NARA INSTITUTE of SCIENCE and TECHNOLOGY

Noisy-Channel Minimum Bayes Risk Decoding

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

ICML International Conference On Machine Learning
HALL A #3513, Tue, Jul 7, 2026 • 10:30 AM – 12:15 PM

TL; DR We introduce a noisy channel decomposition for Minimum Bayes Risk (MBR) decoding, providing a unified interpretation and improving interpretability through channel weighting.

Introduction & Motivation — What is the MBR Decoding?

Background: MBR decoding yields more *robust and higher-quality text generation* than maximum a posteriori (MAP) decoding by maximizing **expected utility** over sampled pseudo-references.

$$h^{\mathrm{MAP}} = \arg\max_{h \in \mathcal{H}} P(h|x;\phi)$$

$$h^{\mathrm{MBR}_\phi} = \arg\max_{h_i \in \mathcal{H}} \mathop{\mathbb{E}}_{r \sim P(r|x;\phi)} [f_\theta(h_i, r)] \quad \Rightarrow \quad h^{\mathrm{MBR}_{\mathrm{MC}}} = \arg\max_{h_i \in \mathcal{H}} \frac{1}{|\mathcal{R}|} \sum_{r_j \in \mathcal{R}} f_\theta(h_i, r_j)$$

$$\mathcal{R} := \{r_i \in \mathcal{Y} \mid r_i \sim P(r_i|x;\phi)\}$$

x: input sequence
$P(h|x;\phi)$: output probability of model ϕ
$\mathcal{H}$: set of output hypotheses
$\mathcal{R}$: set of pseudo-references
f_θ: utility function, e.g., BLEU, COMET, BERTScore ($\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$)
$\mathcal{Y}$: infinite output space

Problem: Hypothesis selection calculates expected utility scores conditioned on given pseudo-references, while commonly used evaluation metrics, e.g., BLEU, are inherently **asymmetric**.

⇒ Original MBR formulations do not explicitly capture directional asymmetry or prior effects, treating hypotheses and references interchangeably despite directional reference-to-hypothesis relationships.

Derivation — Noisy-Channel-Based Decomposition of MBR Decoding

Noisy-channel-based decomposition naturally accounts for such bidirectional interactions

$$\begin{aligned}
h &= \arg\max_{h_i \in \mathcal{H}} \frac{1}{|\mathcal{R}|} \sum_{r_j \in \mathcal{R}} f_\theta(h_i, r_j) \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} f_\theta(h_i, r_j) \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \frac{f_\theta(h_i, r_j)}{\sum_{y_{j'} \in \mathcal{R}} \sum_{h_{i'} \in \mathcal{H}} f_\theta(h_{i'}, y_{j'})} \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \frac{f_\theta(h_i, r_j)}{\sum_{h_{i'} \in \mathcal{H}} f_\theta(h_{i'}, r_j)} \cdot \frac{\sum_{h_{i'} \in \mathcal{H}} f_\theta(h_{i'}, r_j)}{\sum_{y_{j'} \in \mathcal{R}} \sum_{h_{i'} \in \mathcal{H}} f_\theta(h_{i'}, y_{j'})}
\end{aligned}$$

$$\begin{aligned}
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} P(h_i|r_j) P(r_j) \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \sqrt{P(h_i|r_j)^2 P(r_j)} \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \sqrt{\frac{P(r_j|h_i)P(h_i)}{P(r_j)} P(h_i|r_j) P(r_j)^2} \\
&= \arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \sqrt{P(r_j|h_i) P(h_i) P(h_i|r_j) P(r_j)}
\end{aligned}$$

1. **Hypothesis-to-reference likelihood:** $P(r_j|h_i)$
2. **Reference-to-hypothesis likelihood:** $P(h_i|r_j)$
3. **Hypothesis prior:** $P(h_i)$
4. **Reference prior:** $P(r_j)$

$$\arg\max_{h_i \in \mathcal{H}} \sum_{r_j \in \mathcal{R}} \sqrt{P(r_j|h_i)^\alpha P(h_i)^\beta P(h_i|r_j)^\gamma P(r_j)^\delta}$$

