Investigating the Effectiveness of Multiple Expert Models Collaboration

Ikumi Ito[1], Takumi Ito[1, 2], Jun Suzuki[1, 3], Kentaro Inui[4, 1, 3]

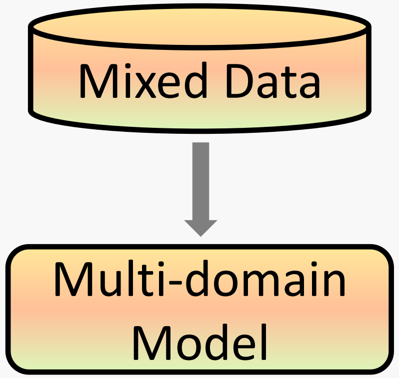
[1]Tohoku university, [2]Langsmith Inc., [3]RIKEN, [4]MBZUAI

Findings of EMNLP2023

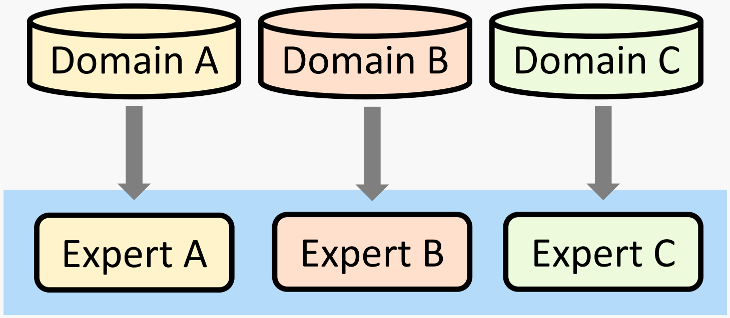
Explore the potential of multiple models in multi-domain translation

- One challenge in machine translation is multi-domain adaptation [Saunders+ ‘22]
- Main approaches for multi-domain
  - a single Multi-Domain Model (MDM)
  - multiple Domain Expert Models (DEMs)
- The contributions of this work
  - Demonstrated effectiveness of DEMs
  - Investigated effective collaboration methods in DEMs

