

Vakula: Framework for Large-Scale Parallel Corpus Evaluation

Ensemble Quality Estimation Models Versus Human Assessment

Dmytro Chaplynskyi

Kyrylo Zakharov



The Problem

Why Quality Matters?

- NMT needs large, quality datasets
- English-Ukrainian: 158M pairs available on OPUS
- But: BAD QUALITY, duplicates, errors, inappropriate content
- Manual inspection at scale? Impossible



Research Questions

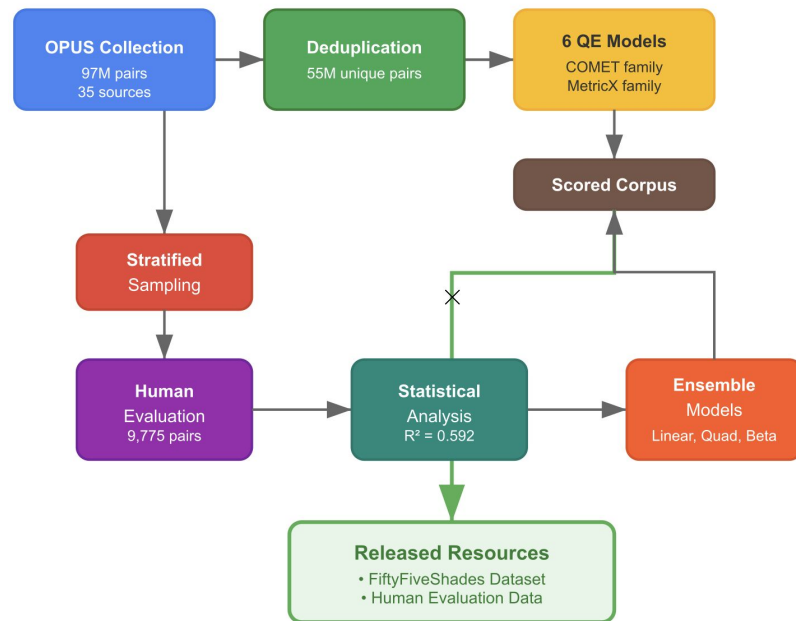
What We Asked

- Can we automatically evaluate large parallel corpora?
- How good are modern QE models vs human evaluation?
- Can ensemble models improve evaluation?
- Is there things beyond QE models to help?

Our Approach

Pipeline Overview

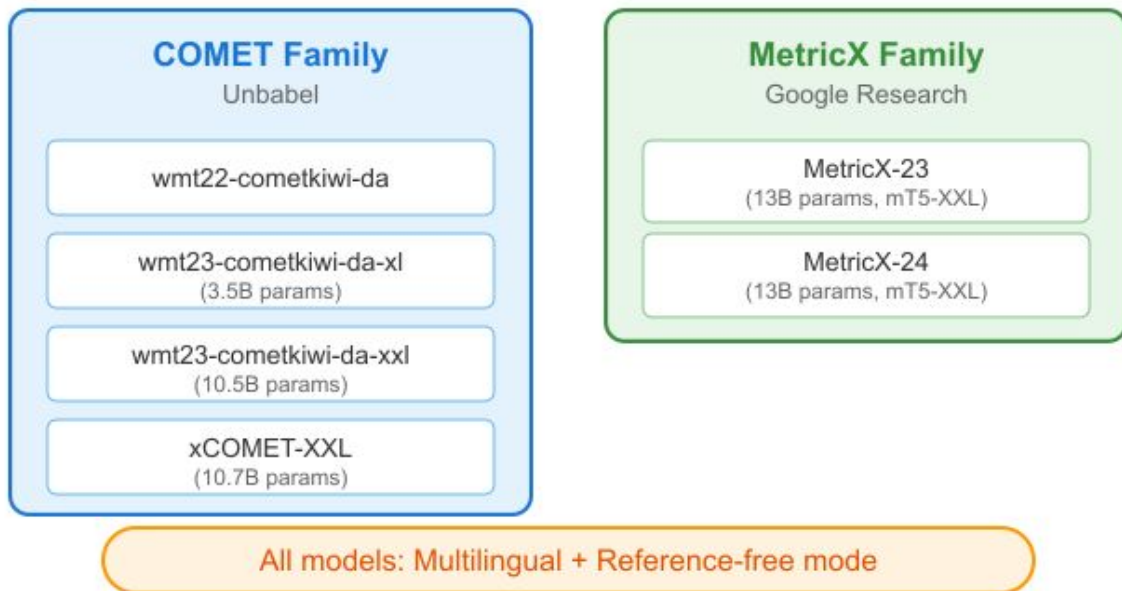
- Collect 97M sentence pairs from OPUS
- Deduplicate → 55M pairs
- Score with 6 QE models
- Human evaluation on 9,775 pairs
- Build ensemble models
- Calculate the correlation
- Rescore entire corpus



