Introduction to Minimum Bayes Risk Decoding

Yuu Jinnai
CyberAgent

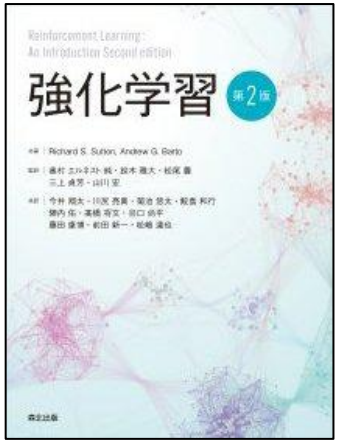
CyberAgent AI Lab

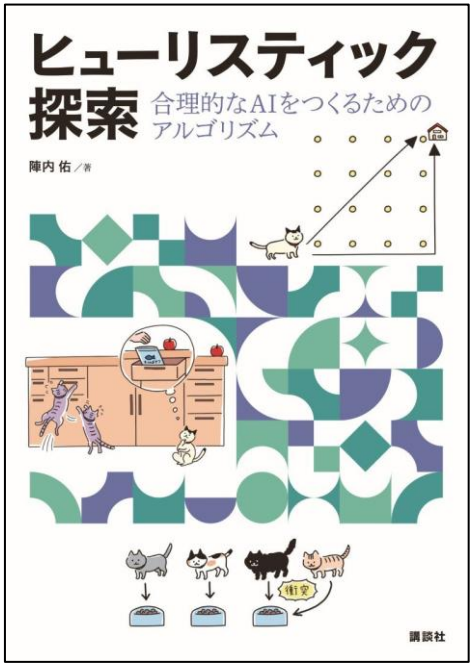
Yuu Jinnai 陣内 佑

CyberAgent AI Lab, Reinforcement Learning Team

Research Interest

- Sequential Decision Making
  - Planning and Search (Text Generation)
  - Reinforcement Learning (RLHF)



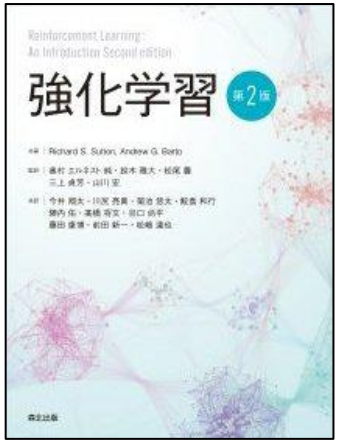


Yuu Jinnai 陣内 佑

CyberAgent AI Lab, Reinforcement Learning Team

Research Interest

- Sequential Decision Making
  - Planning and Search (Text Generation)
  - Reinforcement Learning (RLHF)

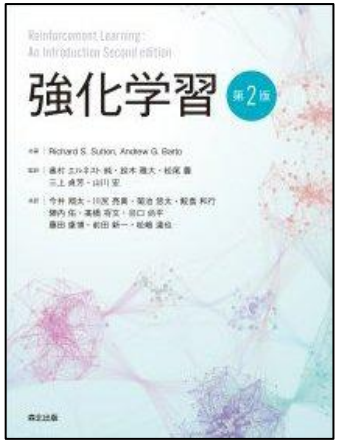


Yuu Jinnai 陣内 佑

CyberAgent AI Lab, Reinforcement Learning Team

Research Interest

- Sequential Decision Making
  - Planning and Search (Text Generation)
  - Reinforcement Learning (RLHF)



Decoding Matters More than You Think!

Q1. Question Time!

What is He?

A. Dog
B. Cat
C. Owl
D. Bear

If you know him pretend that you don't

Problem: Text Generation

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.”

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

Many NLP Tasks Involve Text Generation

Image Captioning

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

“A black cat.”

output context

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

translation

黒猫

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

“A black cat.”

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

Many NLP Tasks Involve Text Generation

Image Captioning

$\mathbf{x}$

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

output = sequence of tokens

context

“A black cat.”

$\mathbf{h}$

translation

黒猫

$\mathbf{x}$

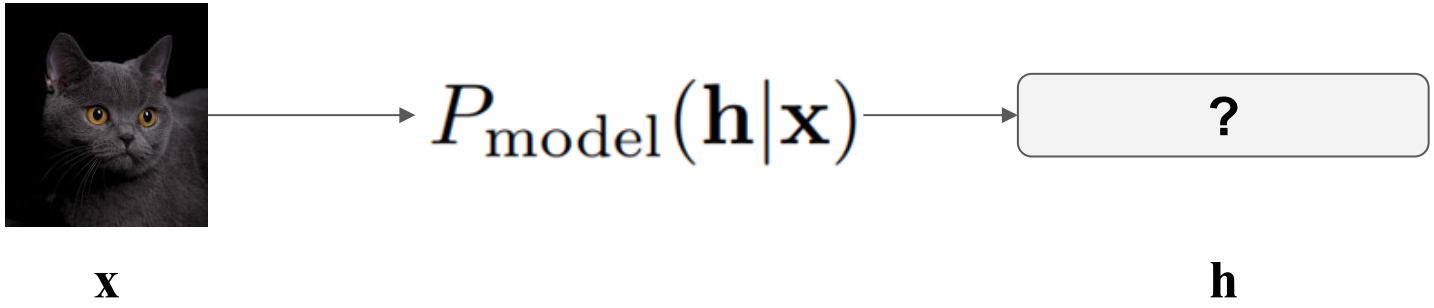
$P_{\text{model}}(\mathbf{h}|\mathbf{x})$

“A black cat.”

$\mathbf{h}$

Text Generation Problem

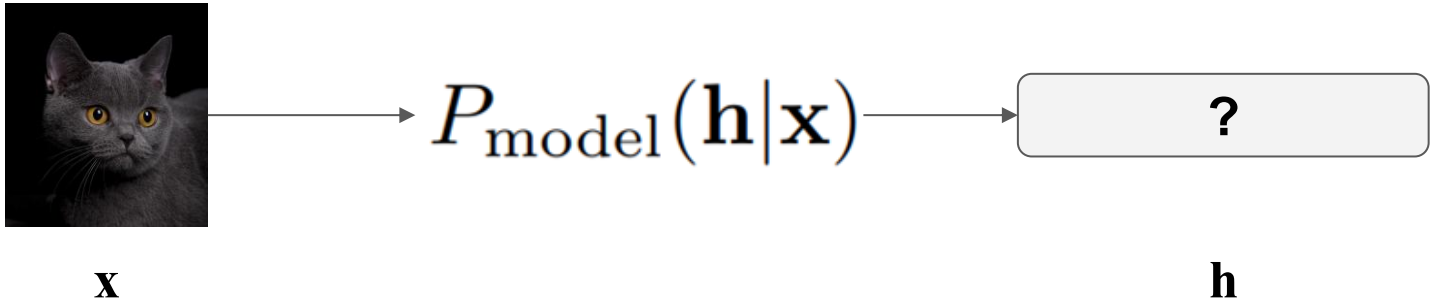
Given a context $\mathbf{x}$ and a model P_{model}, generate a *desired* output



Text Generation Problem

Given a context $\mathbf{x}$ and a model P_{model},
generate a *desired* output

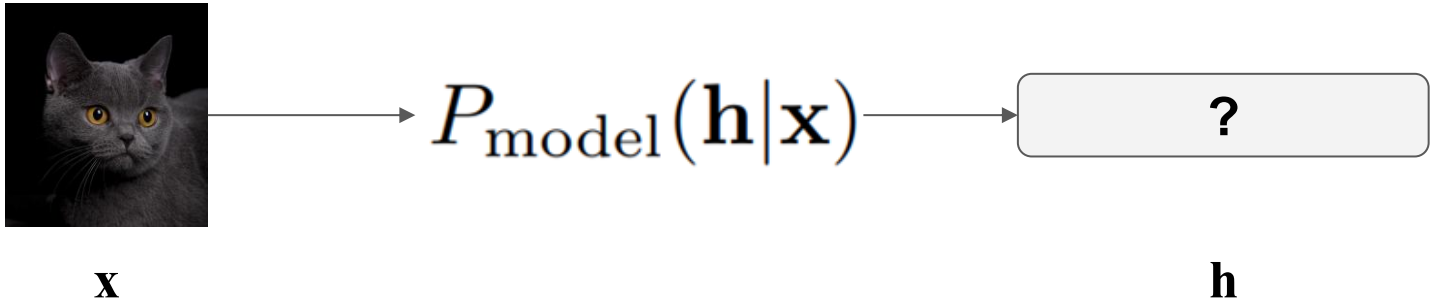
→ This process is called **decoding!**



Text Generation Problem

Given a context $\mathbf{x}$ and a model P_{model},
generate a *desired* output

→ This process is called **decoding!**
…but what is ***desired* output**?



Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.
B. Mostly yes - rare edge cases may exist.
C. No - there are many other aspects to consider.
D. It can never be perfect so the question has no point.

Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.

B. Mostly yes - rare edge cases may exist.

C. No - there are many other aspects to consider.

D. It can never be perfect so the question has no point.

Hint is...

Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.

B. Mostly yes - rare edge cases may exist.

C. No - there are many other aspects to consider.

D. It can never be perfect so the question has no point.

Hint is...

Okay, it's time to think about Him!

What is He?

A. Dog
B. Cat
C. Owl
D. Bear

Maximum-a-posteriori (MAP) Decoding

What is He?

A. Dog
B. Cat
C. Owl
D. Bear

MAP decoding (estimate) selects the most probable option (i.e. highest probability)

Why MAP Wouldn't be Perfect?

Why MAP Wouldn't be Perfect?

Why MAP Wouldn't be Perfect?

What is He?

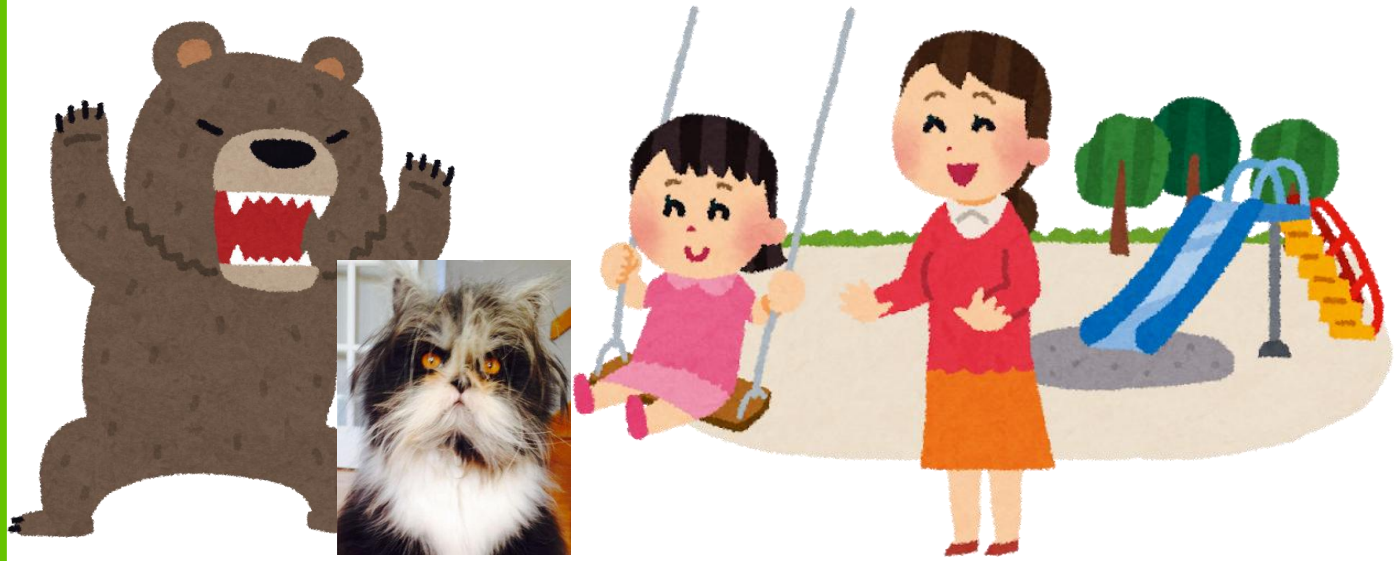
A. Dog
B. Cat
C. Owl
D. Bear

Why MAP Wouldn't be Perfect?

If it's actually a cat/dog/owl, it may not be a big problem, but...

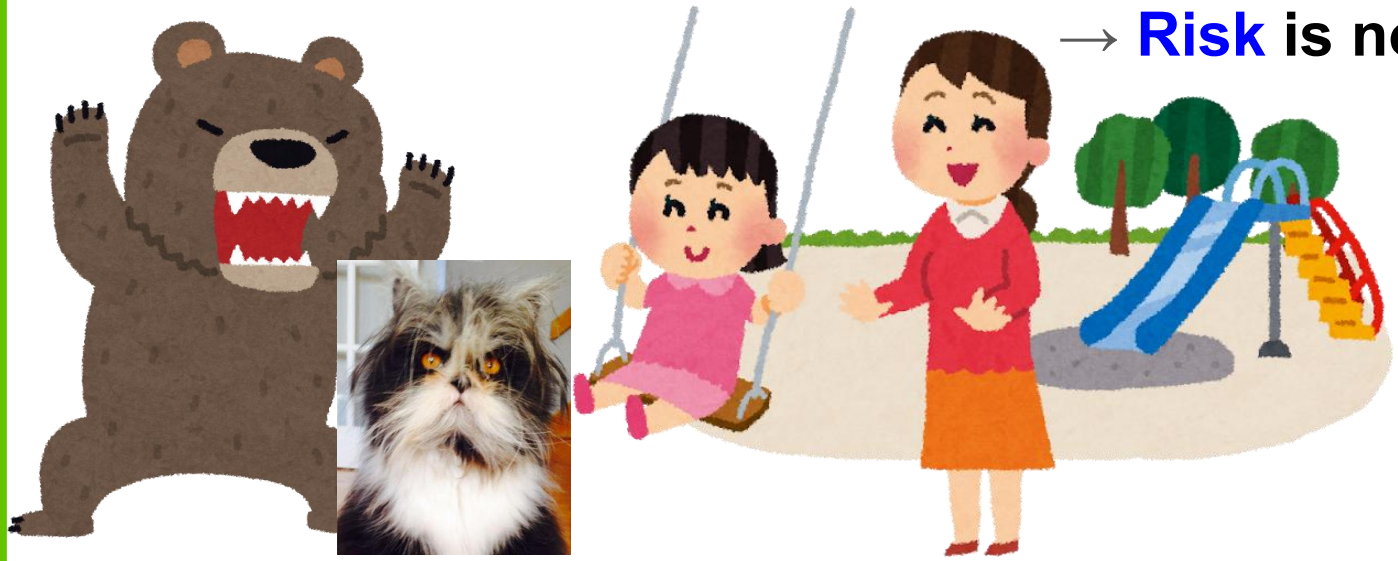
Why MAP Wouldn't be Perfect?

If it's actually a cat/dog/owl, it may not be a big problem, but...
What if he's a baby bear with an angry mom behind him?



Why MAP Wouldn't be Perfect?

If it's actually a cat/dog/owl, it may not be a big problem, but...
What if he's a baby bear with an angry mom behind him?



→ **Risk** is not considered

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. Pet it!

- It is most likely a cat

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. Pet it!

- It is most likely a cat

B. Pull back the kid and take a closer look

- Most likely a cat but it might be a bear

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. Pet it!

- It is most likely a cat

B. Pull back the kid and take a closer look

- Most likely a cat but it might be a bear

C. Ruuuuun!!!

- Worst case, it might be a bear!

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. **Pet it!**
   - It is **most likely** a cat

B. **Pull back the kid and take a closer look**
   - **Most likely a cat but it might be** a bear

C. **Ruuuuun!!!**
   - **Worst case**, it might be a bear!

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. Pet it!

- He is most likely a cat

→ **Maximum-a-posteriori (MAP)**

B. Pull back the kid and take a closer look

- Most likely a cat but it might be some wild animal

→ **Bayes risk minimization (MBR)**

C. Ruuuuun!!!

- There is a chance he is a bear

→ **Minimax**

Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.
B. Mostly yes - rare edge cases may exist.
C. No - there are many other aspects to consider.
D. It can never be perfect so the question has no point.

Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.
B. Mostly yes - rare edge cases may exist.
✓ C. No - there are many other aspects to consider.
D. It can never be perfect so the question has no point.

How you ACT given the probability estimate is up to you

Q2. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

✓ A. Yes - text generation is trivial with a perfect model.

B. Mostly yes - rare edge cases may exist.

C. No - there are many other aspects to consider.

D. It can never be perfect so the question has no point.

But can't we think of a "language model" that takes into account of human preference?

“Language Model” with Human Preference

But can’t we think of a ”language model” that takes into account of human preference? Yes. Reinforcement learning.

Q1. Question Time!

He is Atchoum

What is He?

A. Dog

B. Cat 🎉 🎉

C. Owl

D. Bear

He is Atchoum

What is He?

A. Dog

B. Cat 🎉 🎉

C. Owl

D. Bear

Image from
https://www.atchoumthecat.com/

He is Atchoum

What is He?

A. Dog

B. Cat 🎉 🎉

C. Owl

D. Bear

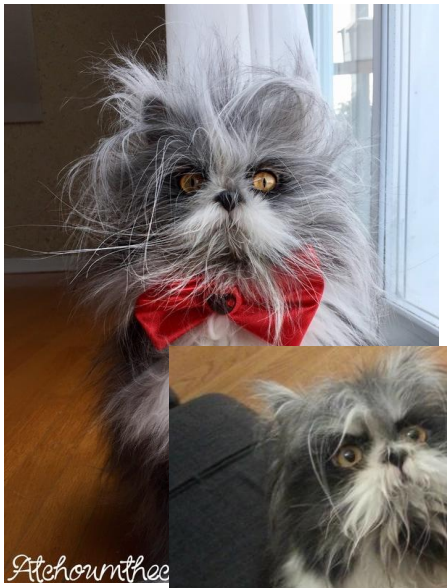


Image from
https://www.atchoumthecat.com/

You Have Multiple Options even with a Perfect Probability Model

Question should be **what you would do**

A. Pet it!

- He is most likely a cat

→ **Maximum-a-posteriori (MAP)**

B. Pull back the kid and take a closer look

- Most likely a cat but it might be some wild animal

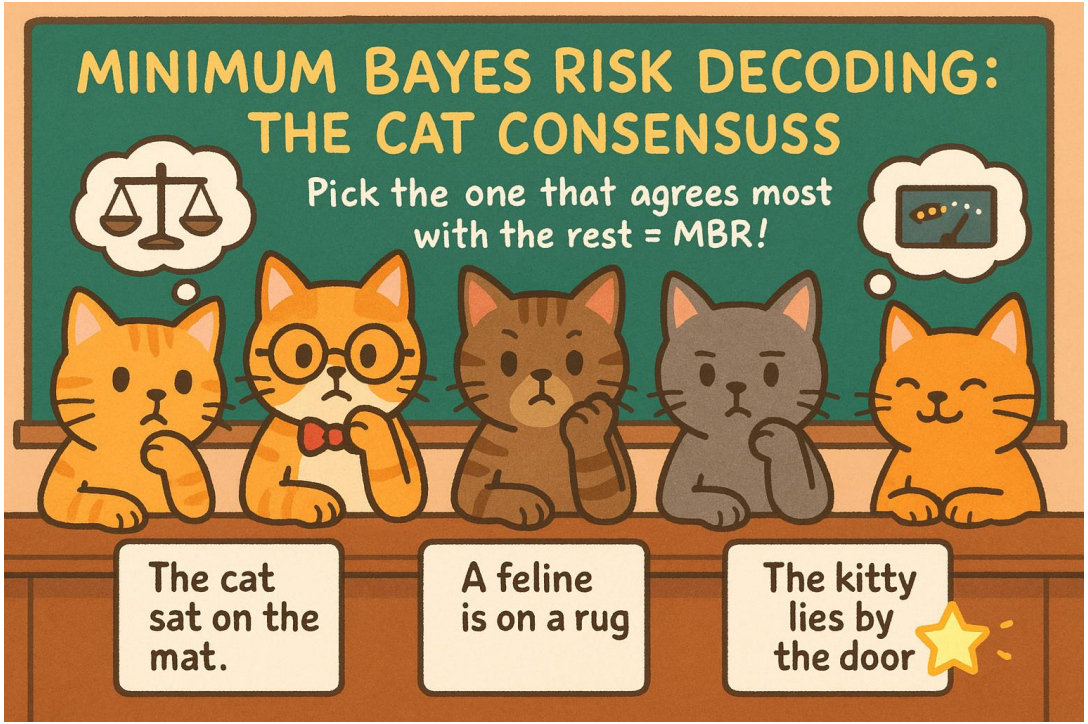
→ **Bayes risk minimization (MBR)**

C. Ruuuuun!!!

- There is a chance he is a bear

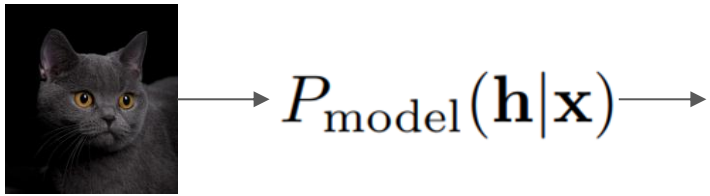
→ **Minimax**

Algorithm: Minimum Baye Risk Decoding



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

1. Sample outputs randomly



A face of a black cat

A cat with brown eyes

A black kitten

A black cat

A cat

My cute little kitty

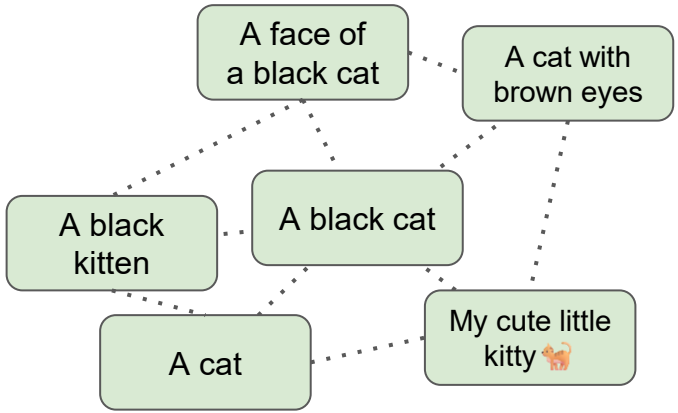
Prompt: "What's in the picture?"

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

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$

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

Prompt: "What's in the picture?"

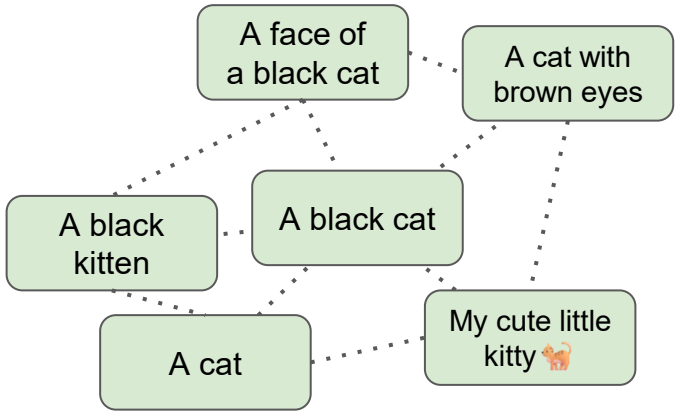


Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
   = - risk

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

Prompt: "What's in the picture?"