MDM

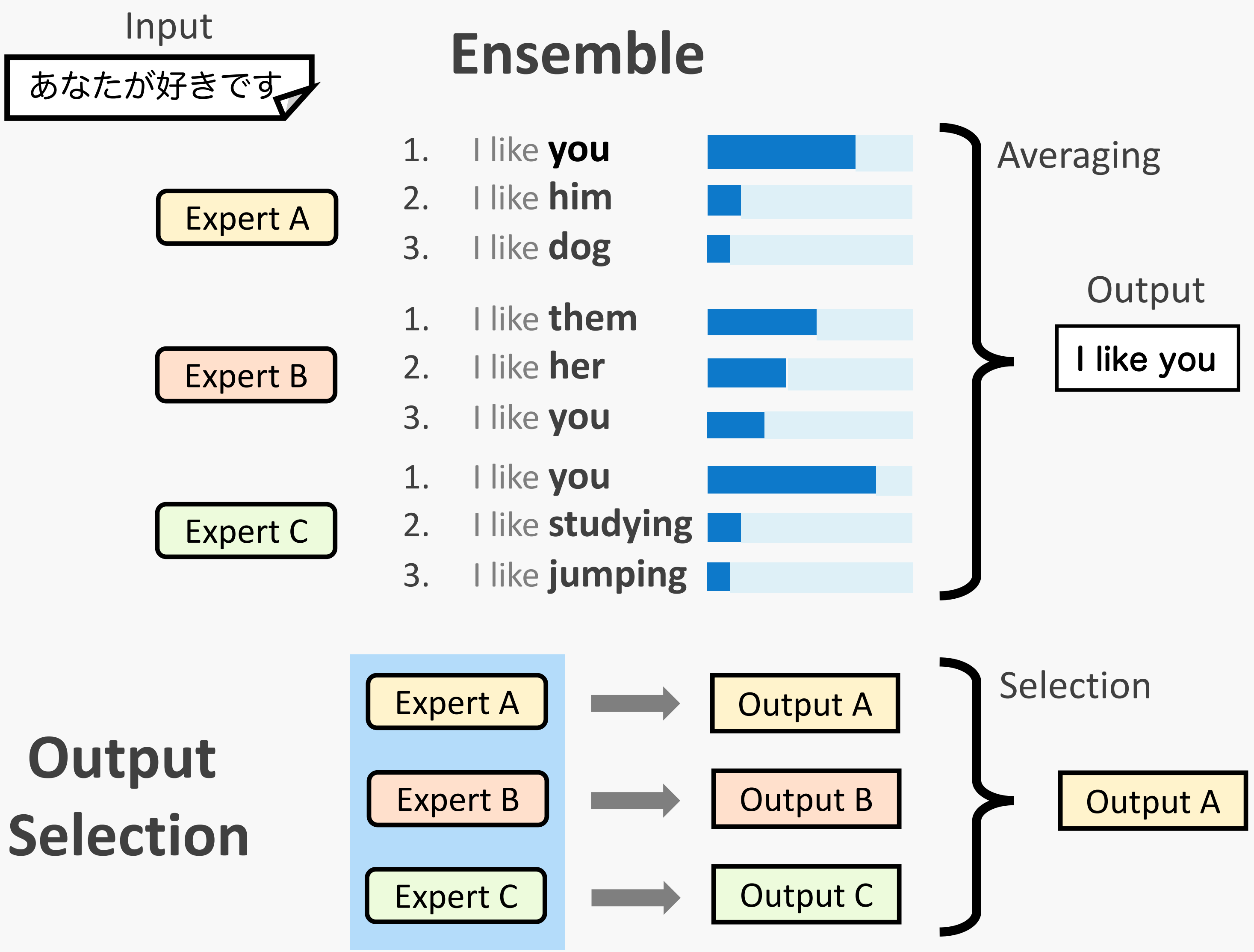


DEMs



How multiple models collaborate in DEMs

- Ensemble
- Output selection
  - Quality Estimation (QE)
  - Minimum Bayes Risk (MBR)



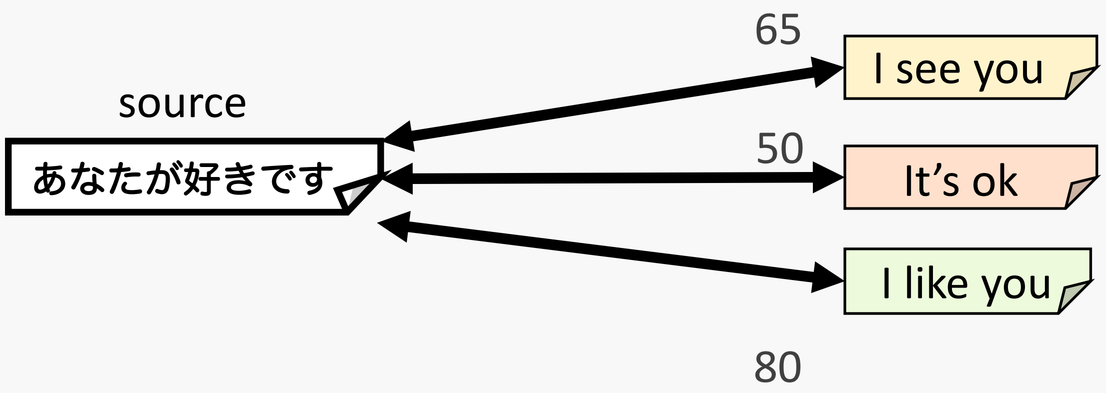
Quality Estimation (QE)

$$\arg\max_{C_i \in C} \mathrm{QE}\,(Src,\ C_i)$$

Quality Estimation (QE)

