Evaluating the Use of Generative LLMs for Intralingual Diachronic Translation of Middle-Polish Texts into Contemporary Polish

Cezary Klamra[1[0000−0003−4321−8862]],
Katarzyna Kryńska[2[0000−0003−3446−122X]], and
Maciej Ogrodniczuk[1[0000−0002−3467−9424]]

[1] Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
maciej.ogrodniczuk@ipipan.waw.pl

[2] Institute of Polish Language, Polish Academy of Sciences
al. Mickiewicza 31, 31-120 Kraków, Poland
katarzyna.krynska@ijp.pan.pl

Abstract. This paper presents efforts towards creating a tool for translating texts from Middle Polish into modern Polish. Archaic texts sourced from the CBDU digital library were translated into modern language using ChatGPT and the resulting parallel corpus was used to train a neural text-to-text model. We assessed the results using automatic metrics and performed human evaluation of translations of the best-performing model and ChatGPT. Even though the performance of the trained models was far from perfect, the quality of translations produced with ChatGPT was good in most cases. Although caution should be exercised, we believe that LLMs have a high potential for text-to-text annotation applications.

Keywords: Intralingual diachronic translation · Automatic annotation · Large Language Models

1 Introduction

Numerous archaic Polish texts have been made available as part of historical corpora or digital libraries [7,14,6]. However, comprehension of these texts may be challenging due to the far-reaching linguistic differences between archaic Polish and its modern counterpart. In this paper, we present our recent efforts towards using generative intelligence to create a tool for converting Middle-Polish texts into modern language — a task often referred to as intralingual diachronic translation.

For this purpose, a set of parallel corpora of archaic and modern texts should be used. Yet, scarcity of such resources is an obstacle that hinders the development of diachronic normalization or translation methods based on machine learning [28] and their development would require labour-intensive and time-consuming annotation.

The ongoing development of large language models (henceforth: LLMs) may soon overcome this difficulty. In our solution, we leverage ChatGPT to translate Middle-Polish texts into modern Polish and use the resulting parallel corpus to train a neural text-to-text model. The main source of such texts is the Digital Library of Polish and Poland-Related News Pamphlets from the 16th to the 18th Century[3] (CBDU, from Polish *Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku*) [7,21].

Although the purpose of this work is rather exploratory, several possible practical applications of a diachronic machine translation tool in the digital library can be proposed. Apart from the most obvious improvements in the comprehensibility of the presented artefacts, the translation model could also be used in the indexing engine, allowing the library to be searched using a modern language query.

The paper is structured as follows. Section 2 provides an overview of the relevant literature. The process of creating a parallel corpus of archaic and modern Polish sentences is outlined in Section 3. The development of the tool, based on the aforementioned corpus, is described in Section 4. Automatic and human-based evaluation of the obtained results is presented in Section 5. Finally, Section 6 contains concluding remarks for the paper.

2 Related Work

Most prior research on intralingual diachronic machine translation has focused on East Asian languages. Various studies have discussed machine translation from ancient to modern Chinese, along with creating relevant language resources for this task [19,32,33]. Similarly, [22] describes machine translation from ancient to modern Korean. The work focusing on Indo-European languages is mainly limited to methods of diachronic normalization of spelling [16,2]. Such methods have also been developed for Polish [11,28], however, they focused on language from the 18th and 19th centuries, much closer to modern Polish than the language of earlier centuries, represented in CBDU.

Using LLMs to perform linguistic annotation has already been proposed (see e.g. [30,4]), however, the idea has been gaining more attention as the capabilities of the models continued to rapidly improve. Some of the recently published language resources annotated with LLMs include [8] and [18]. It has been argued that the quality of LLM-annotated data may surpass that of data annotated by humans, including experts [5,10,29]. At the same time, LLMs significantly reduce the cost of annotation and speed up the process.

Most of the work on LLM-based linguistic data annotation has focused on the classification task performed in English. LLM-based annotation in English and Slovenian has been examined in [17]. However, several studies have evaluated the quality of translations generated using LLMs. [12] assessed the quality and robustness of the translations produced with LLMs and found that using

[3] https://cbdu.ijp.pan.pl/

GPT-4 as the engine significantly boosts the performance over ChatGPT-3.5. [9] conducted a human evaluation and analysis of the quality of translation performed with three GPT-based models and compared the results with the highest ranked systems in WMT22. [13] demonstrates that LLMs may leverage paragraph-level context to produce more coherent translations compared to a sentence-to-sentence setting.