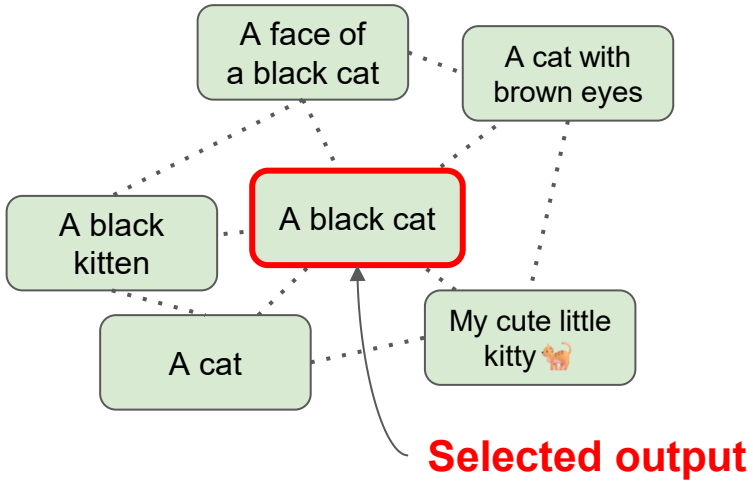


Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

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

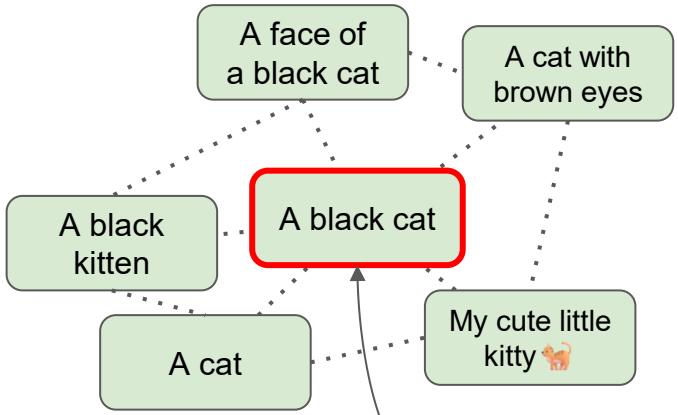
Prompt: "What's in the picture?"



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

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

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



Prompt: "What's in the picture?"

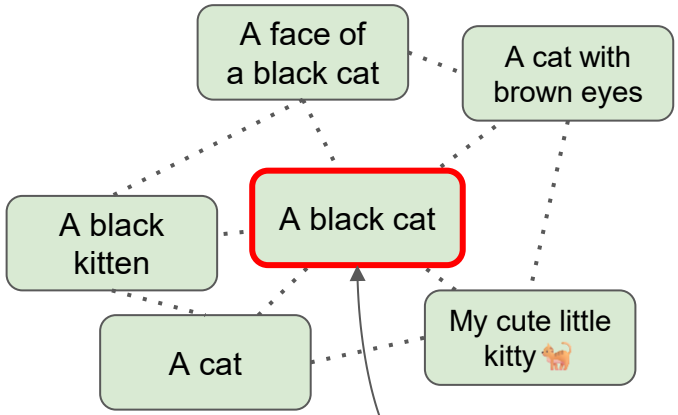
$$\mathrm{h}_{\text{MBR}} = \underset{h \in \text{samples}}{\text{argmax}} \frac{1}{|\text{samples}|} \sum_{\mathrm{y} \in \text{samples}} u(\mathrm{h}, \mathrm{y})$$

Interpretation of MBR Decoding

Assuming the generated samples are the possible “true answers”,
minimize the average risk over them

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

Prompt: “What’s in the picture?”



$$\mathrm{h}_{\text{MBR}} = \underset{h \in \text{samples}}{\text{argmax}} \frac{1}{|\text{samples}|} \sum_{\mathrm{y} \in \text{samples}} u(\mathrm{h}, \mathrm{y})$$

Utility Function Matters!

Using better utility function results in better generation

Method		Automatic Evaluation						Model	Human Eval
		BLEU	sBLEU	CHRF	YISI	BL.1	BL.2	logP	MQM ↓
Human Transl.	Ref-D	31.5	31.6	60.9	84.7	37.1	75.6	−38.0	**0.388**†
Beam 4		34.3	34.2	62.5	85.3	26.8	71.6	−11.5	2.030
MBR	sBLEU	34.7	**34.8**	62.5	85.4	23.4	70.5	−11.2	1.855
	CHRF	34.2	34.3	**64.1**	85.7	25.8	71.4	−13.2	2.139
	YISI	34.2	34.2	62.8	**86.0**	26.4	71.6	−11.4	2.445
	BLEURT v0.1	29.2	29.4	60.0	84.3	**50.0**	77.1	−18.7	**1.571**†
	BLEURT v0.2	25.4	26.0	57.7	83.1	43.9	**79.0**	−24.4	**1.661**†

Freitag et al., TACL 2022

Sampling Algorithm Matters!

Candidate	de-en	en-de	ru-en	en-ru
Ancestral	85.82	86.32	82.11	86.13
Beam	88.47	89.32	84.16	89.44
Epsilon ($\epsilon = 0.02$)	**88.51**	**89.47**	**84.36**	**90.17**
Nucleus ($p = 0.6$)	88.01	89.12	83.76	89.96
Nucleus ($p = 0.9$)	88.02	89.04	83.98	89.57

	Pseudo-Reference	de-en	en-de	ru-en	en-ru
Epsilon ($\epsilon = 0.02$)	Ancestral	85.82	87.51	82.02	88.41
	Beam	85.62	87.40	81.64	87.78
	Epsilon ($\epsilon = 0.02$)	85.89	87.74	82.01	88.46
	Epsilon ($\epsilon = 0.02$)*	85.87	87.74	81.98	88.46
	Nucleus ($p = 0.6$)	85.69	87.57	81.76	88.26
	Nucleus ($p = 0.9$)	**86.04**	**87.82**	**82.18**	**88.61**
	Beam Search	84.38	86.13	80.76	85.69
	Beam Search (ensemble)	84.30	86.06	80.91	85.74

Ohashi et al., NAACL 2024

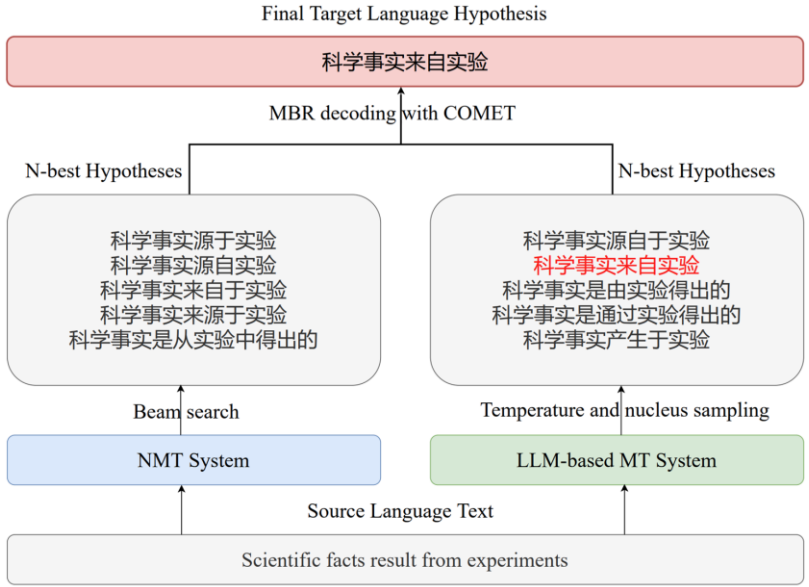
Applications of MBR Decoding

MBR Decoding for Machine Translation

Many submissions to WMT'24 use MBR Decoding

Models	en→xx		
	METRICX ↓	XCOMET↑	COMETKIWI ↑
Baselines			
NLLB-54B	7.61 7	66.90 7	57.01 7
GPT-4O	1.50 6	83.74 6	77.04 5
CLAUDE-SONNET-3.5	1.40 5	84.85 5	78.09 4
DEEPL	—	—	—
TOWER			
TOWER-V2 7B	1.48 5	83.77 5	77.02 5
TOWER-V2 70B	1.32 4	84.87 4	78.29 4
TOWER + QAD			
TOWER-V2 70B+MBR	0.92 2	88.78 2	81.39 3
TOWER-V2 70B+TRR	1.03 3	87.95 3	82.13 2
TOWER-V2 70B 2-step	**0.89** 1	**89.25** 1	**82.54** 1

