Model-Based Minimum Bayes-Risk Decoding for Text Generation

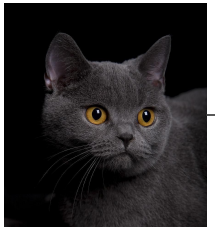
Yuu Jinnai, Tetsuro Morimura, Ukyo Honda,
Kaito Ariu, Kenshi Abe

CyberAgent AI Lab

Text generation problem

Many NLP tasks involve text generation

Image Captioning



$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

“A black cat.”

$\mathbf{x}$ $\mathbf{h}$

translation

黒猫

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

“A black cat.”

$\mathbf{x}$ $\mathbf{h}$

Beam search decoding

Beam search selects the text that maximizes the model probability (Maximum a-posteriori estimate)

$$\mathbf{h}^{\mathrm{MAP}} = \arg\max_{\mathbf{h} \in \mathcal{Y}} P_{\mathrm{model}}(\mathbf{h}|\mathbf{x})$$

- $\mathbf{x}$ Input
- $\mathcal{Y}$ All possible outputs
- $\mathbf{h}$ Candidate output

Beam search decoding

Beam search selects the text that maximizes the model probability (Maximum a-posteriori estimate)

$$\mathbf{h}^{\mathrm{MAP}} = \arg\max_{\mathbf{h} \in \mathcal{Y}} P_{\mathrm{model}}(\mathbf{h}|\mathbf{x})$$

- $\mathbf{x}$ Input
- $\mathcal{Y}$ All possible outputs
- $\mathbf{h}$ Candidate output

However, **sequences with the highest model probability is often a bad sequence** (Ott+18, Stahlbert+19)

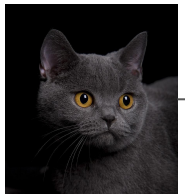
Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

The goal is to maximize the quality of the text

Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

The goal is to maximize the quality of the text

Prompt: "What's in picture?"



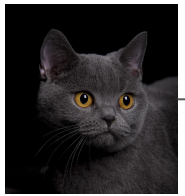
$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

- A face of a black cat
- A cat with brown eyes
- A black kitten
- A black cat
- My cute little kitty🐈
- A cat

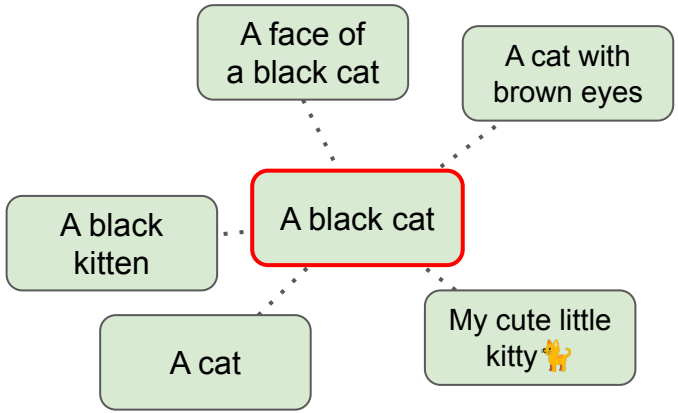
Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

The goal is to maximize the quality of the text

Prompt: "What's in picture?"



$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

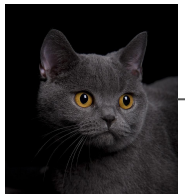


Estimate the "similarity" between the samples with $u(\mathbf{h}, \mathbf{y})$

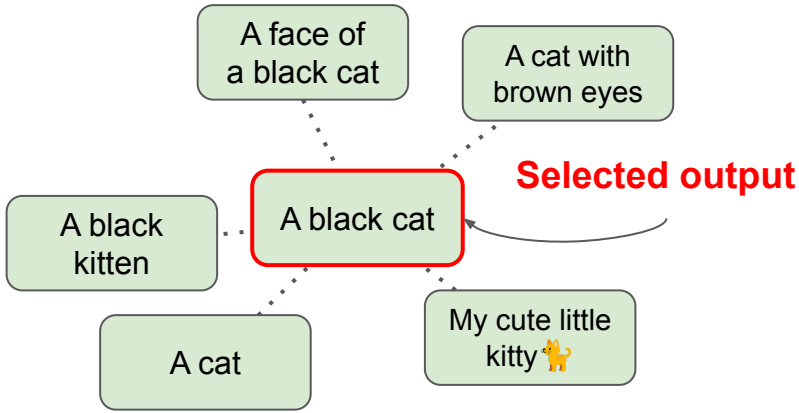
Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

The goal is to maximize the quality of the text

Prompt: "What's in picture?"



