NLP Lab Assignment : Analysing MQM Error Profiles and Metric Correlations in MT Systems

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

15th October 2025

1 Introduction

Evaluating the output of machine translation (MT) systems has evolved significantly over the past two decades. Traditional approaches such as adequacy and fluency scales, and later direct assessment (DA), provide valuable information about overall quality but offer limited insight into *why* systems perform as they do.

In recent years, more fine-grained error annotation schemes have emerged. One of the most influential is **Multidimensional Quality Metrics (MQM)**, which allows human annotators to classify and weight errors according to categories such as *Accuracy*, *Fluency*, *Style*, or *Terminology*. This makes it possible to go beyond overall scores and understand the specific linguistic weaknesses of different systems.

In this lab, we will work with the MQM dataset released alongside :

> Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). *Experts, Errors, and Context : A Large-Scale Study of Human Evaluation for Machine Translation.* TACL.

The dataset contains expert MQM annotations for multiple MT systems submitted to the WMT 2020 shared task. Ricardo Rei provide an extended version of the dataset (https://huggingface.co/datasets/RicardoRei/wmt-mqm-human-evaluation) that maps each sentence to its source and reference translation. By using these two datasets, it is possible to add automatic metric scores (BLEU, chrF, COMET) to each translation. The aim of this assignment is to analyse **error profiles of systems**, compare them, and explore how they align with automatic metrics.

2 Objectives

By completing this assignment, you will :
— Understand how MQM error categories can be used to characterise MT systems.
— Build and visualise error distribution profiles for multiple systems.
— Perform system clustering based on their MQM error profiles and interpret the results.
— Compare human evaluation (MQM) with automatic metrics (BLEU, chrF, COMET).
— Explore how surface and linguistic features of the text affect the correlation between human and automatic metrics.

3 Dataset and Tools

We will use the **public MQM dataset** provided by Freitag et al. (2021), available at :

https://github.com/google/wmt-mqm-human-evaluation

This dataset includes :
— Source segments in English, German and Chinese.
— System translations (En→De and Zh→En).
— Expert MQM annotations : category, sub-category, severity, and span.
— Automatic metric scores (BLEU, chrF, COMET).

4 Methodology

Data Preparation Begin by loading the MQM annotation files for one language pair (for example, En→De). Extract the following columns from the TSV file :
— `system`
— `category` (e.g. Accuracy, Fluency, Style)
— `severity` (e.g. Minor, Major)
— `source`, `target`
Clean the data to ensure that category labels are consistent. You may also want to filter out entries with missing annotations.

Building and Visualising Error Profiles For each system, calculate the frequency of errors in each MQM category. Normalise these frequencies by the number of segments or total number of tokens. Represent each system as a vector of category frequencies :

$$v_{\text{system}} = [\text{Accuracy}, \text{Fluency}, \text{Style}, \text{Terminology}, \ldots]$$

If desired, you may also break down sub-categories, such as *Mistranslation*, *Omission*, or *Addition*, within *Accuracy*.

To compare systems, visualise their error profiles as radar (spider) charts. This will allow you to easily spot systems with :

— a high proportion of accuracy errors (e.g. mistranslations),
— relatively fewer style or fluency errors,
— distinctive patterns across error categories.

This visual step is crucial for understanding how different systems behave linguistically, beyond overall scores.

Clustering Systems by Error Profiles Use the error frequency vectors to cluster systems. You may apply : either K-means clustering or Hierarchical clustering. To visualise the clusters, consider dimensionality reduction (e.g. PCA or t-SNE). What do the resulting clusters reveal about similarities and differences between systems ? Do systems from the same organisation or with similar architectures exhibit similar error patterns ?

MQM and Automatic Metrics Aggregate MQM scores per system by applying the weighting scheme used in Freitag et al. (2021), e.g. Minor = 1, Major = 5. Then compute Pearson and Spearman correlations between MQM and BLEU, chrF, COMET. Do this at both system level and segment level, and visualise the relationships with scatter plots. Which automatic metrics align most closely with MQM scores ? Does COMET outperform BLEU in terms of correlation with human judgements ?

Linguistic Factors and Correlation To explore how surface and linguistic properties affect metric alignment :

— Compute basic segment-level features : length (in tokens or characters), type–token ratio, optional neural embeddings.
— Cluster sentences based on these features using TF–IDF or sentence embeddings.
— Recompute MQM–metric correlations within each cluster.

Does correlation differ between short and long segments ? Do metrics align less well with human judgements for more linguistically complex sentences ?

5 Reflection and Interpretation

In your report, address the following :

— What can error profiles reveal that overall MQM or BLEU scores cannot ?
— Why might clustering systems by error type be useful for system development ?
— Which types of errors appear most closely associated with low MQM scores ?
— How stable is the alignment between automatic metrics and human evaluation across different sentence types ?

6 References

— Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context : A Large-Scale Study of Human Evaluation for Machine Translation. *TACL*.

— Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics.

— Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013, 2015). Direct Assessment of Machine Translation.

— Bojar, O. et al. Findings of the WMT Shared Tasks.