

Toward Domain-Specific Machine Translation and Quality Estimation Systems

Layman's Talk

Javad Pourmostafa Roshan Sharami

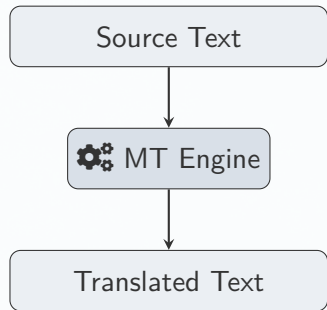
Cobbenhagen B – Aula
Tilburg University

March 4, 2026





Machine Translation (MT)

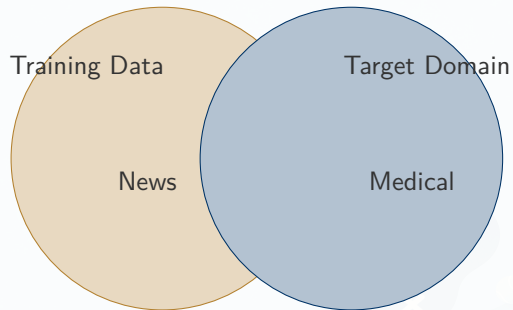


How do we evaluate MT?

1. Reference-Based Evaluation
Compare with a human translation
(e.g., BLEU, COMET)

2. Quality Estimation (QE)
Predict translation quality
without using a reference

Domain Mismatch in MT and QE



Limited overlap between domains → performance drops.



Challenges in Domain-Specific MT and QE

1. Limited In-Domain Data

MT systems (incl. LLMs) require domain data — but such data is scarce or unavailable.

2. QE Sensitivity to Domain Shift

QE is more fragile than MT — domain mismatch leads to unreliable estimates.

3. Tokenization & Vocabulary Mismatch

New terminology may not be properly represented by existing subword models.

4. Selecting Effective ICEs

ICL can replace fine-tuning — but selecting optimal examples requires human judgment.

Tackling Challenge 1: Lack of Domain-Specific Data



The Problem

Specialized data is needed — but often scarce and expensive.

Our Approach

Select the most relevant sentences from large mixed-domain data pool.

What We Found

Small, carefully selected data outperforms large mixed-domain training.

Less data → same performance → **reduced energy consumption** 🌿

Takeaway

Smart data selection can replace large-scale training.

Challenges in Domain-Specific MT and QE

1. Limited In-Domain Data

MT systems (incl. LLMs) require domain data — but such data is scarce or unavailable.

2. QE Sensitivity to Domain Shift

QE is more fragile than MT — domain mismatch leads to unreliable estimates.

3. Tokenization & Vocabulary Mismatch

New terminology may not be properly represented by existing subword models.

4. Selecting Effective ICEs

ICL can replace fine-tuning — but selecting optimal examples requires human judgment.

Tackling Challenge 2: Making QE Domain-Aware



The Problem

QE models struggle when the *topic changes* or when *data is limited*.

Our Approach

Gradual adaptation: general data → mixed data → domain-specific data.

What We Found

QE, like MT, needs domain adaptation.

Adapted QE improves reliability — especially for *low-resource* and *unseen languages*.

Better domain fit → more trustworthy and stable scores.

Takeaway

Without domain awareness, QE can be misleading — and expensive.

Challenges in Domain-Specific MT and QE

1. Limited In-Domain Data

MT systems (incl. LLMs) require domain data — but such data is scarce or unavailable.

2. QE Sensitivity to Domain Shift

QE is more fragile than MT — domain mismatch leads to unreliable estimates.

3. Tokenization & Vocabulary Mismatch

New terminology may not be properly represented by existing subword models.

4. Selecting Effective ICEs

ICL can replace fine-tuning — but selecting optimal examples requires human judgment.

Tackling Challenge 3: Tokenization and Vocabulary Design

Tokens are the building
blocks of Large Language
Models (LLM).

The Problem

New domain terms emerge — but tokenizers may fragment them.

Our Approach

A systematic analysis of tokenizer and vocabulary sources: pre-training vs. in-domain data.

What We Found

We derive practical guidelines:

- Mismatched setups → unstable training
- In-domain tokenization → higher accuracy at higher cost
- Hybrid setups → balanced trade-off

Takeaway

For smaller NMT models, careful tokenizer design improves translation quality.

Challenges in Domain-Specific MT and QE

1. Limited In-Domain Data

MT systems (incl. LLMs) require domain data — but such data is scarce or unavailable.

2. QE Sensitivity to Domain Shift

QE is more fragile than MT — domain mismatch leads to unreliable estimates.

3. Tokenization & Vocabulary Mismatch

New terminology may not be properly represented by existing subword models.

4. Selecting Effective ICEs

ICL can replace fine-tuning — but selecting optimal examples requires human judgment.

Tackling Challenge 4: Selecting Effective ICEs Automatically



The Problem

LLMs can use ICL instead of fine-tuning — but selecting the right examples requires human judgment, and no single prompt works across domains.

Our Approach

A domain-specific QE-guided search algorithm that:

- Scores candidate ICEs
- Selects the most impactful examples

What We Found

- Higher performance with fewer, high-quality examples
- Faster inference and more consistent translations, *reducing post-editing effort*

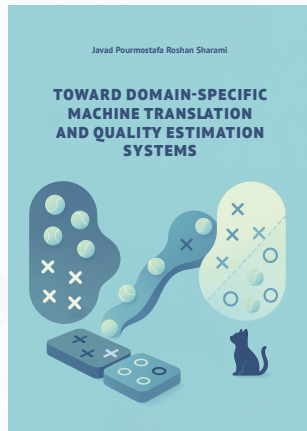
Takeaway

Quality-aware ICE selection enables domain adaptation — without fine-tuning.

Key Insights

- 1 Smart data selection replaces large-scale training.
- 2 Domain-adaptive QE improves reliability.
- 3 Tokenization design impacts adaptation success.
- 4 QE-guided ICL enables domain-specific MT in LLMs.

Read more: javad.pourmostafa.me/phdthesis



Available on Bol.com :-D
but **OP = OP**