LLM-based Translation for Latin: Summaries Improve Machine Translation

Dominic P. Fischer, Martin Volk
University of Zurich
Department of Computational Linguistics
dominicphilipp.fischer@uzh.ch

Abstract

Recent studies demonstrated that modern Large Language Models set a new state-of-the-art in translating historical Latin texts into English and German. Building upon this foundation, we investigate the impact of incorporating text summaries into prompts for LLM-based translation tasks. Having both the historical text and a modern-language summary is a typical setup for classical editions. Our findings reveal that integrating summaries significantly enhances translation accuracy and coherence.

Keywords: Large Language Models, Machine Translation, Summaries, Latin, Historical Letters

1 Introduction

Summarizing is an essential task for editors of historical texts. Editors create summaries in modern languages to distill the complex and extensive information found in historical documents, ensuring that the core message and significant details are preserved. This editorial practice not only aids in efficient information retrieval but also maintains the integrity and context of historical records.

Historical texts are therefore often accompanied by summaries. This provides a valuable opportunity to exploit the modern language summary when processing the historical text with an AI system. We suspect that LLMs may profit from the expert-distilled information in the summaries.

This paper explores the impact of manual summaries in LLM-based machine translation for Latin to English and German translation of 16th century letters, following up on (Volk et al., 2024a). We hypothesize that providing LLMs with well-crafted summaries will enrich the translation process, yielding superior quality text in the target language. The translation of the full text provides the complete rendering of the original content and thus allows for a more comprehensive analysis than only relying on the summaries.

Pairing the letter text with its summary as input to machine translation not only highlights the practical applications of LLMs in historical research but also underscores the value of editorial practices in the digital age. By combining the strengths of human expertise and advanced AI, we push the boundaries of what can be achieved in the translation of historical texts.

Our research is in the line of research on prompting strategies for LLM-based machine translation (Zhang et al., 2023; He et al., 2024) which focuses on the impact of providing translation examples. We are the first to test the impact of providing a target language summary together with the source text.

Adding the summary is a form of knowledge injection through the prompt. Similar to the integration of domain-specific terminology to a prompt (as in Bogoychev and Chen (2023)), and similar to adding translation suggestions from lexical footnotes (Volk et al., 2024b) or from bilingual dictionaries to the prompt (Ghazvininejad et al., 2023). The latter show that LLM prompting provides an effective solution for rare word translation, by using knowledge from bilingual dictionaries. Yao et al. (2024) introduce various strategies to incorporate external and internal cultural knowledge into the prompt. Strategies include self-explanation and self-ranking to activate the relevant knowledge of the LLM.

2 Corpus and Methodology

For our evaluation, we utilized the test set from (Fischer et al., 2022), which includes eight Latin letters manually translated into German by a domain expert. This test set comprises 121 Latin sentences, ranging from short greetings to sentences as long as 47 words, totaling 1240 words in Latin and 1768 words in the corresponding German translations. To adapt this test set for translation into English, we

used GPT-4 to automatically translate the German texts into English.

The letters are taken from the 16th century letter exchange of the Zurich Reformer Heinrich Bullinger. 3200 of the letters have been published in 21 printed volumes over the last 50 years by the Institute for Swiss Reformation Studies[1], professionally edited, summarized and extensively commented.[2] Depending on the volume, the summary length and format varies. The German summaries are as short as a few sentences in the first volumes (published in the 1970s) and then increase in length to being more comprehensive.

For example, the letter from Berchtold Haller to Heinrich Bullinger (February 1532, published in volume 2; not part of the test set) consists of 609 tokens in Latin (5 lengthy paragraphs plus initial greetings and letter closing). But the editors summarized it with only one paragraph (68 tokens) in German:

- *Berichtet von der Lage nach der Berner Synode, deren Akten bald im Druck erscheinen werden und worüber er Bullingers Meinung erfahren möchte. Bittet um Antwort auf verschiedene Fragen, um die Zusendung von Bullingers und Pellikans Werken, macht Vorschläge für eine Annäherung zwischen Bern und Zürich und betont, daß Zwinglis Sohn Wilhelm in Bern unter den besten Voraussetzungen erzogen wird. Grüße.*
  (Reports on the situation after the Bern Synod, whose records will soon be published, and wishes to hear Bullinger's opinion on the matter. Requests answers to various questions, the sending of works by Bullinger and Pellikan, and makes suggestions for a rapprochement between Bern and Zurich. Emphasizes that Zwingli's son, Wilhelm, is being educated in Bern under the best conditions. Sends greetings.)