$$\underset{C_i \in C}{\arg\max}\ \boxed{\mathrm{QE}\,(Src,\ C_i)}$$

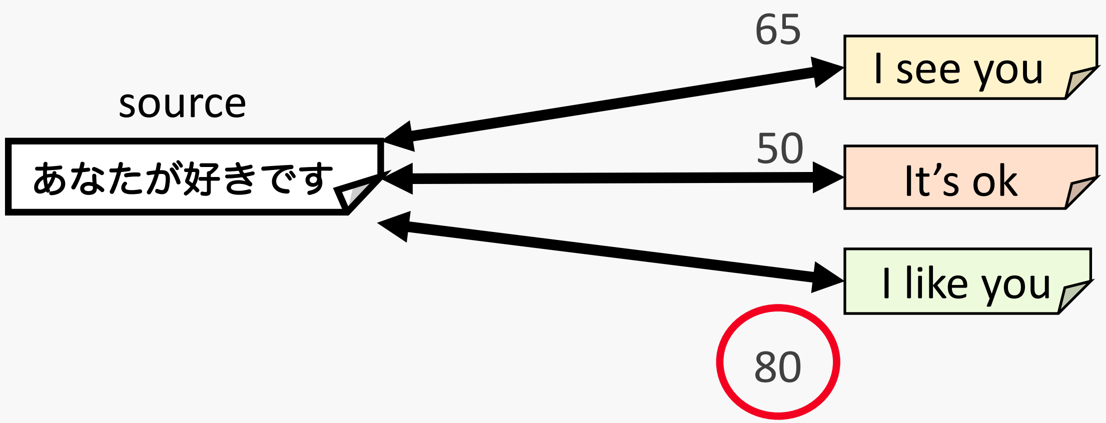
Calculate quality estimation metric score
we used MS-COMET-QE-22 [Kocmi+ '22]



Quality Estimation (QE)

$$\boxed{\underset{C_i \in C}{\arg\max}}\, \mathrm{QE}\,(Src,\ C_i)$$

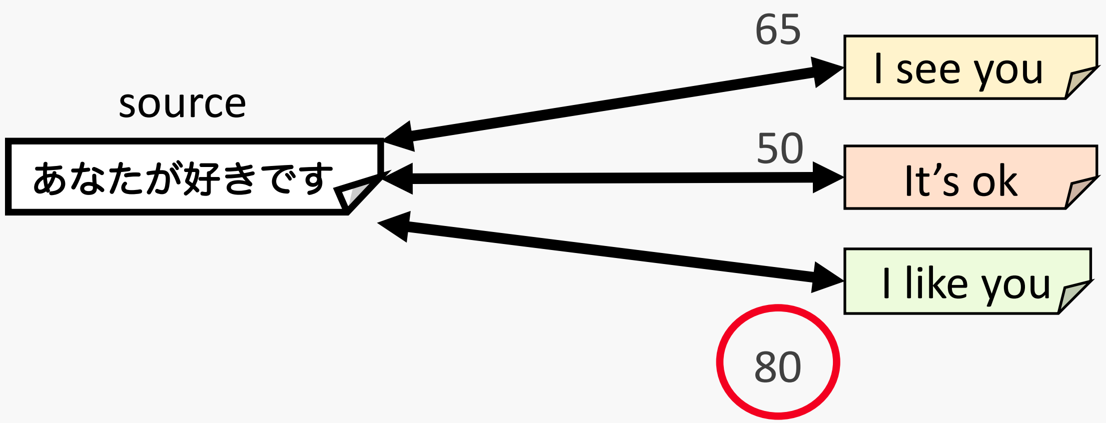
Select the one with the highest score



Quality Estimation (QE)

$$\underset{C_i \in C}{\arg\max}\, \mathrm{QE}\,(Src,\, C_i)$$

Select the candidate translation
with the highest
quality estimation score



What you need

- Source
- Each expert model’s outputs

→ Practical method

Minimum Bayes Risk (MBR)

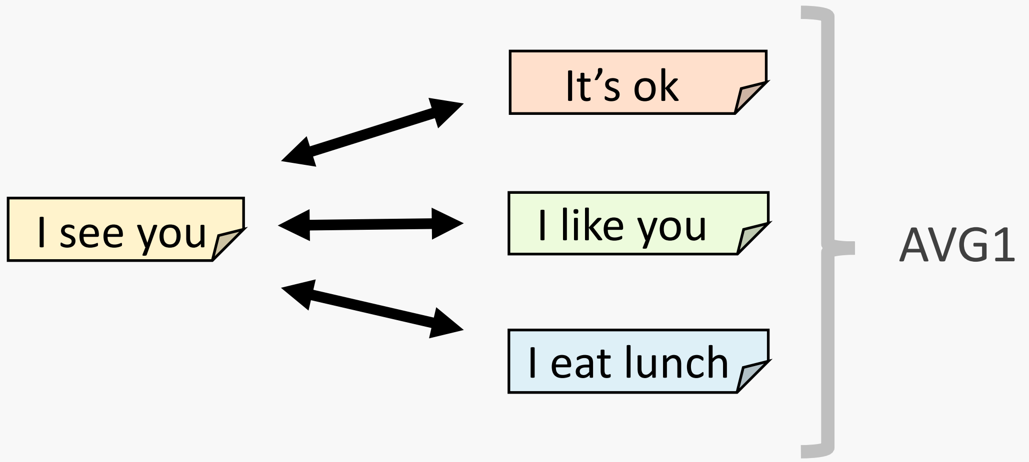
$$\underset{C_i \in C}{\arg\max} \frac{1}{|C|} \sum_{j \neq i} \mathrm{EVAL}\left(Src, C_i, C_j\right)$$

