Second MNLP Homework

Mattia Di Marco 2019367
Simone La Bella 1995847
Muhammad Ramish Khan 2171618

1 Introduction

In this project, we decide to work on the first task: **"Ancient to Modern Italian Automatic Translation"**. In addition, we gained access to Cineca and we get our hands a little dirty, learning how to use it and how capable it is.
We used different LLMs (Llama, Zephyr, Prometheus) and a transformer model for the task of generating translations and judging the translated performance.
For the translations, we used Llama and Zephyr with different in context learning configurations. In addition, we fine-tuned the llama model. We also trained a transformer model NLLB (No Language Left Behind) for this task.
As for the task of judging the models, we used Prometheus.

2 Methodology

2.1 LLM without In-Context Learning

We first approach the task using Zephyr-7B-β as LLM, because we wanted to test a different LLM than the one suggested. It is a fine-tuned version of Mistral-7B-v0.1 that was trained on a mix of publicly available synthetic data sets using direct preference optimization (DPO).

2.2 LLM with In-Context Learning

Secondly, we tried to understand how in-context learning could influence the translation task. In particular, we noticed that, in the dataset, most of the sentences were written in Florentine or Tuscan, so it was a quite obvious choice to use phrases from the "Divine Comedy" of Dante Alighieri. This choice was also made because of the ease of finding a correct translation in modern Italian. We have therefore chosen a total of seven sentences between Hell, Purgatory, and Paradise, this to exploit also the character of the change of linguistic register that Dante implements in the three cantici. With these seven sentences and their translations [B], and using as LLM "Llama-3.2-3B-Instruct", we performed five different translations, respectively with zero (No In-Context), one, three, five and seven shots (shot is the number of sentence-translation pairs passed into input to the model together with the sentence to be translated).

2.3 LLM with Fine Tuning

In addition to in-context learning, we explored fine-tuning using the llama 3.2B instruct. The finetuning was performed using a Peft framework which allows the adaptation of llms using LoRa. We save the LoRa weights locally, to merge it with the base model for later inference.

2.4 Transformer-based Machine Translation Systems

For this approach we decided to use the NLLB model, in particular we used the version "NLLB-200-Distilled-1.3B". However, among the languages supported by this model, there is obviously no archaic Italian, so it was necessary to use "ita_Latn" as the language of both the original phrase and the translated phrase. This obviously means that the translations will generally be of lower quality, not so much for the translations of individual words, which, in most cases are correctly translated into the modern version, the more for the structure of the sentence that remains unchanged and therefore, often, does not respect those which are the standards of the modern Italian language.

3 Experiments

So we did all the translation for the entire dataset, using various approaches and models [2]. In particular, regarding the LLM models we used a specific prompt [A.1].

For fine-tuning the llama model, we used a custom data set consisting of our translated sentences [B] and some additional sentences.

4 Results

4.1 LLM-as-a-Judge

For translation evaluation, we used the LLM-as-a-Judge method, using Prometheus. Using it, was a little more complex, in fact it is an LLM specifically designed to work as a "judge". Therefore, it needs not only the instruction, the original sentence and the translated one but also the gold reference, in our case the correct translation of the sentence. It is possible to bypass this by downloading directly from HuggingFace and using a prompt formatted in a particular way. This means that, of course, the quality of the judgments will be lower than having gold references, but in the absence of them, you can still use Prometheus. The full prompt given to the judge can be read at Sec. [A.2] and the rubrics are described in Sec[A.3].

4.2 Score correlation

To generate correlations, we decided to score the first 20 sentences from each of the translations generated by our models (llms and transformer). We used the same scoring rubrics [A.3]. We calculate the correlation between the our scores and ones generated by Prometheus, using the *Spearman's rank correlation coefficient* [Tab. 1].

5 Appendix

References

A Prompt

A.1 Translation

"You are an expert translator specializing in Medieval Italian texts (13th-15th century). Translate the following text from Archaic Italian to Modern Italian following these precise rules:
PRESERVE the original punctuation exactly as written,
MAINTAIN the original sentence structure and word order,
UPDATE only archaic vocabulary and grammatical forms to modern equivalents,
DO NOT paraphrase or interpret - translate literally,
KEEP the same level of formality and register,
Respond with ONLY the translated text, no explanations or comments."

A.2 Judging

prompt = "Task Description:
An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing a evaluation criteria are given.
Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
The output format should look as follows: "Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 5)" Please do not generate any other opening, closing, and explanations.

The instruction to evaluate: instruction

Response to evaluate: response

Score Rubrics: rubric

A.3 Rubric data

rubric_data = { "criteria":"Is the model the converts old and archaic italian to Modern Italian",
"score1_description":"Fails to convey the meaning: gibberish or irrelevant output",
"score2_description":"Significant loss of meaning: Major errors in translation or grammar, Misleading or confusing phrasing",
"score3_description":"Some meaning preserved: Modern grammar/vocabulary partly correct, One or more semantic inaccuracies, Unnatural sentence structure",
"score4_description":"Core meaning preserved: Mostly accurate and fluent modernization with minor errors, May sound a bit unnatural but still comprehensible",
"score5_description":"Preserves all core ideas and intent: The modernized version is fluent, fully faithful to the original meaning, and idiomatically natural in modern Italian" }

B Sentences used for Learning

Sentences with their translation:

1. *Amor, ch'a nullo amato amar perdona, mi prese del costui piacer sì forte, che, come vedi, ancor non m'abbandona.*
   Amore, che non tollera che chi è amato non ami a sua volta, mi rapì della bellezza di questi [Paolo] in modo così potente, che, come vedi, ancora lo amo.