Starting from volume 16 (published in 2014), the summaries are done paragraph by paragraph, covering the entire letter. These summaries can be seen as shortened German paraphrases of the letter. Still, as from the first volume of the edition, the summaries are written as a description of the letters ("The author X reports on the situation after the Bern synod, ...") in contrast to the letters themselves that are written from a personal perspective ("After the synod was concluded, ...").

The eight letters in our test set are taken from volumes 14, 15 and 16; three of them have paragraph-wise summaries. The summary lengths range from 54 to 428 tokens with the ratios of summary length to letter length ranging from 0.43 to 0.96 (cf. Table 5).

For the LLM-based translation of the test set without and with the summary, we employed the following two prompts:

- Without summary: Translate the following Latin text into German/English while keeping the formatting as it is: *Latin text*.
- With summary: Translate the following Latin text into German/English: *Latin text*. Keep the formatting as it is. As a help for your translation, consult this summary: *summary*.

Additionally, we tested whether GPT-4 performs better at translating a letter when it is aligned with the sentences of the corresponding summary. For this purpose, we manually aligned the sentences of the summary with the letters, inserting them in brackets after the sentence they refer to, like in Table 1.

[...] Nihil certi ex comitiis audio. Expectatur adhuc Ferdinandus rex. **(The Reichstag [in Speyer] is still waiting for King Ferdinand.)** [...]

Table 1: Letter with aligned summary sentence (Johannes Gast to Heinrich Bullinger on 1.4.1544)

Automatic alignment with GPT-4 provided results with only minor discrepancies with regards to the human alignment, indicating a promising avenue for automatic text-summary alignment. For the purpose of the experiment, however, we used human alignment to avoid inducing any errors.

For the translation with aligned summaries, another two prompts - one as short as possible, one more descriptive - were used:

- Translate the following Latin sentences into German/English. Use the sentences in brackets to guide your translation. Preserve the formatting: *Latin text with aligned summary*
- Translate the following Latin letter into German/English. The lines in brackets are from a

[1] https://www.irg.uzh.ch/

[2] The complete preserved Bullinger correspondence consists of 12,000 letters.

summary of the letter and have been aligned, so that they explain the preceding lines. Take them into account, but do not output them in your translation. Keep the line breaks as they are: *Latin text with aligned summary*

3 Main Findings

The analysis of translation quality revealed notable improvements when summaries were included, as indicated by both BLEU (SacreBLEU) and ChrF scores[3]. However, this only applies if the summary is in the same language as the target text, as is illustrated in Tables 2 and 3. When translating the test set into English, the BLEU score increased only marginally from 32.1 to 32.5 when given the German summary, but increased significantly by 2.3 points to 34.4 with an English summary (which we automatically translated from German). When translating into German, the increase in BLEU is 2.0 points when given the German summary.

Table 2: Translation Quality Scores (BLEU) on the test set with and without summaries.

Testset	No Summary	W/ Summary
DE	25.8	**27.8** (DE)
EN (GPT-4)	32.1	32.5 (DE)
		34.4 (EN)

Table 3: Translation Quality Scores (ChrF) on the test set with and without summaries.

Testset	No Summary	W/ Summary
DE	51.6	**53.3** (DE)
EN (GPT-4)	52.6	53.4 (DE)
		54.7 (EN)

While these BLEU score increases of 2.3 for English and 2.0 for German were similar, the absolute values of the BLEU scores are higher for translations into English. We will discuss the reasons for this in section 4. With regards to the ChrF scores, we see the same trend - an increase of about 2 points when summaries are included, yet interestingly, the difference in absolute values between English and German is only marginal (cf. Table 3).