Minimum Bayes Risk (MBR)

$$\arg\max_{C_i \in C} \boxed{\frac{1}{|C|} \sum_{\boxed{j \neq i}} \mathrm{EVAL}\left(Src, C_i, \boxed{C_j}\right)}$$

pseudo-references

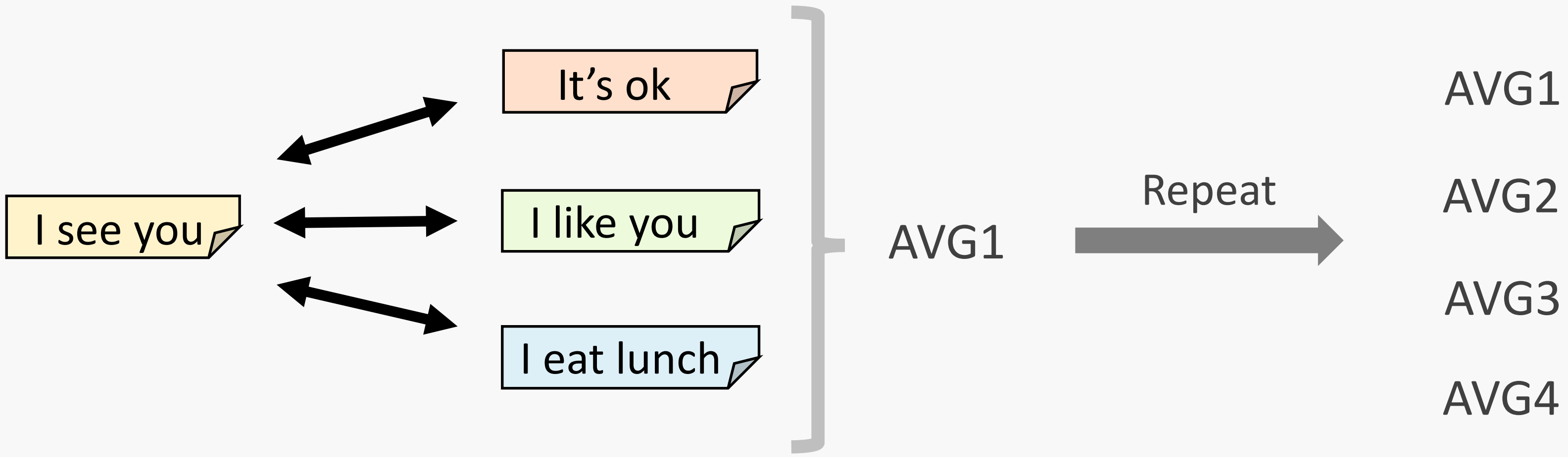
Evaluate against all other candidate translations
and take an average score.
We used MS-COMET-22 [Kocmi+ ’22]



Minimum Bayes Risk (MBR)

$$\underset{\boxed{C_i \in C}}{\arg\max}\ \boxed{\frac{1}{|C|}\sum_{j \neq i} \mathrm{EVAL}\left(Src, \boxed{C_i}, C_j\right)}$$