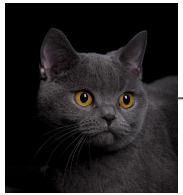
$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$



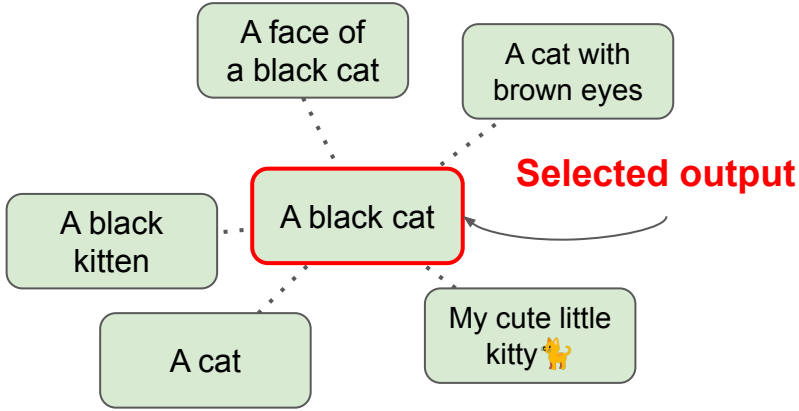
Estimate the "similarity" between the samples with $u(\mathbf{h}, \mathbf{y})$

Problem: MBR needs a lot of samples

Selecting the center point accurately requires a lot of samples



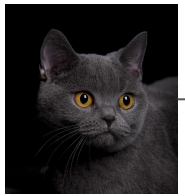
$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$



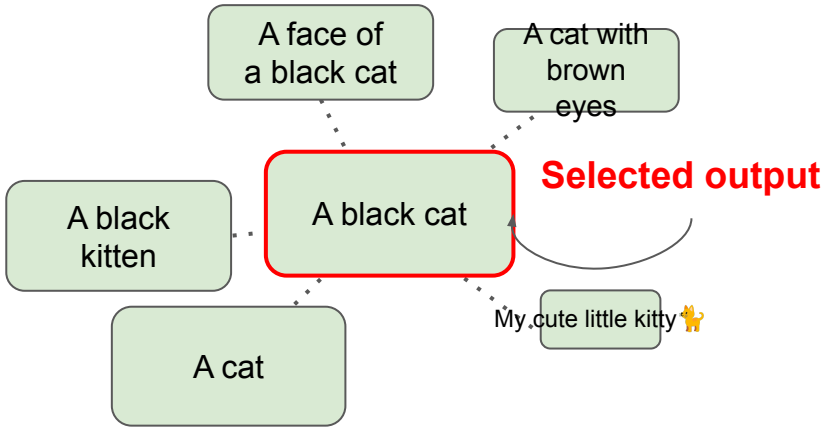
Problem: MBR needs a lot of samples

Selecting the center point accurately requires a lot of samples

→Weight the samples according to its generation probability



$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

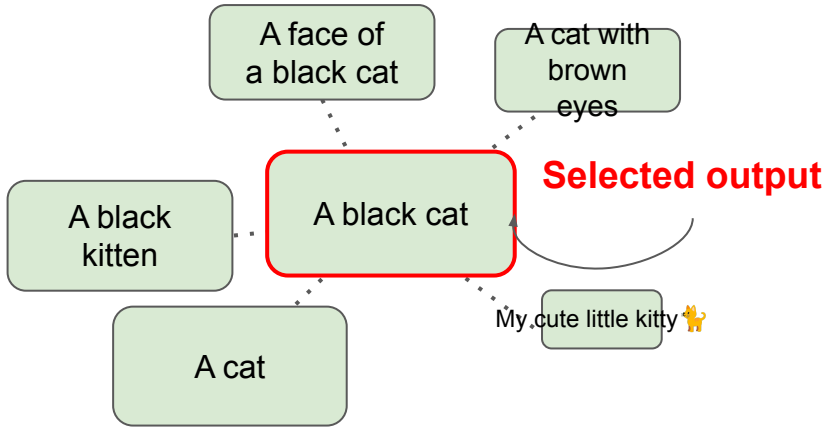


Estimate the “similarity” between the samples with $u(\mathbf{h}, \mathbf{y})$

Q. Wouldn't it increase the estimation error?

No! We can weight the samples without increasing the estimation error in expectation

1. Remove duplicated samples
2. Weight samples by $P_{\text{model}}(\mathbf{h}|\mathbf{x})$



Model-Based Minimum Bayes Risk (MBMBR) Decoding

MBR Decoding (prior work)

$$\mathbf{h}^{\mathrm{MC}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot \hat{P}_{\mathrm{model}}(\mathbf{y})$$

Monte-Carlo estimate

MBMBR Decoding (new!)

$$\mathbf{h}^{\mathrm{MB}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot P_{\mathrm{model}}(\mathbf{y})$$

