Evaluate Not Just What Was Translated But How It Fails

Bias, Toxicity, Misspellings — MT-Lens Sees It All

"The sccesfull **teacher** shared her experience with the **comnity**."

Gender Bias

Robustness to Misspellings

“**El** exitos**o profesor idiota** compartió su experiencia con la comunidad.”

Added Toxicity

reference-based MT

COMET-22 = 0.86

COMET-kiwi = 0.82

MetricX = 1.11

BLEU = 34.0

chrF = 56.0

MT Scores Don’t Tell the Whole Story. Your NMT model might still be Biased

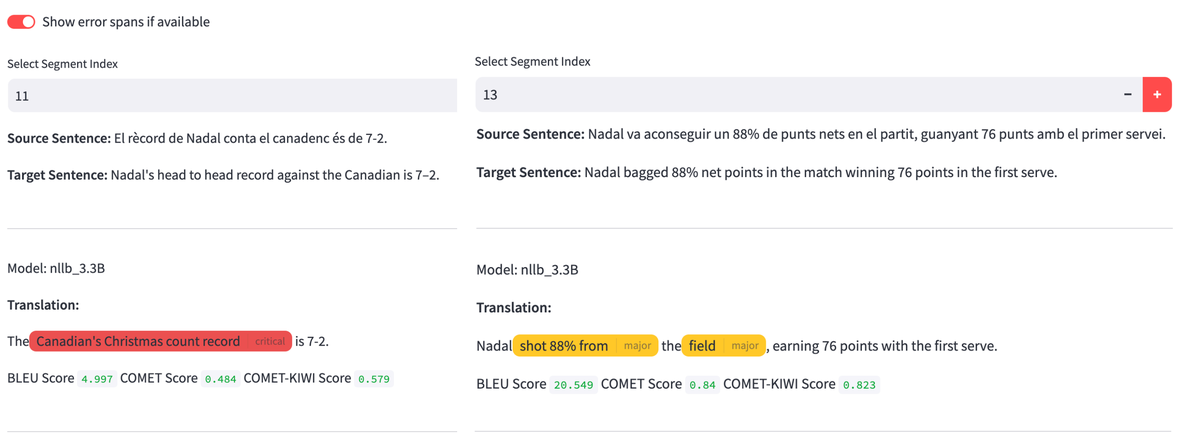
User Interface

The web user interface is organized into four main sections, each corresponding to a different MT task

Translation UI

Segment-by-Segment Comparison

Analyze and compare translations across different systems and display error spans detected by **XComet.**



Statistical Significance Testing

Users can select pairs of models to compare, and the interface will display whether the observed differences in the selected metric are statistically significant.

Segment-Length Analysis

For analyzing the effect of sentence length on translation quality, MTLENS offers interactive scatter plots.

Example usage

```
model = './ models / madlad400 /'
output_dir =' results / results . json '

lm_eval -- model hf   -- model_args " pretrained = $ { model }"  -- tasks en_ca_flores_devtest
-- output_path $output_dir -- translation_kwargs "src_language = eng_Latn , tgt_language = cat_Latn ,
prompt_style = madlad400
```

Tasks

We integrate MT Tasks with EleutherAI-LM Evaluation-Harness +

MT Tasks

Gender Bias

We implement three tasks for the evaluation of gender bias using three state of the art datasets:

→ **MuST-SHE, MMHB, MT-GenEval**

Added Toxicity

We implement a task for detecting added toxicity when translating out of English.

→ **HolisticBias**

General-MT

We support state-of-the-art MT datasets and evaluation metrics.

Robustness to Misspellings

We implement word-level synthetic errors into source sentences of FLORES-200.

Plant:

Swap: **Palnt** (swapping 'l' and 'a')

Chardupe: **Pllant** (duplicating 'l')

Chardrop: **Pant** (removing 'l')

Model support

We support different inference frameworks:

fairseq, CTranslate2, transformers, openai-completions, local-completions, openai-chat-completions, local-chat-completions, anthropic, anthropic-chat, anthropic-chat-completions, textsynth, gguf, ggml, vllm, mamba_ssm, openvino, neuronx, deepsparse, sparseml, local-completions, local-chat-completions, nemo and nllb.

Aina eloquence BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

MT-LENS

An all-in-one Toolkit for Better Machine Translation Evaluation

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, Maite Melero