Ancient to Modern Italian Automatic Translation

Alemanno Giulia, Lanzellotti Flaminia
Correspondence: alemanno.1953800@studenti.uniroma1.it, lanzellotti.2201485@studenti.uniroma1.it

1 Introduction

Translating a text is not an easy task, it requires a deep understanding not only of the language and cultural context in which the text was written in, but also a meticulous research aimed to find the best suited terms that represent the subtle nuances of meaning that can be easily lost in translation. How can an LLM overcome lexical, syntactic, and semantic challenges intrinsic in historical texts? This paper is going to try to provide a more accessible option to enhance the value of literary heritage, experimenting and exploring different techniques and models available. The translations obtained were judged by an automatic scoring system enforced by an LLM and later on compared with the opinion of an expert in the field, that allowed a much deeper analysis on the accuracy metrics and fluidity of the text generated.

2 Methodology

Modern LLMs trained for text translation do not consider non-modern Italian as a language on its own so, due to the specificity of the task, the main issue was to find an alternative way to simple Seq2Seq Machine Translation models that require specific input and output language pair. In order to tackle this task accurately the initial hypothesis was to compare different LLMs' translations with different prompting techniques. The core idea was to give the same prompt to different decoder-only models and compare their respective output while, at the same time, testing the output's sensibility to different prompting styles.
In an attempt to approach the problem in the most comprehensive way an additional strategy was implemented: finetuning a Seq2Seq Multilingual Machine Translation model. The idea behind this strategy was to compare the translations obtained from the finetuned model with the LLMs translations to try and understand if there was a significant difference between the two approaches.
Finetuning a model means training further a pretrained model on a task-specific dataset to improve its overall performance. This meant that it was necessary to define a new dataset containing original non-modern Italian sentences and their respective translations.
The translations were then judged by two different LLMs chosen distinctively from the ones considered earlier. The score appointed by the LLMs was based on a grading system from 1 to 5 penalizing mainly semantic and syntactic errors. The highest graded translation set was then also graded by a human and the final scores were compared using the Cohen's kappa coefficient to assess the overall agreement.
The two main approaches were then combined by taking advantage the highest graded LLM to revise the translations obtained by the finetuned model. This resulted in a new and improved set of translations that was then judged by one LLM. A subset of this dataset was also judged by a human and an expert in the field: a newly Master's graduate from the faculty of *"Filologia Moderna (LM-14)"* at "La Sapienza".

3 Experiments

The LLMs used for the text translation were **Qwen3-14B** (qwe) and **Meta-Llama-3-8B-Instruct** (lla) both open-source and available in the Hugging Face library. The two prompts used were engineered with two different prompting techniques: zero-shot prompting and role prompting with context.
The first one contained context on the year in which the sentences were written in and it was as concise as possible, it reads as follows: *"Translate the following sentence from non-modern Italian, written between 1250 and 1350 into modern Italian"*.