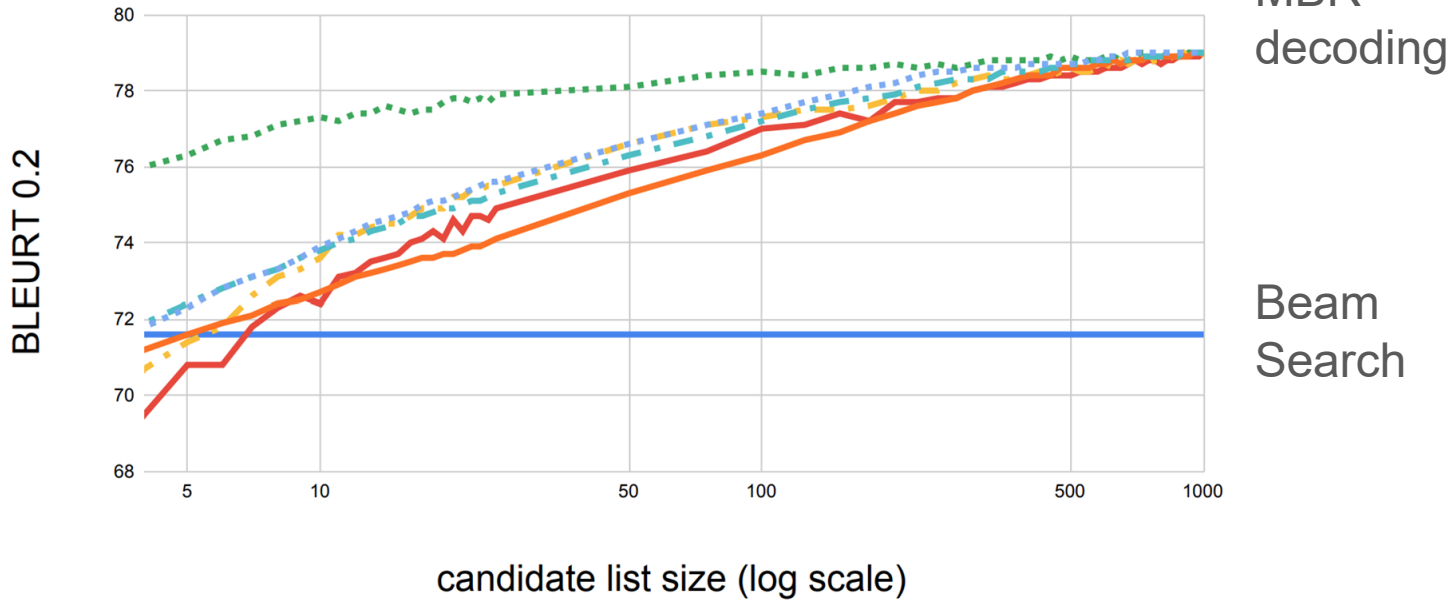
Rei et al., WMT 2024



Wu et al., WMT 2024

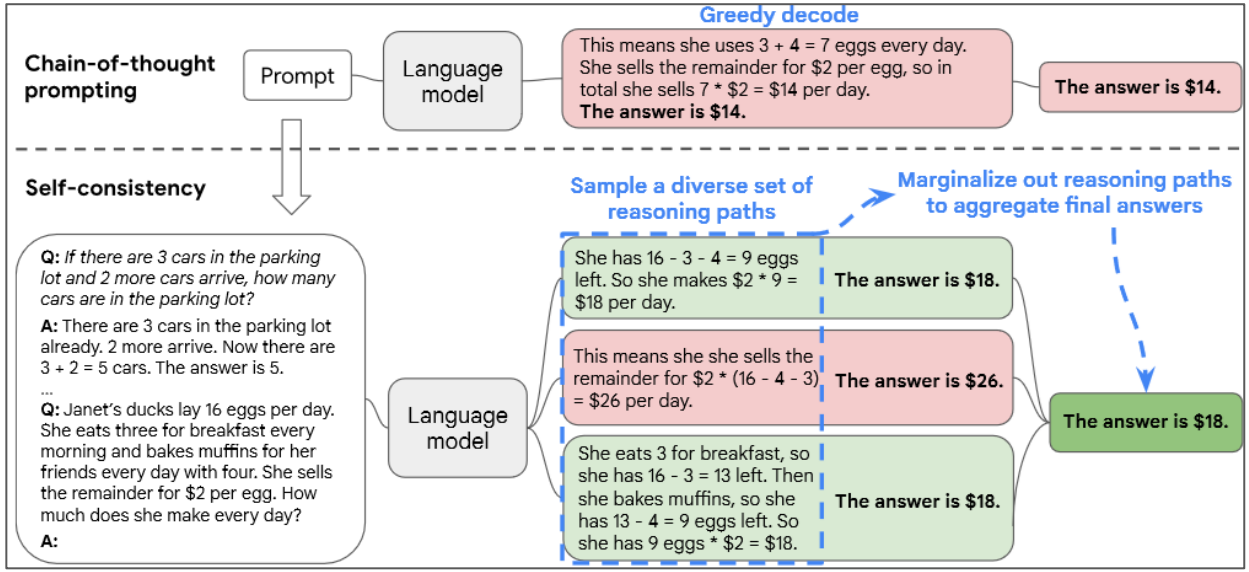
MBR Decoding for Machine Translation

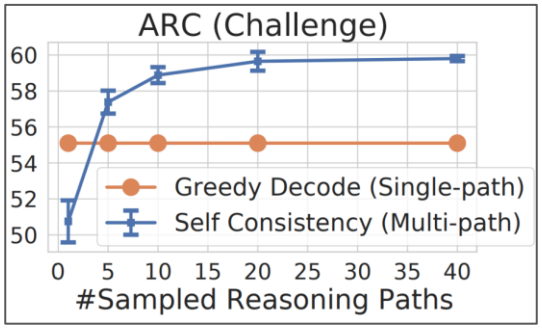
MBR Decoding is better than beam search



Freitag et al., TACL 2022

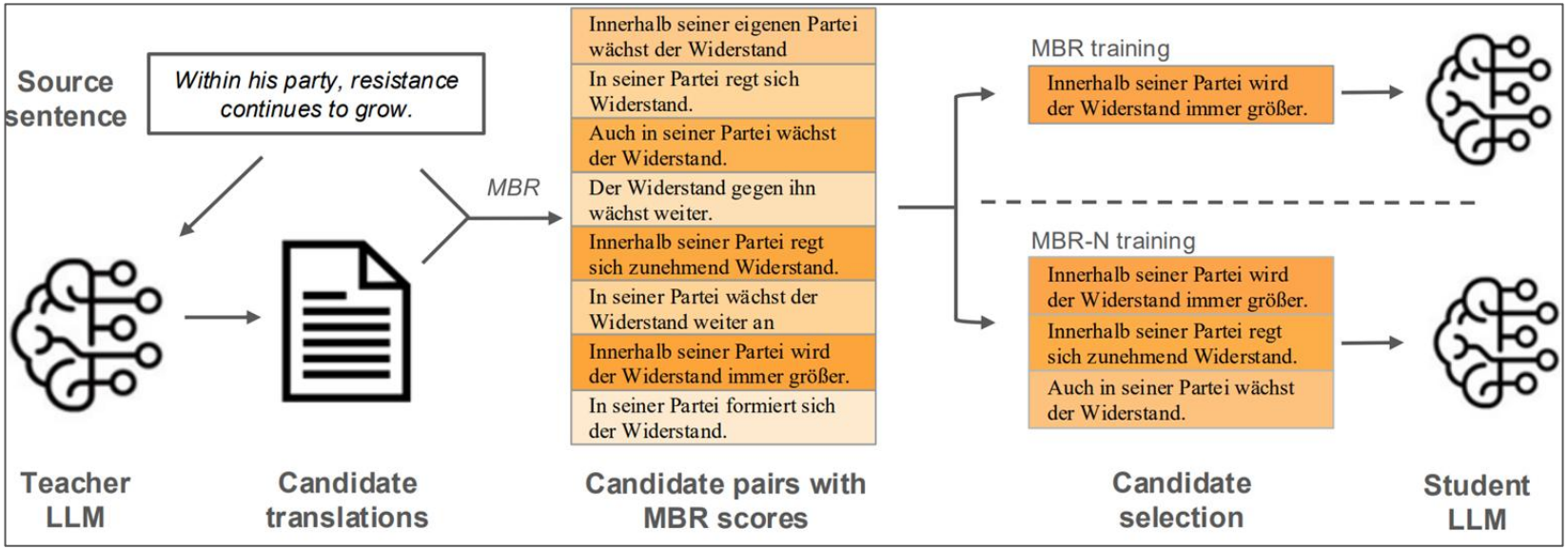
MBR with Chain-of-Thought a.k.a. Self-Consistency