α, β, γ, and δ are hyperparameters for adjusting the importance of each term

Channel Weighing & Unified Interpretation

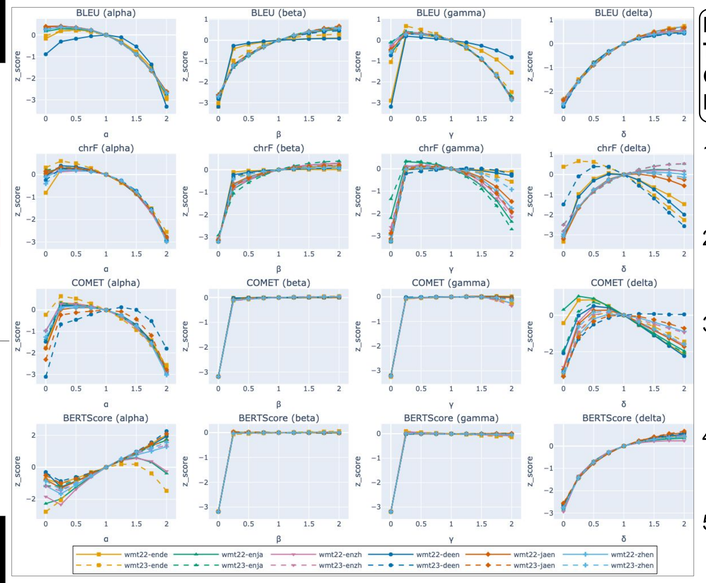
By adjusting the channel weights, it is possible to produce various interpretations of MBR decoding

Corresponding Method		MAP	Original	Ours (NEW)			
				Conditional	Swap	Inverse	Others
Formulation		$\mathcal{U}(0,1)$	$P(h_i\|r_j)P(r_j)$	$P(h_i\|r_j)$	$P(r_j\|h_i)P(h_i)$	$P(r_j\|h_i)$	$P(r_j\|h_i)^\alpha P(h_i)^\beta P(h_i\|r_j)^\gamma P(r_j)^\delta$
Hyperparameters	α	0	0	0	1	1	Others
	β	0	0	0	1	0	
	γ	0	1	1	0	0	
	δ	0	1	0	0	0	

- By parameterizing each term, we can capture changes in **"Others"**.
- By performing a **grid search** over the channel weights α, β, γ, and δ, we can understand the behavior of the MBR in greater detail.

Key Findings (RQ1: Utility Functions)

Do different "utility functions" exhibit consistent or characteristic trends across channels?

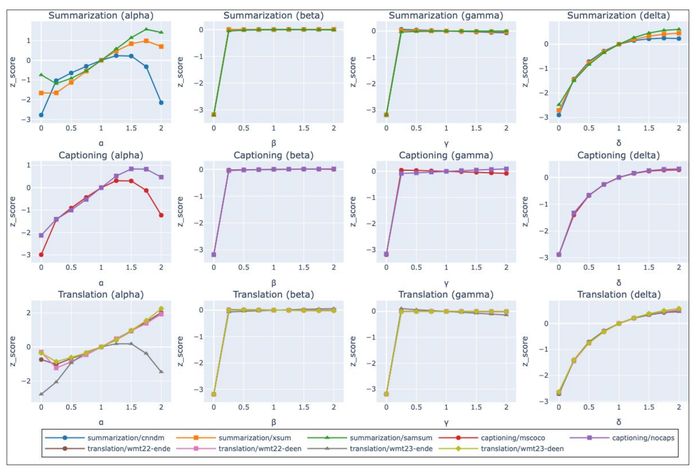


Setup
Metrics: BLEU, chrF, COMET, BERTScore
Task: Machine Translation (WMT22/23 en ↔ {de, zh, ja})
Grid Search: [0.0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.00]
Ref. & Hyp. : 256 sentences by eprision sampling (ε=0.02)

1. **Overview:** Channel-wise trends characterize individual utility functions, while the overall MBR behavior follows an almost consistent pattern.
2. **α:** The hypothesis-to-reference likelihood P(rj | hi) shows monotonic effects with respect to its weight and generally prefers low weighting, except for BERTScore.
3. **β:** The **hypothesis prior** P(hi) has limited impact for semantic-aware metrics such as COMET, but provides modest gains for surface-level metrics such as BLEU by acting as a ***semantic correctness term***.
4. **γ:** The **reference-to-hypothesis** likelihood P(hi | rj) is most effective with moderate low weights, while high weights can lead to performance degradation.
5. **δ:** The reference prior P(rj) dynamically influences performance, exhibiting metric-dependent stability while responding monotonically to its weight.

Key Findings (RQ2: Task-Agnosticity)

Do different "types of tasks" exhibit consistent channel-wise trends under the same utility function?



Setup
Metrics: BERTScore
Task: Summarization (CNN/DM, XSum, SAMSum), Captioning (MSCOCO, NoCaps), MT (WMT22/23)

6. The overall trajectory remains consistent across tasks and datasets, indicating that MBR decoding behavior is task-agnostic and dominated by the choice of utility function
7. Optimal channel weights are largely task-independent, and the hypothesis-toreference likelihood P(rj | hi) and the reference prior P(rj) dominate performance (BERTScore)

Conclusion:

- **Utility-function choice dominates behavior**
- **Channel trends are largely task-agnostic**
- **P(r|h) and P(r) are the most influential terms**
- **Appropriate channel weighting improves decoding**

(e.g., best weight wmt22 → wmt23: charF 50.18;vs MBR 50.05)