The second one gave a specific role to the LLM and added more context to the task which resulted in a much more complex and detailed prompt, it reads as follows: *"You are an expert at translating text from non-modern Italian to modern Italian. Non-modern Italian is a form of the Italian language that was directly derived from ancient Latin. Provide the translation of the following sentence written between 1250 and 1350, specifically the Tuscan and Umbrian variety, into modern Italian"*.
The results from this experiment proved that a much simpler prompt delivered more accurate results.
For the second approach a new gold annotated dataset was introduced. This dataset was created analyzing sentences from the most prominent authors between the years 1250–1350: Dante Alighieri, Giovanni Boccaccio, Francesco Petrarca, Guido Cavalcanti, Guido Guinizzelli, and San Francesco d'Assisi. The dataset contained the original sentence and their official translation in modern Italian.
This dataset was later tokenized and used to finetune an already pretrained open-source model: **facebook/mbart-large-50-many-to-many-mmt** (mba), available in the Hugging Face library. The key parameters set for training are as follows: a batch size of 8 for both training and evaluation, a weight decay of 0.01, a learning rate of 2e-5, and 5 training epochs. The training logs recorded information at each step such as training and validation loss and the rouge metrics. As shown in the Table 1 the validation and training loss is decreasing consistently, indicating that the model is generalizing well, and the rouge metrics' values show meaningful improvements across the epochs. All the translations generated by the different models were saved in JSONL files, one for each model.
Regarding the combined approach the translations generated by the finetuned model were then corrected by **Qwen3-14B** using a few-shot prompting technique, incorporating one example of an acceptable correction, it reads as follows: *"Revise the translation of the following sentence in old Italian maintaining as much as possible relevance to the original sentence's meaning. Paraphrase the translation using modern italian terms, an example of acceptable correction is: Old Italian: Marco Cornelio ch'era de' dieci compagni, studiosamente si riservò di parlare all'ultimo. Translation: Marco Cornelio che era dei dieci compagni, si riservò di parlare all'ultimo con studio. Corrected Translation: Marco Cornelio, che era tra i dieci compagni, scelse di parlare per ultimo intenzionalmente."*.
The LLMs used to judge the outputs were: **Gemini 2.0 Flash** (gem), an LLM provided by Google available with requests limits, and **Prometheus-7b-v2** (pro), an open-source LLM available in the Hugging Face library.
The marks appointed to the translations were achieved through a specific evaluation grid, consistent between the LLMs, with scores from 1 to 5. In **Gemini**'s case the grid needed to be included inside the prompt, meanwhile for **Prometheus** it was set up as a rubric. All the scores generated by the different judges were saved in JSONL files, one for each combination of model and judge. The scores from the two judges varied greatly, with **Prometheus** resulting in much harsher marks.

4 Results

The Table 2 shows the average score of each model's translation for both **Gemini** and **Prometheus**.
Gemini had an overall average of 3.93 while **Prometheus** had an overall average of 2.71.
However, the translations obtained with the LLM correction were only judged by **Gemini 2.0-flash**. The highest scoring model was **Qwen-14B** with the first prompt for this reason it was chosen for the human metrics correlation part of the task. As shown in the confusion matrix in Figure 1 the human scores compared with **Gemini**'s score were much lower with a perfect correspondence only on 18 samples, the Cohen's kappa coefficient was 0.024, indicating a slight agreement. For 74% of the samples **Gemini** assigned a perfect score meanwhile only 17% of the samples marked by the human obtained the highest score. As for **Prometheus** the Cohen's kappa coefficient only improved slightly with a value of 0.051, with a confusion matrix shown in Figure 2. **Prometheus**' scores were more diversified resulting in a much coherent assignment of the marks which was more in agreement with the human grading approach.
The combined approach's subset evaluated resulted in an average of 4.11 and, as seen in Table 3, in a Cohen's kappa coefficient of 0.5 evaluated between the scores of the LLM and the human scores, and a Cohen's kappa coefficient of 0.396 between the LLM scores and the expert's scores.

5 Appendix

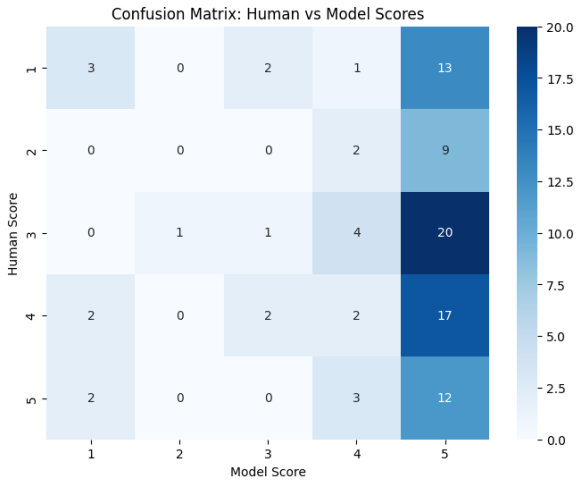

Figure 1: Confusion Matrix for Human-Gemini correlation (Qwen first prompt)