[3]BLEU evaluates translation quality by measuring the overlap of sequences of n words (so-called n-grams) between the machine-generated and a reference translation, while ChrF uses overlapping *character* n-grams.

The experiments with the aligned summaries showed interesting results. With the simple prompt, the results were the same or only slightly better (~1 BLEU/ChrF point) than the translation *without* summary.

The longer, more descriptive prompt yielded different results in German and English. In German, the results were worse than with the simple prompt, with almost the same scores as without summary. For English, this resulted in the best translation yet, surpassing the translation scores with target language summary by 0.9 BLEU points and 0.7 ChrF scores (cf. Table 4). Nevertheless, this approach did not yield consistent improvements, as illustrated by the wrong translation in the last row of Table 6.

Table 4: BLEU and ChrF scores for translation with aligned summary in the target language.

	With Aligned Summary	
	BLEU	**chrF**
P1: DE	**26.8**	**52.2**
P2: DE	25.7	51.8
P1: EN	32.4	53.2
P2: EN	**35.3**	**55.4**

Table 5 shows that longer summaries, or summaries that cover more of a given letter do not necessarily lead to greater improvements in translation. At the same time, even short summaries (as short as a single sentence) can lead to significant quality increases. It therefore stands to reason that situating the letter, outlining its content and the actors therein is an efficient way of injecting crucial information for translation quality gains.

letter id	#tok.L	#tok.S	ratio	ΔBLEU
12151	244	105	0.43	4.3
11916	180	96	0.53	5.31
11898	98	54	0.55	3.33
12838	98	54	0.55	1.31
11930	179	109	0.61	-0.39
12378	106	67	0.63	0.02
12154	172	157	0.91	2.62
12509	444	428	0.96	0.11

Table 5: Comparison of letter (L) and summary (S) token counts, ratio, and BLEU improvement measured between without and with summary. The entries are ordered ascendingly by ratio. (11898 and 12838 happen to have the same counts for summaries and letter texts.)

Latin original	Commissum habeas adolescentulum; polliceor et ego me non ingratum fore.
EN reference	I recommend the young boy to you; I assure you that I too will not be ungrateful.
EN without summary	You may have committed the young man; I also promise that I will not be ungrateful.
EN with summary in DE	You have the young man in your care; I promise that I will not be ungrateful.
EN with summary in EN	**You may have the young man in your care**; I promise that I will not be ungrateful.
EN with aligned summary in EN	You have a committed young man; I promise that I will not be ungrateful.

Table 6: Translations of the Latin sentence without summary and with summary in German and English. The Latin conjunctive 'commissum habeas' only gets correctly translated with the English summary in the prompt to 'You may have'.

This is supported by our qualitative analysis of the summary-induced effects in the German translations (cf. Table 7). To that end, we manually compared the 121 test set sentences when translated with and without summary. It results that 51 stayed the exact same, while 70 contained changes. Out of these 70, 36 contained minor neutral (word choice) changes, 23 minor positive changes, and only 4 minor negative changes. 7 sentences contained major positive changes, including changes crucial to the correct understanding of the sentence and major changes in the sentence structure.

	amount	**percentage**
the same	51	42
different	70	58
... *of which*		
neut. (≈)	36	30
pos. (+)	**23**	**19**
neg. (-)	4	3
major pos. (++)	**7**	**6**

Table 7: Overview of changes induced by including the summaries in the prompt.

Classified as "minor" were changes of often one, sometimes multiple (max. 3) words. Minor positive changes contained predominantly name corrections/normalizations (*Marcus* → *Markus*, *Caesar* → *Kaiser Karl V.*), clarifications of pronouns (*these* → *these news*, *he* → *it*), and previously missed precisions (*an answer* → *any answer*). Negatives included wrongful reversals of such things, like *these questions* → *these*, *pray to the Lord* → *pray*. The major positive changes greatly affected the understanding of the sentence, including changes of modus (imperative → conjunctive) or of an entire (part of a) sentence, such as in table 6. Major negative or neutral changes were not present.

4 Discussion

The observed improvements in translation quality with the inclusion of summaries can be attributed to the additional context provided by the summaries. This context helps the LLMs generate more accurate and coherent translations by offering clear guidance on the essential points and context of the text.