As [26] points out, using ChatGPT for text annotation has its drawbacks. ChatGPT is nondeterministic (i.e., identical inputs might lead to different outputs) and vulnerable to adversarial examples, and hence the consistency of the predictions is limited. In some cases, changing even a single character can have a negative impact on the reliability of the outputs [27]. Although these studies examined the classification task, these concerns are valid for text-to-text tasks as well: according to [23], ChatGPT tends to hallucinate when performing a non-English-centric translation.

3 Annotation with ChatGPT

Documents from CBDU (see Section 1) were exported from the original TEI P5 XML format. In total, 258 transcribed documents were used, jointly consisting of over 380,000 segments. During export, the paragraphs were extracted and unnecessary tags were removed – specifically `<gap>` (used to mark a gap in the text, for example, when the original text was unreadable), `<foreign>` tag (used to mark foreign words) and `<pb>` tag (used to mark page breaks). Some words were available in two variants: original and regularized (variants were labelled with `<orig>` and `<reg>` tags and placed inside a `<choice>` tag). In such cases, the latter version was preferred. Furthermore, in some of the documents, certain words were suffixed with the slash symbol ("/") which, if present, was also removed.

The texts were then annotated using ChatGPT (model: `gpt-3.5-turbo`). The paragraphs were split into sentences using Wtpsplit[4] [20] — we decided to use a language-agnostic tool, as punctuation rules in archaic and modern Polish are not the same and applying Polish-specific solutions yielded unsatisfactory results. The translation was performed in batches of sentences; the sentences within each batch originated from the same paragraph, maintaining their original order to ensure coherence. Only the longest paragraphs were split into several batches.

The model was instructed to provide the results in JSON format, which facilitated the parsing of the results. Preliminary experiments have shown that the model occasionally transfers the textual structure (e.g. enumerations) onto the JSON format, resulting in inconsistent key usage. Providing the model with a set of examples successfully prevented that in most cases, hence we decided to use a few-shot prompt. Sporadically, some of the translations were missing from the response. In such cases, the translation was repeated until a correct response

[4] We used the model `wtp-canine-s-12l` and set the threshold parameter at 0.05

was obtained. Some of the texts contained Latin interjections or even consisted solely of Latin text. Most of the translations of such sentences were accurate, nonetheless.

The prompt used consisted of five conversation turns and included two sets of example sentences (same for all examples):[5]

User: Przetłumacz poniższe zdania na współczesny język polski. Podaj odpowiedź w formacie JSON (pod kluczem „translations"). Zdania: *[Translate the below sentences into modern Polish. Provide the answer in JSON format (with the "translations" key). Sentences:]*
„*example source sentence A1*"
„*example source sentence A2*"
...

Assistant:
```
{
  "translations": [
    "example target sentence A1",
    "example target sentence A2",
     ...
  ]
}
```

User: Zdania: *[Sentences:]*
„*example source sentence B1*"
„*example source sentence B2*"
...

Assistant:
```
{
  "translations": [
    "example target sentence B1",
    "example target sentence B2",
     ...
  ]
}
```

User: Zdania: *[Sentences:]*
„*sentence to translate 1*"
„*sentence to translate 2*"
...

The desired model response had the following form:

Assistant:
```
{
  "translations": [
    "translated sentence 1",
    "translated sentence 2",
     ...
  ]
}
```

[5] The example translations included in the prompt can be accessed at https://github.com/ipipan/cbdu-idt

Finally, the translations were parsed. The resulting corpus consists of 12043 pairs of archaic and modern Polish sentences. The corpus was divided into training, development, and test subsets (using an 80:10:10 proportion). The cost of annotation was USD 11.25.

4 Experiments

As the scope of the problem addressed is limited to a single language, we decided to focus on models pre-trained on Polish texts. We tuned plT5-base, plT5-large[6] [3] and plBART[7] for 10 epochs. In all experiments, we held the learning rate at 10^{-4} and used AdamW optimizer ($\varepsilon = 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with linear scheduler (with 500 warm-up steps). The batch size was 2 for plT5-large and 8 for other models. The models were evaluated on the validation subset of the corpus every epoch; we report results on the checkpoint corresponding to the highest BLEU score (i.e. the checkpoint after the last epoch for plT5-base and plBART and after the sixth epoch for plT5-large).

