Multi-Dimensional Machine Translation Evaluation: Model Evaluation and Resource for Korean

Dojun Park, Sebastian Padó

IMS, University of Stuttgart, Germany

LREC-COLING 2024

Different Aspects of Translation Quality

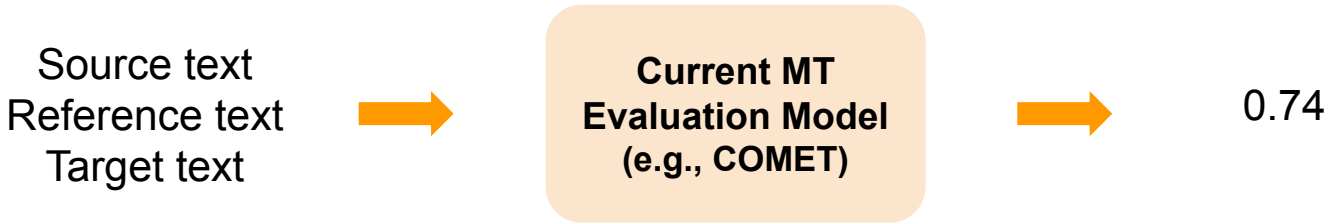
Die Katze jagte die Maus.
(The cat chased the mouse.)

➡ The cat chased the ball.
Inaccurate meaning. (Accuracy)

➡ The cat the mouse chased.
Not grammatical. (Fluency)

➡ The cat found itself in pursuit of the mouse.
Too formal. (Style)

Current Machine Translation Evaluation Paradigm



MQM: Multidimensional MT Evaluation

Major Cat.	Minor Cat.	Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize distinct errors.

Shortcomings and Contributions

- MQM annotation is well established in manual translation evaluation (Freitag et al., 2021)
- **Few computational studies**: Scarcity of MQM-annotated corpora

- Our contributions:

1. A 1200-sentence MQM evaluation benchmark for English-Korean in three key dimensions: *accuracy, fluency, and style.*
2. MT evaluation as **multi-task prediction** of the three MQM scores
    - Result: More details and also better overall (single-score) evaluation

Part 1: Constructing an English–Korean Parallel Dataset

Parallel Corpora

- Selected two parallel corpora: Global Voices & TED Talk 2020
- Randomly sampled 600 translation pairs, summing up to 1200

Paraphrasing

- Paraphrased the English source by gpt-3.5-turbo

Translation

- Translated into Korean using Google Translate

Annotation

- Selected three error dimensions: accuracy, fluency and style.
- Annotated in two severity levels: major and minor errors.

Validation of the MQM Scores

We ask two questions:

(a) Are the MQM scores that we have obtained reliable?
(b) Do they provide us with additional information compared to single-score metrics?

(a) Reliability of annotated MQM scores

Accuracy	Fluency	Style
0.54	0.57	0.34

- Inter-annotator agreement for annotated scores measured as correlation (Kendall's tau)
- Result: Fluency and Accuracy reliable (strong correlation), style harder to annotate reliably

(b) Information content of annotated MQM scores

	Accuracy	Fluency	Style	BLEU
Accuracy	1			0.17***
Fluency	0.29***	1		0.15***
Style	0.10***	0.01	1	0.08***

The correlation coefficients show that MQM scores capture nuances potentially missed by single-score metrics like BLEU.

