MarianMT, M2M100: A Translation Showdown

What drives translation quality across different Transformer architectures?

1. Intro:

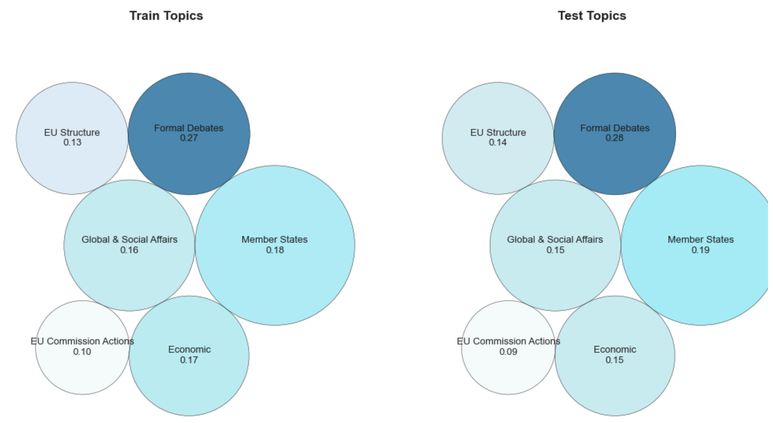
Translation quality varies across model architectures — but how do multilingual vs. monolingual systems flexibility affect this?

- **MarianMT** is a **monolingual** model trained on specific language pairs.
- **M2M100** is a **multilingual** model capable of translating between 100+ languages directly.

They differ in architecture and training philosophy — we compare them in terms of translation accuracy, adaptability responsiveness, robustness and hyperparameters optimization.

2. Dataset:

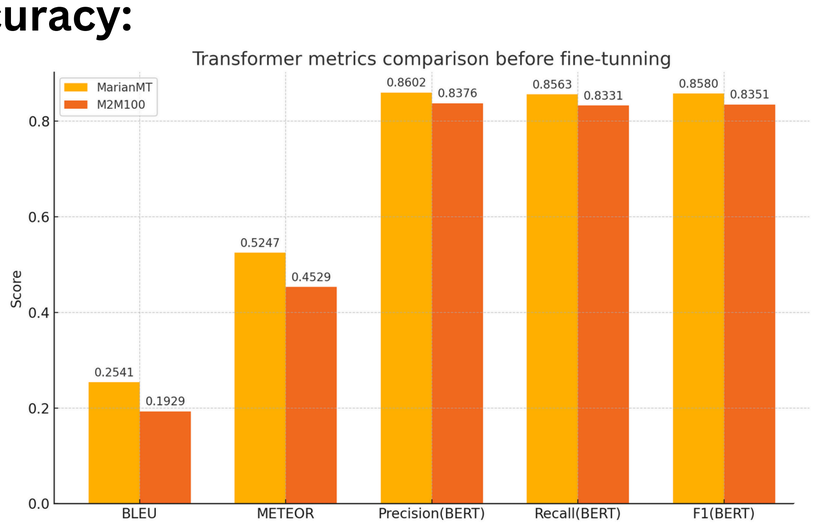
We used the WMT14 English–German dataset.
Clustering algorithm showed that many samples were **politically themed**.



This domain specificity adds value to our experiment — it allows us to observe **how well different transformer architectures adapt to style and preserve meaning during translation.**

3. Baseline Accuracy:

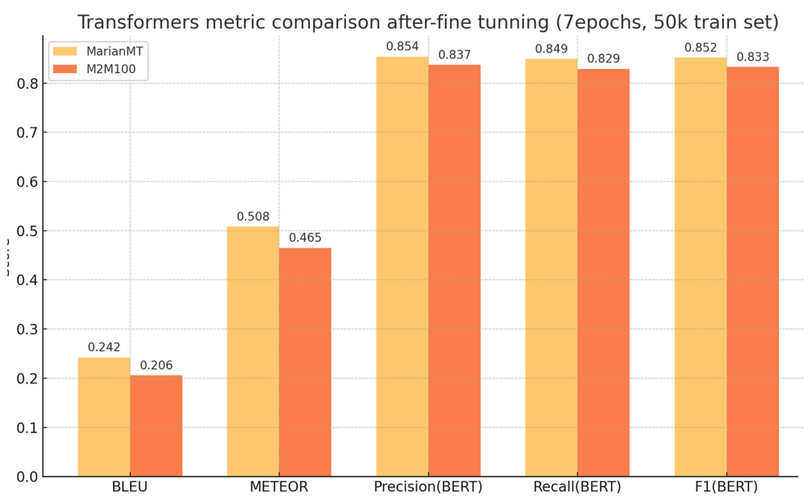
Before fine-tuning MarianMT consistently outperformed M2M100