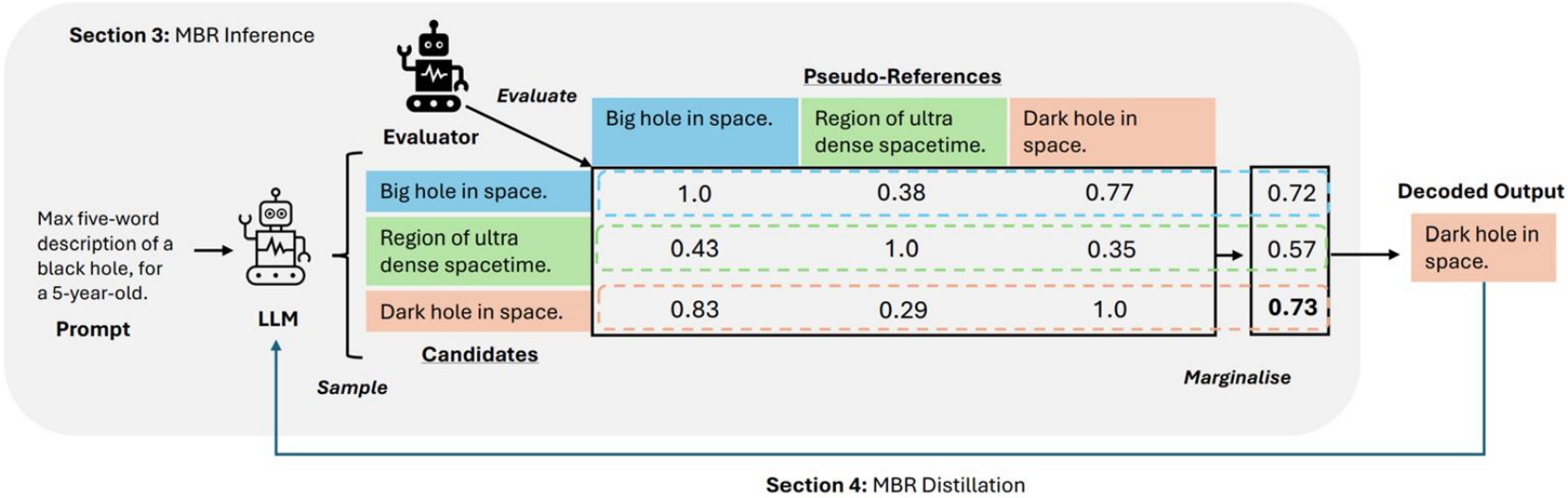
Wang et al. ICLR 2023

MBR for Distillation from Teacher LLM



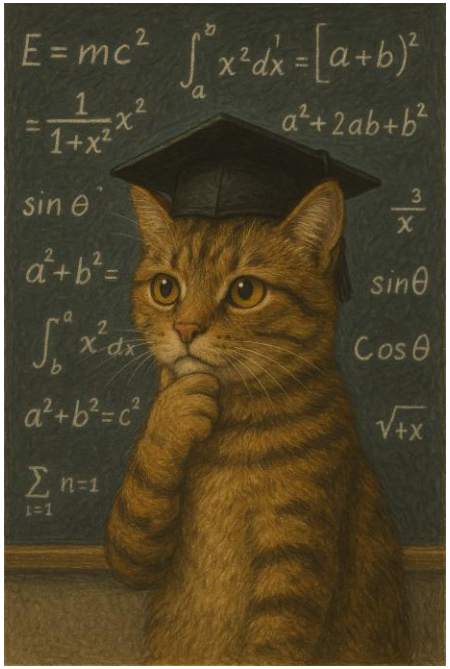
Wang et al. SSI-FM ICLR 2025

MBR for Self-Distillation



Wu et al. ICLR 2025

Why does MBR Decoding Work?

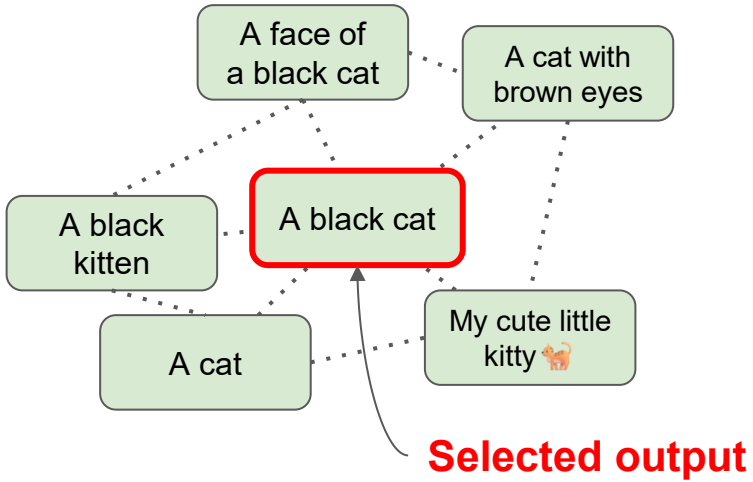


Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

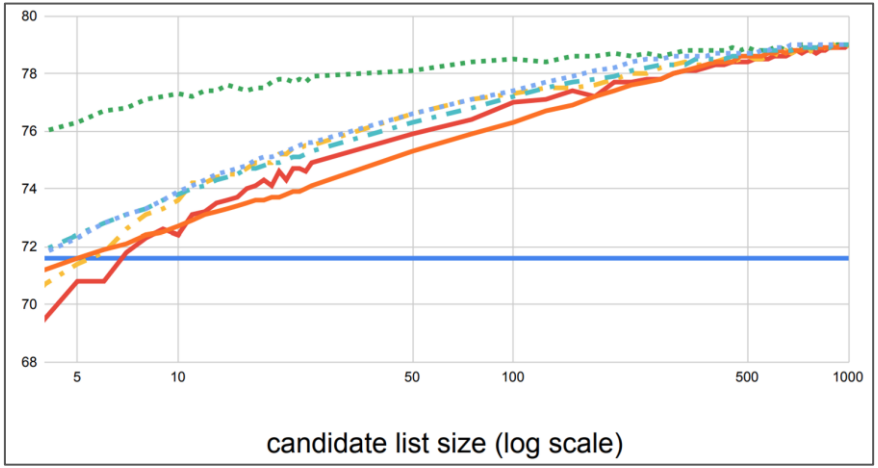
$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

Prompt: "What's in the picture?"



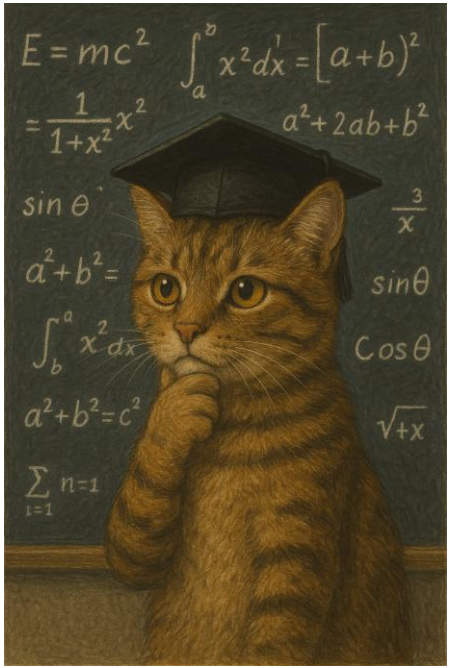
Why does MBR Decoding Work?

MBR Decoding only need finite samples (e.g., 100) to surpass the performance of beam search (state-of-the-art) whereas the number of possible sequences is infinite.