Figure 2: Confusion Matrix for Human-Prometheus correlation (Qwen first prompt)

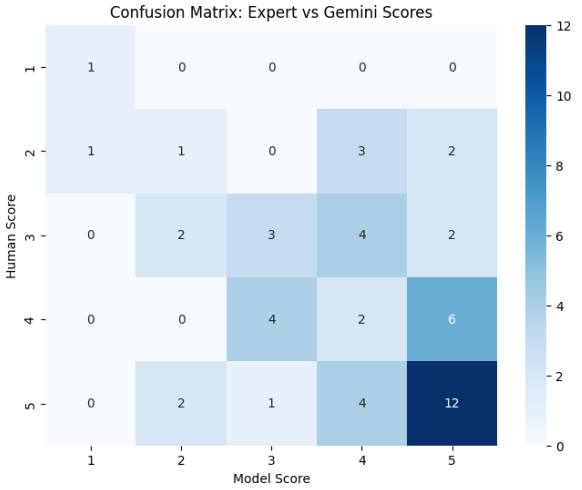

Figure 3: Confusion Matrix for Expert-Gemini correlation

Human-LLM	Expert-LLM	Human-Expert
0.504	0.396	0.444

Table 3: Cohen's Kappa Coefficient

Epochs	Training Loss	Validation Loss	Rouge-1	Rouge-2	Rouge-L	Rouge-L-sum
1	11.380	9.315	0.357	0.123	0.312	0.310
2	8.551	7.812	0.393	0.151	0.343	0.341
3	7.719	6.623	0.383	0.140	0.341	0.339
4	6.382	5.833	0.394	0.154	0.348	0.346
5	5.695	5.486	0.396	0.148	0.347	0.345

Table 1: Performance metrics of the finetuned model.

Judge	Finetuning	Llama	Llama 2nd prompt	Qwen	Qwen 2nd prompt
Gemini	2.896	4.195	4.010	4.453	4.123
Prometheus	2.628	2.855	2.731	2.793	2.608

Table 2: Average scores for each model.

References

Gemini-2.0-flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash?hl=it.

Mbart-large-50-many-to-many-mmt. https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt.

Meta-llama-3-8b-instruct. https://huggingface.co/meta-llama/Meta-Llama-3-8B.

Prometheus-7b-v2. https://huggingface.co/prometheus-eval/prometheus-7b-v2.0/blob/main/README.md.

Qwen3-14b. https://huggingface.co/Qwen/Qwen3-14B.

Dante Alighieri. 1306-1321. *Divina Commedia, Inferno, Canto 1*. https://divinacommedia.weebly.com/.

Giovanni Boccaccio. 1349-1351. *Decameron*. https://moodle2.units.it/pluginfile.php/333230/mod_resource/content/1/BOCCACCIO-DECAMERON.pdf, https://cmathilde.altervista.org/Decameron/Prima/Prima.htm.

Guido Cavalcanti. 1259-1300. *Donna Me Prega*. https://www.roberto-crosio.net/1_AMORE_MEDIOEVO/t_cav_donna.htm.

San Francesco d'Assisi. 1226. *Cantico delle Creature*. https://it.wikipedia.org/wiki/Cantico_delle_creature).

Guido Guinizzelli. 1250-1300. *Io voglio del ver la mia donna laudare*. https://it.wikipedia.org/wiki/Io_voglio_del_ver_la_mia_donna_laudare).

Francesco Petrarca. 1336-1374. *Canzoniere*. https://it.wikisource.org/wiki/Canzoniere_(Rerum_vulgarium_fragmenta), http://petrarca.letteraturaoperaomnia.org/parafrasi/petrarca_parafrasi_canzoniere.html.

(Alighieri, 1306-1321) (Boccaccio, 1349-1351) (Petrarca, 1336-1374) (Cavalcanti, 1259-1300) (Guinizzelli, 1250-1300) (d'Assisi, 1226)