Model-based estimate

Example of MBMBR

Sampled Texts		Target	Monte Carlo Estimate	Model-Based Estimate
Text	#Occurrences	P	$\hat{P}$	$\hat{P}_{\mathrm{MB}}$
But telling the truth is not a crime.	2	0.3	0.4	0.6
However, telling the truth is not a crime.	2	0.1	0.4	0.2
But to tell the truth is not a crime.	1	0.1	0.2	0.2
(All others)	0	0.5	0	0
$D_{\mathrm{KL}}(\cdot \Vert P)$		0	0.808	**0.693**

Monte Carlo estimate (prior work)

$$\mathbf{h}^{\mathrm{MC}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot \hat{P}_{\mathrm{model}}(\mathbf{y})$$

Model-based estimate (new!)

$$\mathbf{h}^{\mathrm{MB}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot P_{\mathrm{model}}(\mathbf{y})$$

Error(Model-based estimate) ≦ Error(Monte Carlo estimate)

Sampled Texts		Target	Monte Carlo Estimate	Model-Based Estimate
Text	#Occurrences	P	$\hat{P}$	$\hat{P}_{\mathrm{MB}}$
But telling the truth is not a crime.	2	0.3	0.4	0.6
However, telling the truth is not a crime.	2	0.1	0.4	0.2
But to tell the truth is not a crime.	1	0.1	0.2	0.2
(All others)	0	0.5	0	0
$D_{\mathrm{KL}}(\cdot\|\|P)$		0	0.808	**0.693**

Monte Carlo estimate (prior work)

$$\mathbf{h}^{\mathrm{MC}} = \arg\max_{\mathbf{h}\in\mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y}\in\mathcal{H}_{\mathrm{ref}}} u(\mathbf{h},\mathbf{y}) \cdot \hat{P}_{\mathrm{model}}(\mathbf{y})$$

Model-based estimate (new!)

$$\mathbf{h}^{\mathrm{MB}} = \arg\max_{\mathbf{h}\in\mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y}\in\mathcal{H}_{\mathrm{ref}}} u(\mathbf{h},\mathbf{y}) \cdot P_{\mathrm{model}}(\mathbf{y})$$

Theorem (informal)

Model-based estimate is guaranteed to be closer to the true model probability than Monte Carlo estimate measured by KL-divergence.

Error(Model-based estimate) ≦ Error(Monte Carlo estimate)

Sampled Texts		Target	Monte Carlo Estimate	Model-Based Estimate
Text	#Occurrences	P	$\hat{P}$	$\hat{P}_{\mathrm{MB}}$
But telling the truth is not a crime.	2	0.3	0.4	0.6
However, telling the truth is not a crime.	2	0.1	0.4	0.2
But to tell the truth is not a crime.	1	0.1	0.2	0.2
(All others)	0	0.5	0	0
$D_{\mathrm{KL}}(\cdot\|\|P)$		0	0.808	**0.693**

?

Monte Carlo estimate (prior work)

$$\mathbf{h}^{\mathrm{MC}} = \arg\max_{\mathbf{h}\in\mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y}\in\mathcal{H}_{\mathrm{ref}}} u(\mathbf{h},\mathbf{y}) \cdot \hat{P}_{\mathrm{model}}(\mathbf{y})$$

Model-based estimate (new!)

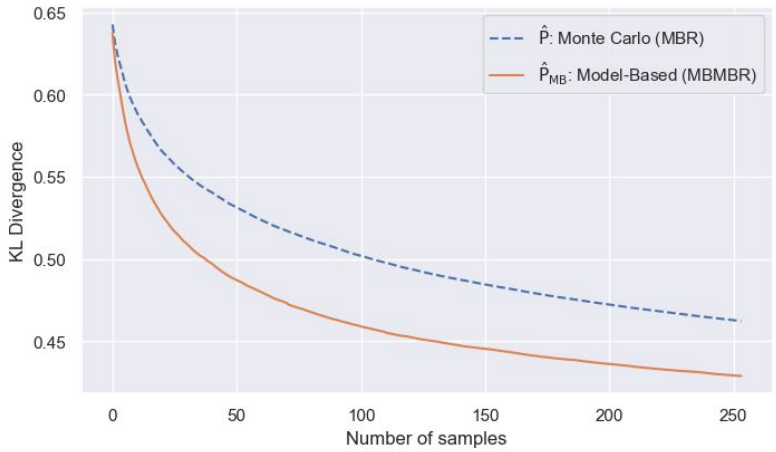
$$\mathbf{h}^{\mathrm{MB}} = \arg\max_{\mathbf{h}\in\mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y}\in\mathcal{H}_{\mathrm{ref}}} u(\mathbf{h},\mathbf{y}) \cdot P_{\mathrm{model}}(\mathbf{y})$$