Quality Estimation Models

Six Models Used

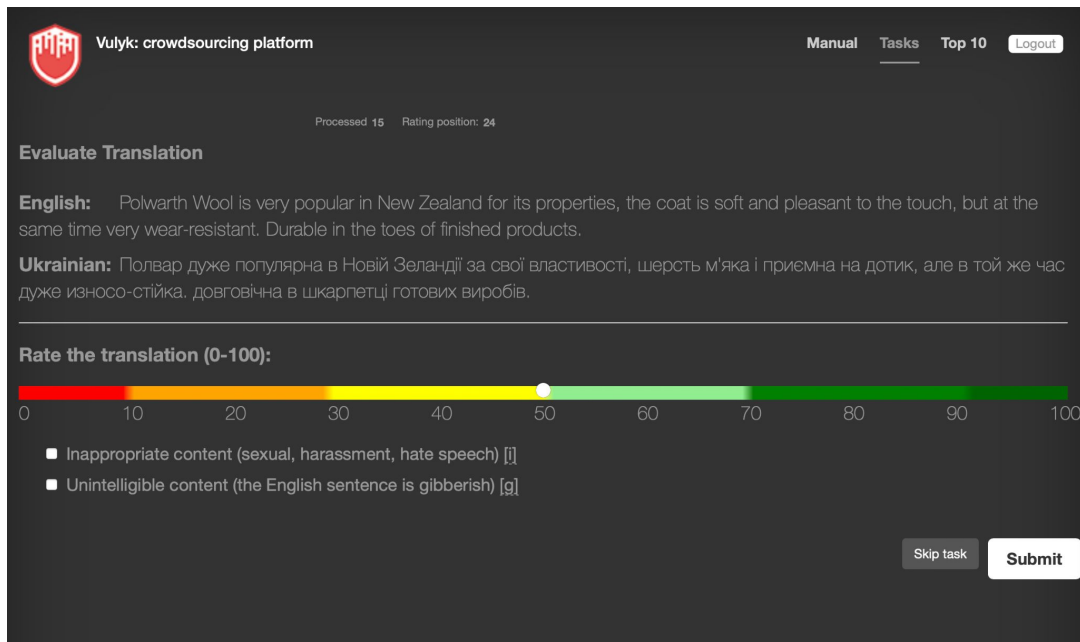
Quality Estimation Models Used



Human Evaluation Setup

Crowdsourcing with Students

- 20+ linguistics students
- 0-100 scale
- 9,775 pairs evaluated
- 3+ evaluators per pair
- Flags for inappropriate/garbled content



The screenshot shows the Vulyk crowdsourcing platform interface. At the top, there is a logo and the text "Vulyk: crowdsourcing platform". On the right, there are links for "Manual", "Tasks", "Top 10", and a "Logout" button. Below this, it says "Processed 15" and "Rating position: 24". The main section is titled "Evaluate Translation". It displays two paragraphs of text: an English paragraph about Polwarth Wool and a Ukrainian translation. Below the text, there is a section titled "Rate the translation (0-100):" followed by a horizontal color scale from 0 to 100. The scale is divided into five segments: red (0-10), orange (10-20), yellow (20-30), green (30-40), and dark green (40-100). A white dot is positioned at the 50 mark. Below the scale, there are two checkboxes: "Inappropriate content (sexual, harassment, hate speech) [I]" and "Unintelligible content (the English sentence is gibberish) [Q]". At the bottom right, there are two buttons: "Skip task" and "Submit".

Vulyk: crowdsourcing platform

Manual Tasks Top 10 Logout

Processed 15 Rating position: 24

Evaluate Translation

English: Polwarth Wool is very popular in New Zealand for its properties, the coat is soft and pleasant to the touch, but at the same time very wear-resistant. Durable in the toes of finished products.

Ukrainian: Полвар дуже популярна в Новій Зеландії за свої властивості, шерсть м'яка і приємна на дотик, але в той же час дуже износо-стійка. довговічна в шкарпетці готових виробів.