2. *Considerate la vostra semenza: fatti non foste a viver come bruti, ma per seguir virtute e canoscenza.*
   Prendete coscienza della vostra condizione [di uomini]: non foste creati per vivere come selvaggi, ma per accrescere [le vostre] virtù e [il vostro] sapere.

3. *Dette mi fuor di mia vita futura parole gravi, avvegna ch'io mi senta ben tetragono ai colpi di ventura.*
   A proposito della mia vita futura mi sono state rivolte parole dure e preoccupanti, sebbene io mi senta ben stabile davanti ai colpi della sorte.

4. *A l'alta fantasia qui mancò possa; ma già volgeva il mio disio e 'l velle, sì come rota*

ch'igualmente è mossa, l'amor che move il sole e l'altre stelle.
All'immaginazione ora mancò la capacità, ma già il mio desiderio ed il volere erano soddisfatti, come una ruota che si muove di moto uniforme, dall'amor che muove il sole e le altre stelle.

5. *Ahi serva Italia, di dolore ostello, nave sanza nocchiere in gran tempesta, non donna di provincie, ma bordello.*
Ahi, Italia, schiava, albergo di dolore, nave senza guida in una tempesta, non donna rispettabile, ma prostituta!

6. *Sovra candido vel cinta d'uliva donna m'apparve, sotto verde manto vestita di color di fiamma viva.*
Posta su un candido velo coronata d'ulivo mi apparve una donna, coperta da un manto verde con un vestito di color rosso vivo.

7. *Era già l'ora che volge il disio ai navicanti e 'ntenerisce il core lo dì c' han detto ai dolci amici addio.*
Ormai si era fatta quell'ora che fa sì che, a coloro che navigano, in quel giorno che hanno detto addio ai cari amici, il desiderio si volga indietro e il cuore si intenerisca.

C Judging Score

C.1 Human scores

Zephyr = [2, 2, 1, 3, 3, 1, 2, 4, 3, 2, 2, 2, 1, 1, 1, 1, 3, 3, 2, 2]
Llama_0Shot = [1, 1, 1, 3, 1, 1, 2, 3, 1, 1, 2, 2, 1, 2, 1, 1, 2, 2, 2, 2]
Llama_1Shot = [3, 2, 1, 2, 3, 3, 2, 4, 3, 4, 1, 1, 2, 2, 4, 2, 1, 2, 2, 3]
Llama_3Shot = [2, 3, 5, 2, 3, 4, 2, 2, 4, 5, 3, 5, 4, 1, 4, 3, 1, 1, 3, 3]
Llama_5Shot = [4, 4, 2, 3, 3, 4, 4, 3, 3, 1, 4, 3, 5, 2, 3, 3, 2, 2, 4, 3]
Llama_7Shot = [3, 1, 2, 3, 3, 4, 1, 1, 4, 3, 5, 3, 5, 3, 4, 3, 2, 3, 3, 4]
NLLB = [2, 1, 2, 2, 2, 2, 1, 3, 1, 1, 3, 4, 3, 2, 2, 3, 5, 1, 2, 2]
Llama_FT = [1, 2, 2, 2, 2, 2, 2, 3, 2, 2, 1, 2, 3, 2, 2, 2, 3, 1, 1, 2]

C.2 Prometheus scores

Zephyr = [2, 5, 3, 3, 4, 2, 4, 4, 2, 3, 3, 2, 1, 2, 1, 1, 5, 3, 1, 3]
Llama_0Shot = [1, 1, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 5, 4, 1, 1, 5, 3, 1, 3]
Llama_1Shot = [2, 3, 1, 1, 4, 1, 1, 5, 2, 3, 1, 1, 1, 1, 2, 1, 5, 3, 2, 4]
Llama_3Shot = [2, 2, 5, 1, 4, 3, 3, 2, 3, 2, 4, 2, 5, 1, 3, 5, 1, 1, 3, 3]
Llama_5Shot = [4, 5, 1, 1, 4, 3, 3, 3, 3, 1, 5, 1, 3, 1, 1, 2, 5, 3, 1, 4]
Llama_7Shot = [3, 1, 1, 1, 2, 3, 1, 1, 5, 2, 4, 1, 5, 1, 2, 1, 1, 3, 2, 2]
NLLB = [2, 1, 1, 2, 2, 2, 2, 4, 1, 1, 1, 2, 1, 2, 2, 1, 5, 1, 1, 2]
Llama_FT = [1, 2, 1, 2, 4, 1, 1, 2, 1, 1, 1, 1, 5, 2, 1, 2, 5, 2, 1, 1]

D Tables

Table 1: Spearman's rank correlation

Models	Rho	P_Value
Zephir	0.6020	0.0050
LLama-0shot	0.4172	0.0673
LLama-1shot	0.4414	0.0514
LLama-3shot	0.5607	0.0101
LLama-5shot	0.3499	0.1304
LLama-7shot	0.7785	0.00005
NLLB	0.3598	0.1192
Llama-finetuning	0.5078	0.0223