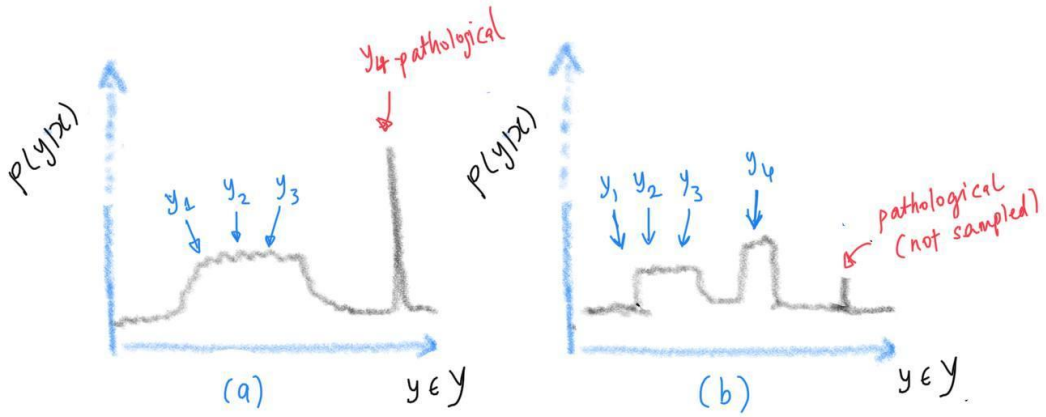
MBR decoding

Beam Search



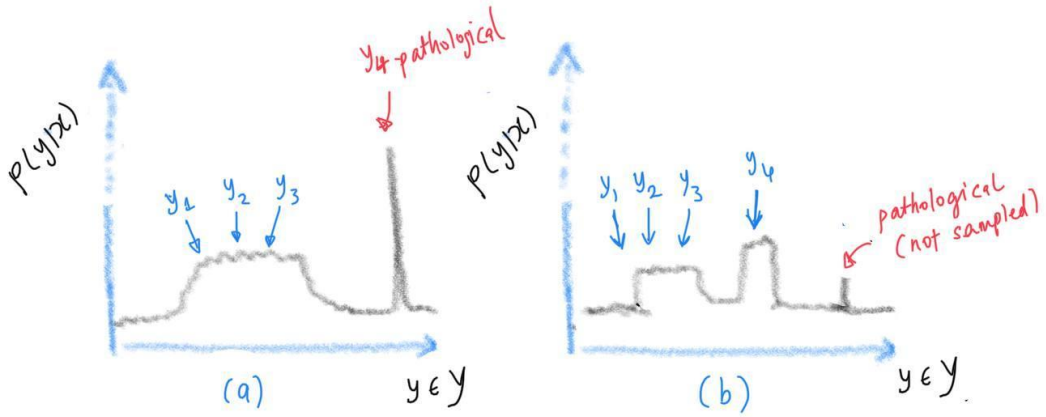
Freitag et al., TACL 2022

Hypothesis: MBR decoding is a Convolutional Filter



https://suzyahyah.github.io/bayesian%20inference/machine%20translation/2022/02/15/mbr-decoding.html

Hypothesis: MBR decoding is a Convolutional Filter



https://suzyahyah.github.io/bayesian%20inference/machine%20translation/2022/02/15/mbr-decoding.html

So far I could not find evidence

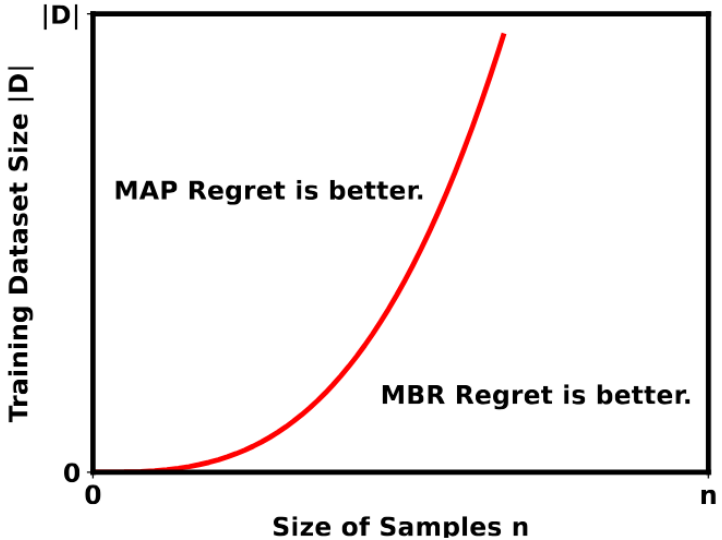
- MBR decoding only using k-neighbors did not improve over MBR
- Neighbors of the MBR output do not always have high probability

Minimum Bayes Risk Decoding Minimizes Bayes Risk (Ichihara et al., ACL 2025)

Which objective functions are easier to optimize, MAP or MBR?

- With large enough number of samples, MBR is likely to be better

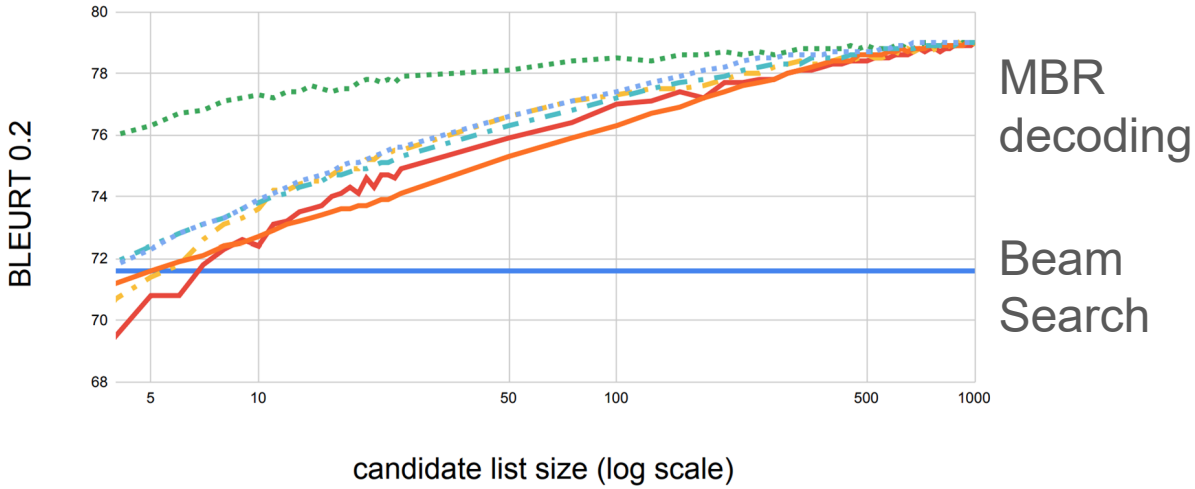
under assumptions



Minimum Bayes Risk Decoding Minimizes Bayes Risk (Ichihara et al., ACL 2025)

MBR decoding converges to the optimal solution with high probability at a rate of $O(1/\sqrt{n})$ where n is the number of samples

under assumptions



Freitag et al., TACL 2022

MBR Decoding as a Medoid Identification Problem (Jinnai&Ariu, Findings 2024)

$$\mathrm{h}_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\operatorname{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

