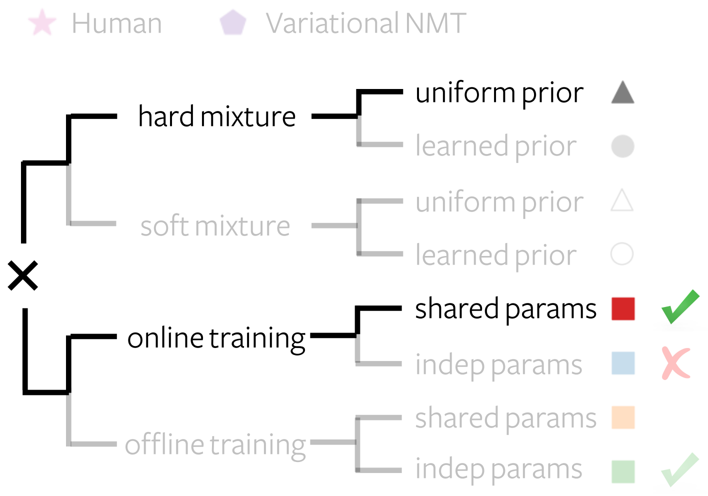
- online shared has higher quality than offline indep
- hard mixture is more diverse than soft mixture

facebook
Artificial Intelligence Research

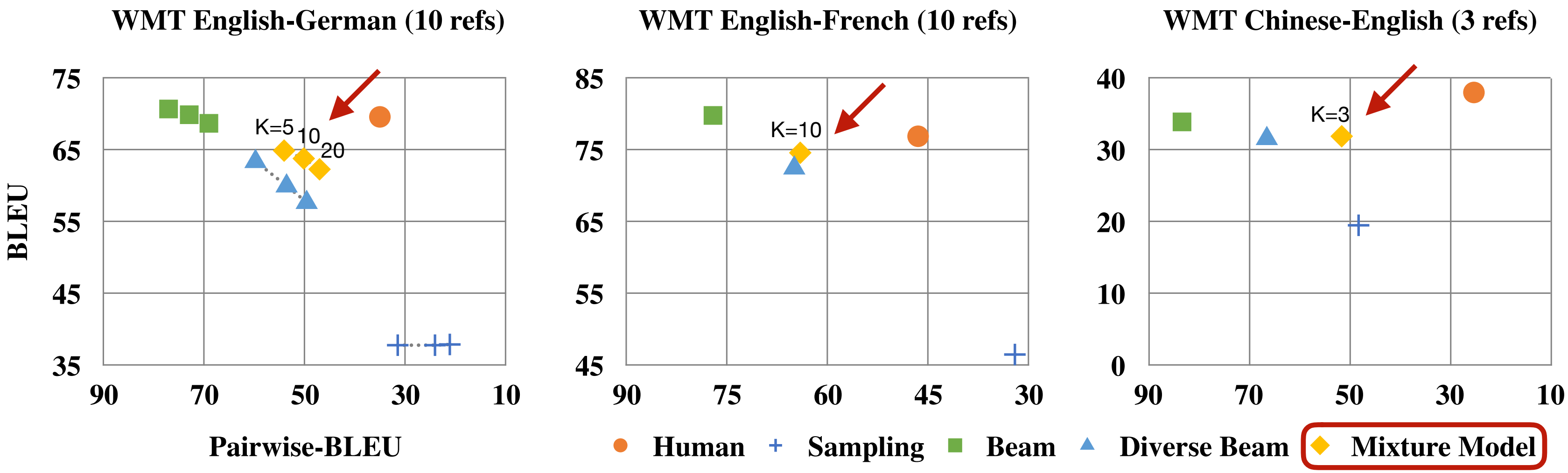
MIT CSAIL

Winning Model





Large Scale Evaluation



Latent Variable Captures Consistent Translation Styles

Source	不断 的 恐怖袭击 显然 已 对 他 造成 很大 打击 。
Reference	`Repeat terror attacks on Turkey have clearly shaken him too .`

hMup	`The continuing terrorist attacks` **`had`** `apparently hit him hard .`
	`He is clearly already being hit hard by the continuing terrorist attacks .`
	`Repeated terrorist attacks` **`have`** `apparently hit him hard .`

Source	他 从不 愿意 与 家人 争吵 。
Reference	`He never wanted to be in any kind of alte`

hMup	`He never` **`liked`** `to quarrel with his family`
	`He never wants to quarrel with his family`
	`He never` **`likes`** `to argue with his family .`

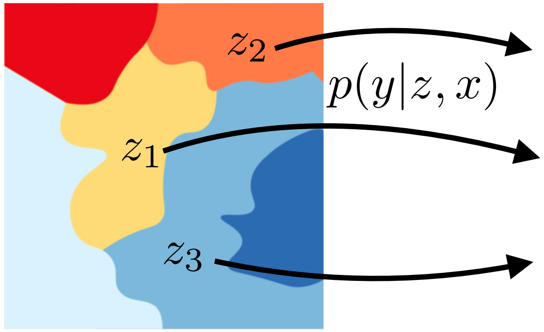
frequency of `was`, `were`, `had`:
z=1's > 3 * z=3's

frequency of `has`, `says`:
z=3's > 2 * z=1's

`this` vs. `that`, `per cent` vs. `%` ...

Conclusions

- Conditional text generation $p(y|x)$ is multi-model
- Search for multiple modes $\arg\max_{y_1,\cdots,y_T} \prod_{t=1}^{T} p(y_t|y_{1:t-1}, x; \theta)$ is difficult
- explicitly model uncertainty with latent variables



It was rejected by 58 % of its members who voted in the ballot .

Of the members who voted , 58 % opposed the contract transaction .

Of the members who participated in the vote , 58 % opposed the contract .

Conclusions

Poster #106 tonight!

- Mixture models work pretty well but hardly explored for text generation
- Training is not obvious, sub-optimal design choices can lead to degeneracies

- Model variants
  - hard Mixture
    - uniform prior
    - learned prior
  - soft Mixture
    - uniform prior
    - learned prior
- Parameterization
  - shared
  - independent
- Training schedule
  - online
  - offline
- Regularization
  - no dropout at E-step, dropout at M-step

- A strong baseline for work on latent variable text modeling
- More applications to dialogue, image captioning, summarization...
- Code: https://github.com/pytorch/fairseq

MIT CSAIL

facebook Artificial Intelligence Research