The better performance of English translations with regards to BLEU could be linked to two main factors. Firstly, the gold standard translation of the letters in English are a GPT-4 translation of the German gold standard, which might have introduced a bias towards higher scores due to the model's own translation capabilities. This could mean that the English summaries were inherently more aligned with the model's strengths.

Secondly, GPT-4 and similar LLMs are extensively trained on English language texts, leading to inherently better performance in English due to the abundance of training data and resources. This extensive training allows the model to produce English text with higher accuracy and fluency, as has been observed in other studies.

The first above point implies that the quality of the English translation is not actually significantly better than the German translation, it merely appears to be because of the skewed English transla-

tion. As the ChrF scores are very close between translations into German and English, ChrF seems to balance this skewness.

A reason might be ChrF's indifference to the structural differences of the two languages. For example, German has a more flexible word order and often requires reordering phrases in translation to maintain grammatical correctness. This can result in lower n-gram overlap in BLEU scores because SacreBLEU heavily relies on exact matches of words and phrases. Similarly, the morphological complexity of German works against the exact matching of word n-grams that BLEU measures, and is better suited to character-level comparisons like ChrF.

In other words: BLEU amplifies the skewness, since it looks for exact matches of n-grams, which are more likely to be present if the reference itself is a product of GPT-translation.

Our findings suggest that including a sentence-aligned summary in the prompt for translation does not lead to significant improvements in the translation quality over feeding the summary as a block of text. While the fleshed-out prompt did lead to the best results for English, the improvement compared to the inclusion of the unaligned summary is not high enough to be significant. Furthermore, the same prompt did not lead to increased, but to clearly decreased translation quality in German, as the translation with aligned summary gets basically the same scores as translation without any summary at all.

5 Conclusion

Incorporating human-made summaries into LLM-based translation tasks significantly enhances translation quality, when the summary language and the target language are equal. This is evidenced by the improved BLEU and ChrF scores of 2+ points when summaries are included in the prompt. Splitting the summaries into sentences and aligning them with the sentences in the letter does not lead to significant improvements and is highly dependent on the prompt and the language. These findings underscore the usefulness of language-specific summaries in improving LLM performance for the translation of historical texts.

This study invites many avenues for further investigation. A baseline experiment could be to regard the summaries as translations and to measure their BLEU scores. For short summaries that are less than half the length of the letter texts, this will inevitably lead to low scores. But for the longer summaries, this might give an interesting lower bound.

Another experiment to investigate the impact of the summary would involve the use of some arbitrary text instead of the summary. This will help us understand the impact of the summary in the automatic translation.

In future work we will also test whether the addition of summaries helps in translating from Early New High German to modern languages, as a follow-up of the work in (Volk et al., 2024b).

Another option is the combination of two LLMs, one that produces a summary (or a draft translation) for the letter in the target language, and another LLM that uses the summary in combination with the letter text for the translation.

Acknowledgments

We would like to thank the two anonymous reviewers for insightful comments. This research is part of the project "Bullinger Digital" funded by the UZH Foundation.

References

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 890–896, Singapore. Association for Computational Linguistics.

Lukas Fischer, Patricia Scheurer, Raphael Schwitter, and Martin Volk. 2022. Machine translation of 16th century letters from Latin to German. In *Proceedings of 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) at LREC-2022*, pages 43–50, Marseille.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arxiv.org/abs/2302.07856*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Martin Volk, Dominic P. Fischer, Lukas Fischer, Patricia Scheurer, and Phillip B. Ströbel. 2024a. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language*

Technologies for Historical and Ancient Languages – LT4HALA (at LREC/COLING), Torino.

Martin Volk, Dominic P. Fischer, Patricia Scheurer, Raphael Schwitter, and Phillip B. Ströbel. 2024b. LLM-based translation across 500 years. The case for Early New High German. In *Konferenz zur Verarbeitung natürlicher Sprache 2024 (KONVENS)*, Wien.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking LLM-based machine translation on cultural awareness. *arXiv:2305.14328v2*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A case study. In *Proceedings of the 40 th International Conference on Machine Learning*, Honolulu.