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

A face of a black cat

A cat with brown eyes

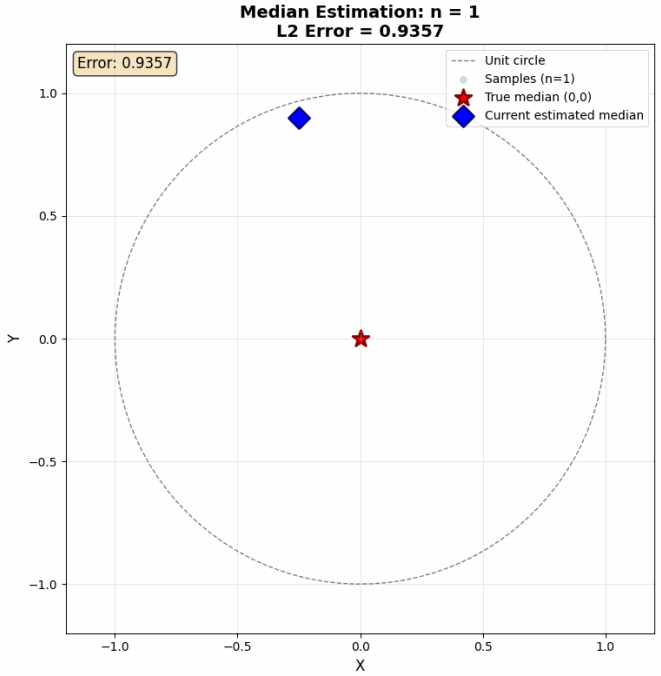
A black kitten

A black cat

A cat

My cute little kitty 🐈

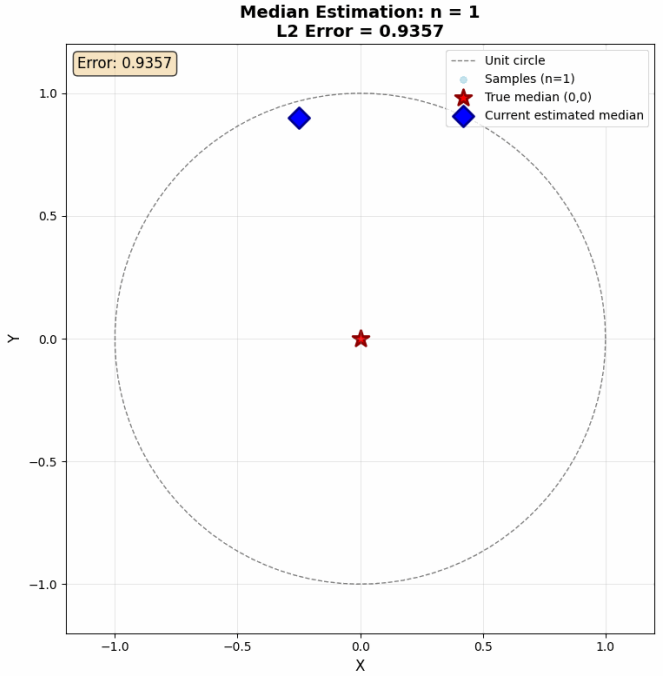
Selected output



MBR Decoding as a Medoid Identification Problem (Jinnai&Ariu, Findings 2024)

$$\mathrm{h}_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\operatorname{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

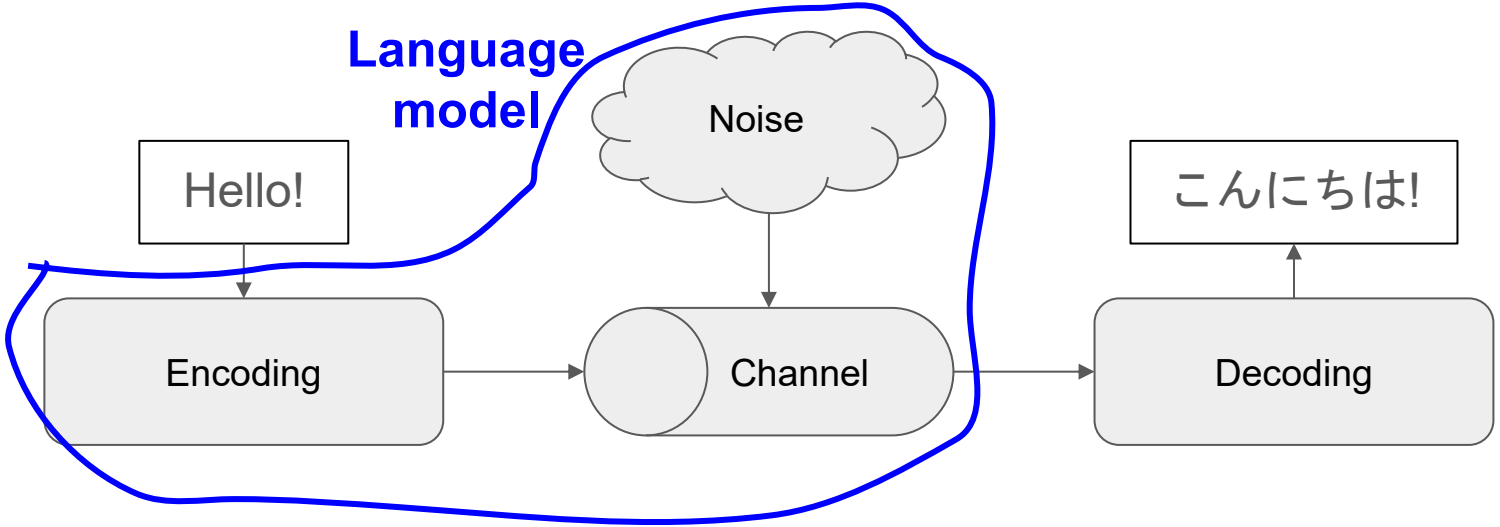
This entails that there exists an approximation algorithm with $\mathrm{O}(\mathrm{n}\ \log\ \mathrm{n})$



MBR Decoding as a Noisy Signal Decoding

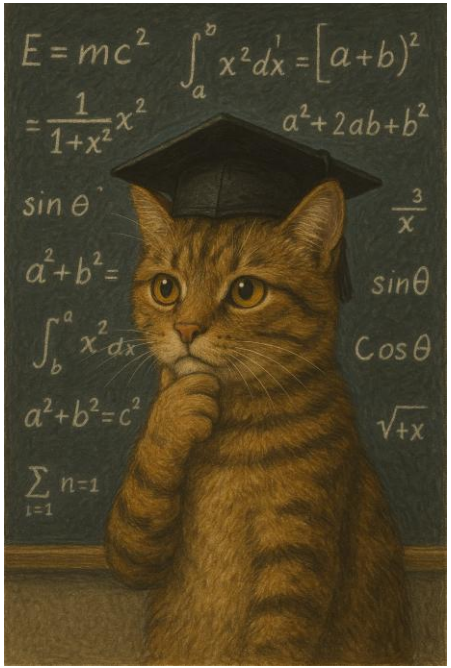
Random sampling has no bias but high variance

Noise (variance) can be ignored by sample-and-aggregate strategy



Why does MBR Decoding Work?

MBR Decoding only need finite samples (e.g., 100) to surpass the performance of beam search (state-of-the-art) whereas the number of possible sequences is infinite.



Still an open question!

Open Problems

MBR x Computational Linguistics

- **Information density, surprisal, exposure bias…**
  - These theories are often mentioned to explain why MAP decoding fails
  - Does it explain why MBR decoding is good?

e.g. The uniform information density hypothesis claims that human prefers to distribute information uniformly

How big is [$_{NP}$ the family$_i$ [$_{RC}$ (~~that~~) you cook for $_i$]]?

Native English speakers don’t want to omit this “that”

MBR for En-Ja, Ja-En Translation

- pfnet/plamo-2-translate is available for research!
  - But not for me… :(
- pfnet/plamo-2-translate-eval is a pairwise evaluation model which may be used in a different way than MBR decoding

Faster Inference: MBR Decoding is SLOW

MBR Decoding is 10-100 times more slower than beam search

Possible solutions

- **Faster computation of utility function** (e.g., Cheng&Vlachos, EMNLP 2023)
- **Efficient use of autoregressive model**
  - e.g. Speculative decoding (e.g., Sun et al., NeurIPS 2025)
- **Efficient sampling**
  - Non-iid sampling algorithm?

Summary

Questions: jinnai_yu@cyberagent.co.jp

Decoding Matters More than You Think!

Where Should I Start?

Starter kit for MBR decoding

- Sampling-Based Approximations to Minimum Bayes Risk Decoding for Neural Machine Translation (Eikema & Aziz, EMNLP 2022)
- High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics (Freitag et al., TACL 2022)
- Minimum Bayes-Risk Decoding for Statistical Machine Translation (Kumar & Byrne, NAACL 2004)

Implementations (Library)

- https://github.com/naist-nlp/mbrs
- https://github.com/ZurichNLP/mbr