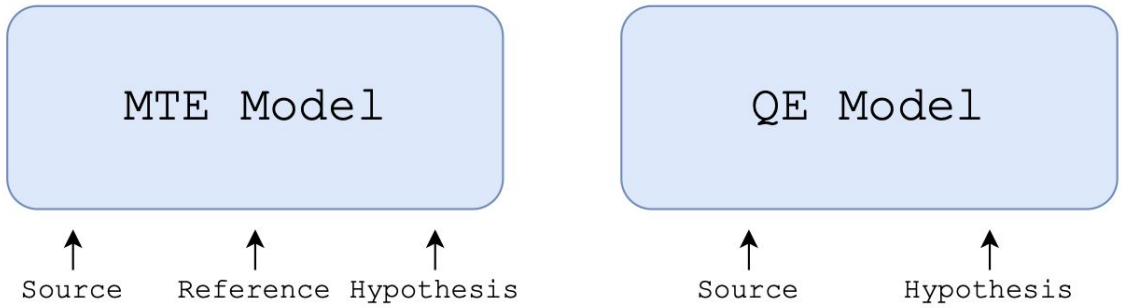
Example of the MQM-annotated Dataset

	type	corpus	en_source	ko_reference	ko_target	annotation	accuracy	fluency	style	total
0	Training	GlobalVoices	To let Japanese buy FT will be the last thing ...	중국은 닛케이가 파이낸셜타임스를 인수하는 걸 절대 바라지 않는다.	중국은 일본 고객의 FT 구매를 최종 선호로 보고 싶지 않을 것입니다.	Accuracy: 일본 고객의 (mistranslation/major), 최종 선호로...	35.0	5.0	0.0	40.0
1	Training	GlobalVoices	If our uncritical engagement with media is any...	미디어와의 이런 안식없는 관계는 우리가 스스로와도 밀접하게 연관되어 있지 못하는 것...	우리가 언론을 세심하게 살피지 않는 걸 보면 우리 일에는 관심조차 없는 것 같다.	Accuracy: 우리가 언론을 세심하게 살피지 않는 걸 보면(mistranslat...	7.0	0.0	0.0	7.0
2	Training	GlobalVoices	China: Rising prices and rooftop gardens · Glo...	중국: 물가상승에 따른 새로운 현상- 옥상텃밭	글로벌 보이스(Global Voices)는 중국의 치솟는 물가와 옥상 정원의 등장에...	Accuracy: 등장에(addition/minor), (Global Voices)...	3.0	2.0	10.0	15.0
3	Training	GlobalVoices	And one year of a brave stance against great e...	그리고 일 년 간 거대한 악에 대항하는 용기도 볼 수 있었다. 그러나 시리아인들은 ...	1년 내내 강력한 세력에 맞서는 용기를 보여줬음에도 불구하고, 시리아 시민들은 자유...	Accuracy: 세력에 (mistranslation/major), 존경을(mistr...	10.0	1.0	0.0	11.0
4	Training	GlobalVoices	Amina's story deeply touched and outraged Moro...	모로코 누리꾼들은 아미나의 이야기를 듣고 분노했고, 트위터 해쉬태그 #RIPAmin...	Amina의 이야기는 Twitter에서 해시태그 #RIPAmina를 사용하여 소녀를...	Accuracy: Amina, Twitter(untranslated text/maj...	10.0	10.0	0.0	20.0
...	...	...	...	...	...	...	...	...	...	...

Part 2: Modeling Multidimensional Translation Quality

Using a handful of SOTA models as base models, we compared different setups including:

- MTE Models taking the source, reference, and target texts.
- QE Models that operate without a reference text.



Result of Model Performance

Model	Accuracy	Fluency	Style	Overall
RemBERT-MTE	**0.40**	0.38	0.26	0.35
RemBERT-QE	0.35	**0.43**	**0.33**	**0.37**

- We conducted experiments with multiple models across various settings; RemBERT, in both the MTE and QE setups, stood out.
- While the MTE model outperformed in accuracy, the QE model excelled in fluency and style, resulting in a higher overall average score.

Comparison against COMET

	Model	Accuracy	Fluency	Style	Overall
MTE	RemBERT	0.40	0.38	0.26	0.35
	COMET-22	**0.30**	0.10	0.02	0.28
QE	RemBERT	0.35	0.43	0.33	**0.37**
	CometKiwi	**0.40**	0.09	0.17	**0.37**

- **Balanced Model Performance**: Our models show a more balanced correlation across dimensions.
- **Style Sensitivity**: CometKiwi exhibits an enhanced correlation for style.
- **Trade-Off in Model Training**: Extensive dataset vs. language-specific training.

Conclusion and Key Findings

- **RemBERT Performance:** RemBERT stood out among other models, particularly in the QE setup, demonstrating a notable edge over competitors.

- **Fine-grained Evaluation:** This approach enhances interpretability, allowing for a fine-grained analysis of various translation quality dimensions.

- **Balanced Evaluation:** Our methodology provides a balanced evaluation across different aspects of translation quality, achieving results comparable to established COMET metrics in terms of overall quality scores.

Limitations

- **Single Language Pair**: Our focus on the English-Korean language pair may limit the generalizability of the results to other language pairs.

- **Dimension Weighting**: A simple average method for weighing different MQM dimensions for the overall quality may not reflect the nuanced needs of other translation scenarios.

- **Lack of Cross Lingual Methods**: Transfer learning can be employed to leverage existing MQM data from other language pairs.