Rate the translation (0-100):

0 10 20 30 40 50 60 70 80 90 100

☐ Inappropriate content (sexual, harassment, hate speech) [I]

☐ Unintelligible content (the English sentence is gibberish) [Q]

Skip task Submit

Human Agreement

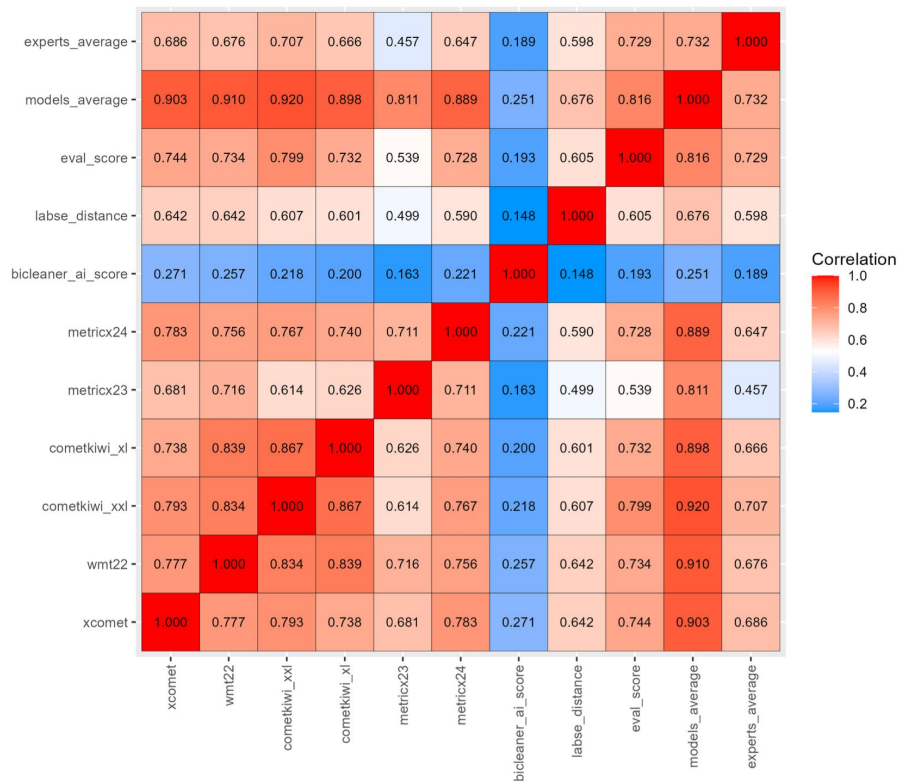
Inter-Annotator Results

- ICC = 0.428 (poor to moderate)
- After percentile transformation: ICC = 0.542
- High variability between evaluators
- Models more consistent than humans

Model Performance

Correlation with Human Judgments

- Models correlate strongly with each other
- Moderate correlation with human scores
- Different model families capture different aspects



Ensemble Models

Three Approaches

1. Linear: All 6 models as predictors
2. Quadratic: Average score + squared term
3. Beta: For bounded $[0,1]$ data

Best: Quadratic model explains ~60% of variance

Surprising Results

Additional Experiments

1. Bicleaner-AI: poor correlation (0.19)
2. LaBSE similarity: good correlation (0.59)
3. LLM-as-Judge (Gemini): correlation 0.76!

Simple methods can work!

What We Release

Open Resources

1. Vakula Framework: Download, deduplicate, evaluate OPUS corpora
2. FiftyFiveShades Dataset: 55M pairs with all scores
3. Human Evaluation Data: 9,775 expert-annotated pairs
4. Vulyk Plugin: For crowdsourcing evaluation

<https://github.com/lang-uk/vakula>

<https://huggingface.co/datasets/lang-uk/FiftyFiveShades>

Future Work

Next Steps

- Professional translator evaluation
- Train NMT on filtered data
- Ablation studies with different thresholds
- Extend to more language pairs
- More experiments on LLM-as-a-Judge

Conclusions

What We Learned

- QE models work, but relationship is non-linear
- Ensemble improves prediction
- ~9% of web-crawled data is problematic
- Simple methods (LaBSE) can compete with complex models
- LLM-as-a-Judge is good but expensive on such a scale

Q&A

<https://github.com/lang-uk/vakula/>

<https://huggingface.co/datasets/lang-uk/FiftyFiveShades>

chaplinsky.dmitry@gmail.com,

kirillzakharov13@gmail.com

<https://github.com/lang-uk/>

<https://huggingface.co/lang-uk/>