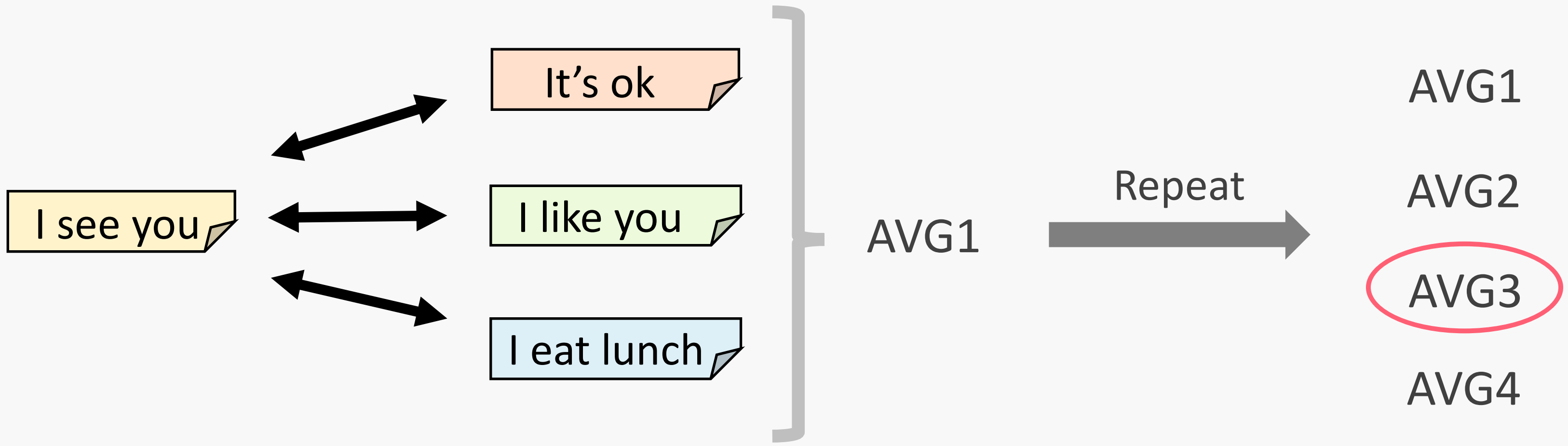
Repeat the AVG score calculation for all candidates



Minimum Bayes Risk (MBR)

$$\boxed{\underset{C_i \in C}{\arg\max}} \frac{1}{|C|} \sum_{j \neq i} \mathrm{EVAL}\,(Src, C_i, C_j)$$

Select the one with the highest score



Minimum Bayes Risk (MBR)

$$\arg\max_{C_i \in C} \frac{1}{|C|} \sum_{j \neq i} \mathrm{EVAL}\,(Src, C_i, C_j)$$

↓

Select a consensus output

What you need

- Source
- Each expert model's outputs

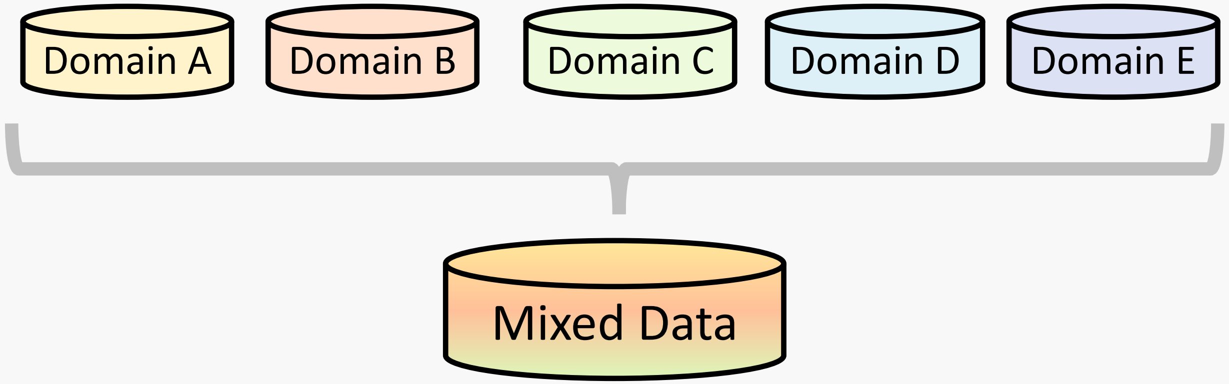
→ Practical method

Experimental settings

- Model: Transformer (90M, 290M, 1B)
- Dataset (En-Ja and Ja-En)
  - Pre-train: JParaCrawl v3.0
  - Fine-tuning: five-specific domain

Name	Params	Encoder & Decoder			
		layers	d_{model}	d_{ffn}	heads
SMALL	90M	6	512	2048	8
BASE	290M	6	1024	4096	16
LARGE	1B	6	2048	8192	32