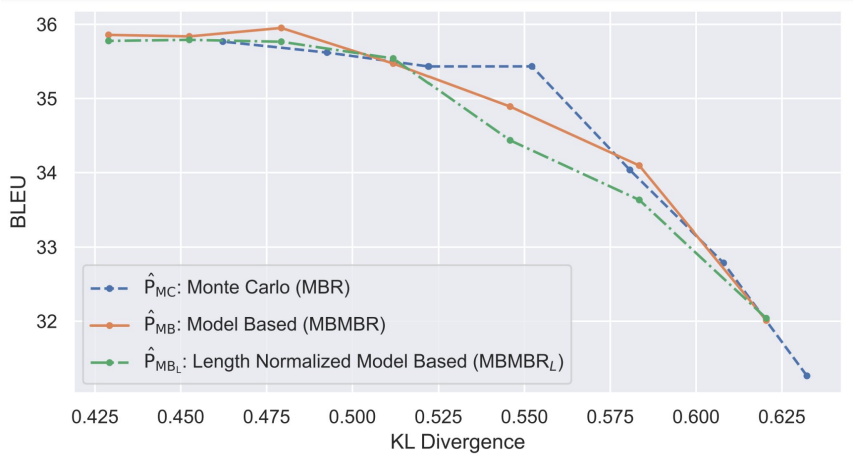
Theorem (informal)

Model-based estimate is guaranteed to be closer to the true model probability than Monte Carlo estimate measured by KL-divergence.

Accuracy of the probability estimate matters

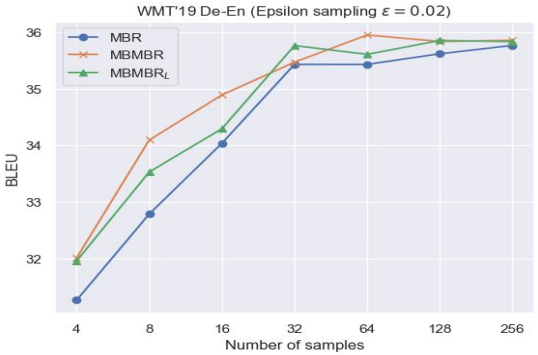
Divergence from the model probability correlates with the text quality



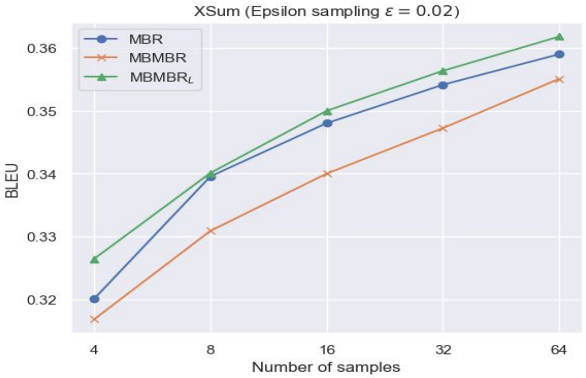


CyberAgent AI Lab

Experimental Evaluation



Machine translation (WMT19 De-En)



Text summarization (XSum)

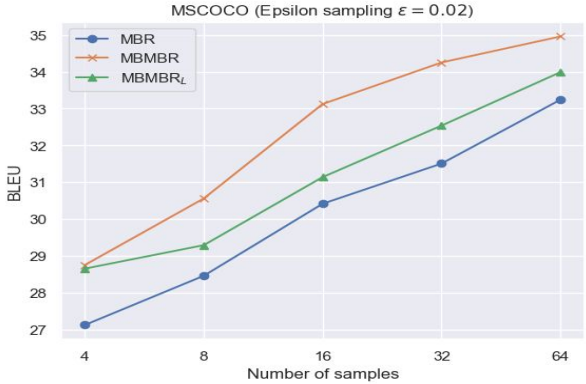


Image captioning (MS COCO)

Summary

- MBMBR uses the model probability instead of Monte Carlo estimate
- MBMBR improves the estimation of the model probability which leads to improved text quality
- Experiments show that MBMBR is effective in machine translation, text summarization, and image captioning
- Implemented in mbrs (Deguchi, 2024) `(pip install mbrs)`

Monte Carlo estimate (prior work)

$$\mathbf{h}^{\mathrm{MC}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot \hat{P}_{\mathrm{model}}(\mathbf{y})$$

Model-based estimate (new!)

$$\mathbf{h}^{\mathrm{MB}} = \arg\max_{\mathbf{h} \in \mathcal{H}_{\mathrm{cand}}} \sum_{\mathbf{y} \in \mathcal{H}_{\mathrm{ref}}} u(\mathbf{h}, \mathbf{y}) \cdot P_{\mathrm{model}}(\mathbf{y})$$