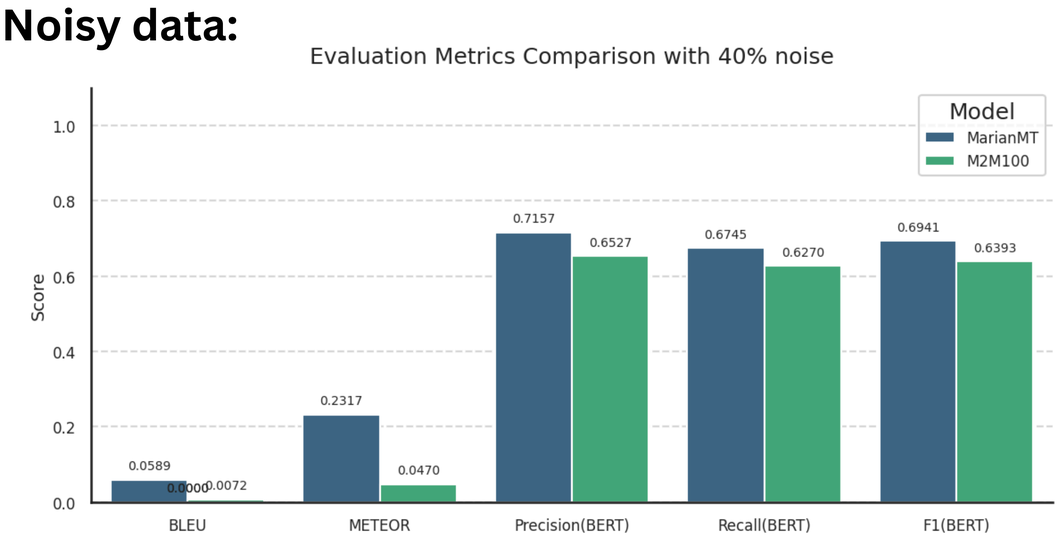
This suggests that **monolingual models like MarianMT offer higher initial accuracy on in-domain data**.

4. Fine-Tuning Impact:

Fine-tuning brings **limited gains** when the model is already language-aligned (MarianMT), while multilingual models (M2M100) **may benefit more in general use cases.**

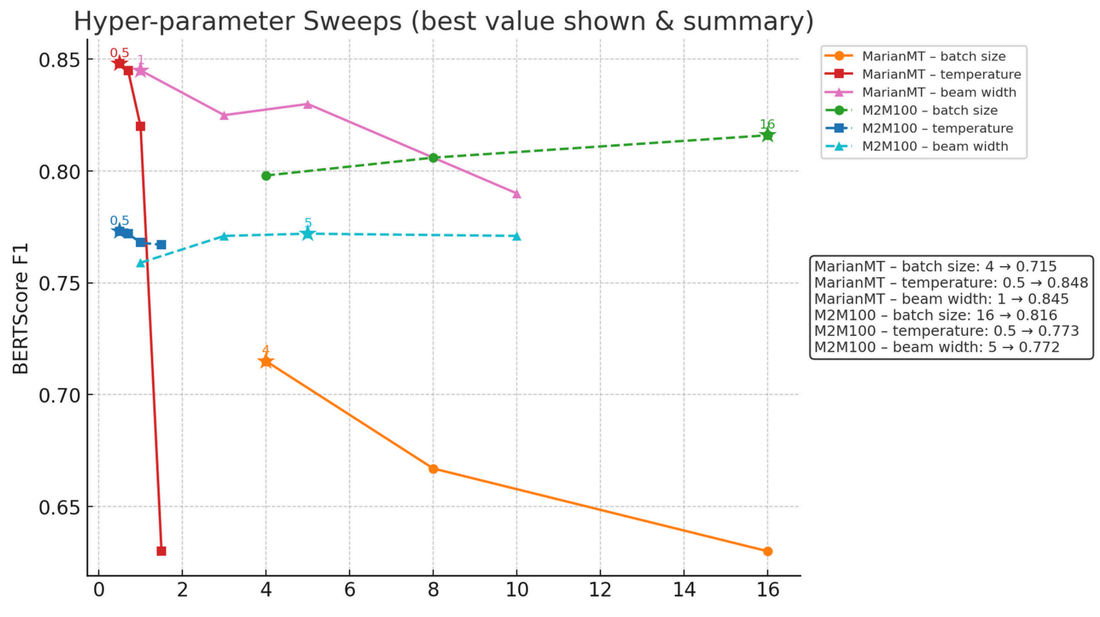


5. Noisy data:



noise = typos, deleted words, change/delete letters, synonyms, etc. The MarianMT model performs better than M2M100 when introducing noise

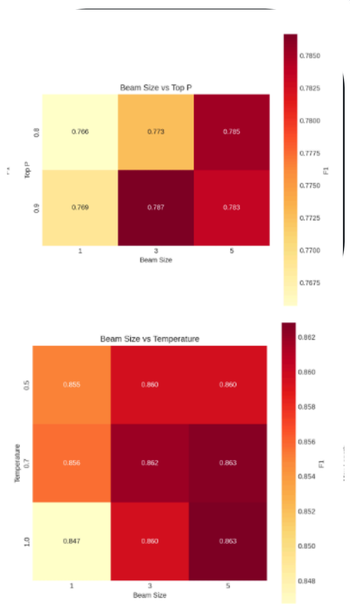
6. Hyperparameters:



7. Hyperparameters (contd.):

- **MarianMT** reaches its highest BERT-F1 when the search stays **compact but not deterministic**: a small beam (**3**) with moderate temperature (**0.7**) and nucleus sampling at p = **0.9**.
- **M2M-100** benefits from a **slightly deeper search -** beam = **5**, yet needs a cooler temperature (**0.5**) to keep outputs on-topic and p = **0.9**.

For both models, giving the decoder up to **150** tokens maximises quality.



8.Conclusion:

- **MarianMT** : **highest in-domain accuracy** “out of the box” and remains **robust on noisy data** thanks to pair-specific architecture and focused training.
- **M2M-100 : greater adaptability.** Fine-tuning and a deeper, cooler search (beam ≈ 5, T ≈ 0.5) closes most gaps while maintaining vast language coverage.

For both models, quality depends mostly on sensible decoding (beam 3–5, T 0.5–0.7, top-p 0.9, max 150).
Choose **MarianMT for instant, noise-tolerant precision** and **M2M-100 when flexibility and multilingual reach are crucial**.

Nastasă Denisa- Elena, Timur Jerčaks, Hitesh Satwani, Cecilia Mazzola, Viktoria Ivanova, Wenjie Liao, Jiang Haoting

DSAI 04