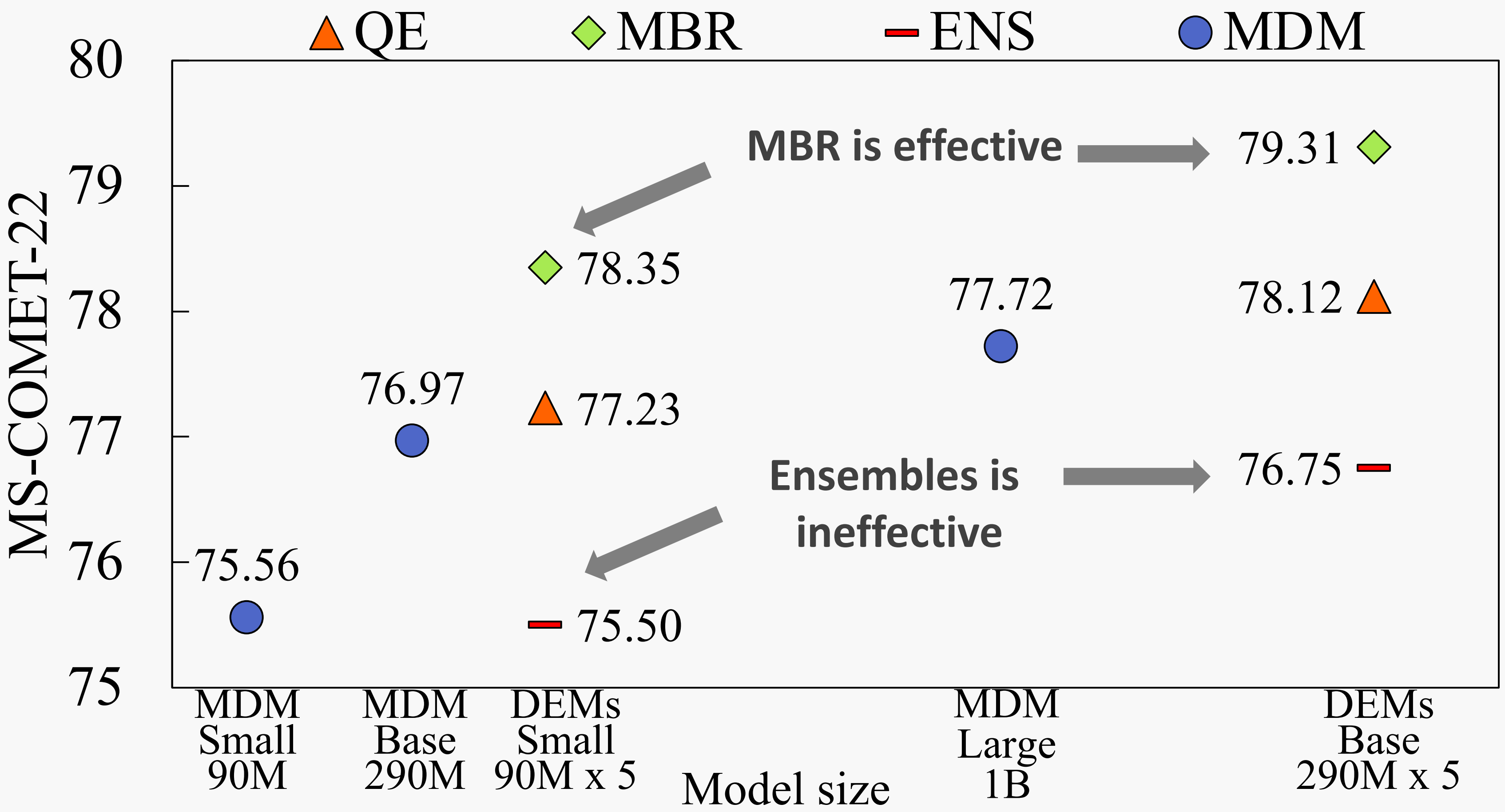
Model configurations



Dataset	#Sent Pairs
JParaCrawl v3.0	25.7M
The Kyoto Free Translation Task (KFTT)	440k
Japanese-English Legal Parallel Corpus (LAW)	260k
TED talks (TED)	225k
Asian Scientific Paper Excerpt Corpus (ASPEC)	200k
The Business Scene Dialogue corpus (BSD)	20k

Data information

Result: output selection from small experts is effective



※ Ja-En setting (same trend in En-Ja)

Summary: DEMs can be a hopeful direction in multi-domain

- The performance of 90M x 5 models was comparable to the 1B model
- Collaboration by MBR is effective, especially MBR
  - ※ The outputs of the small experts must include an output comparable to the output of the large model
  - ※ Selection model (e.g., MS-COMET-22) should address multi-domain