5 Evaluation

5.1 Automatic Evaluation

We assessed the performance of the trained models on the test subset of the corpus. Standard reference-based machine translation evaluation metrics, BLEU, ChrF and TER, were computed using the SacreBLEU library [24] (default parameter values were used). Additionally, we report the values of a recall-oriented metric Rouge-L and a neural-based metric COMET[8] [25]. We also calculated perplexity as a measure of the fluency of the generated text.[9] Note that all the above metrics, except perplexity, require a list of human-produced translations as references, whereas we use sentences translated with ChatGPT. Therefore, it is not possible to compare the performance of ChatGPT and the trained models, and the reliability of the metrics may be limited.

The results of the evaluation are presented in Table 1. Clearly, plT5-large is the best-performing model. Although all models were trained on texts translated using ChatGPT, it is not ChatGPT but plT5-large that produces the most fluent text. This can be explained by the fact that plT5 is a monolingual model trained on a large Polish corpus. Translations produced with ChatGPT tend to contain paraphrases. The translations produced with the trained models are more similar to the original.

---

[6] https://huggingface.co/allegro/plt5-base, https://huggingface.co/allegro/plt5-large

[7] https://huggingface.co/sdadas/polish-bart-base

[8] Model used: `Unbabel/wmt22-comet-da`. The metric is based on the XLM-R model, which only supports modern Polish, hence the performance of the metric may be constrained.

[9] Perplexity was computed using the Polish GPT-2 XL model (https://huggingface.co/sdadas/polish-gpt2-xl)

Table 1. Automatic evaluation results.

Model	BLEU	ChrF	TER	Rouge-L	COMET	Perplexity
Source sentences	—	—	—	—	—	556.06
ChatGPT-3.5	—	—	—	—	—	297.82
plBART-base	15.96	44.25	76.16	0.44	0.65	383.53
plT5-base	18.41	46.20	74.47	0.47	0.66	310.65
plT5-large	**19.55**	**48.18**	**72.04**	**0.49**	**0.70**	**290.50**

5.2 Human evaluation

To compare the performance of ChatGPT and the trained models and to overcome the shortcomings of the reference-based machine translation metrics, we performed a human evaluation of the translations performed with ChatGPT and the best-performing model, plT5-large. A random sample of 100 texts has been selected, each of which has been assessed in two dimensions, adequacy and fluency. Adequacy refers to how accurately the translated sentence conveys the intended meaning of the original sentence; an accurate translation should preserve the core meaning of the original sentence. Fluency assesses how well the translated sentence reads to a native speaker of the target language, i.e. whether it is natural and linguistically correct.

Table 2. Human evaluation: 5-point scale.

Score	Adequacy	Fluency
5	all meaning	flawless Polish
4	most meaning	good Polish
3	much meaning	non-native Polish
2	some meaning	disfluent Polish
1	no meaning	incomprehensible

Following the approach described in [31], a 5-point scale has been used (see Table 2). All texts were evaluated by a single annotator familiar with the Middle-Polish language. Five texts were incomprehensible for the annotator without their context; we decided to replace them with other texts randomly sampled from the test subset of the corpus.

The results of the human-based evaluation are presented in Table 3. Although automatic evaluation suggests that plT5-large produces more fluent translations, here ChatGPT outperforms plT5-large in both adequacy and fluency, the difference in performance is particularly strong for the latter. In Table 4 it can be clearly seen that the language quality of ChatGPT is very good or good in 92% of the cases, compared to 76% for sentences translated with plT5-large.

Table 3. Human evaluation results: average scores.

Model	Adequacy	Fluency
ChatGPT	**3.78**	**4.48**
plT5-large	3.54	4.07

Table 4. Human evaluation results: score frequencies.

Score	Adequacy (%)		Fluency (%)	
	ChatGPT	plT5-large	ChatGPT	plT5-large
5	24	17	57	36
4	38	37	35	40
3	30	31	7	21
2	8	13	1	1
1	0	2	0	2

Some additional remarks can be drawn from the annotator's notes, which were collected during the manual evaluation process. plT5-large could not successfully translate some of the short sentences, even if ChatGPT handled them correctly. For some medium-length sentences, which scored very high (5) in fluency for ChatGPT, the translation generated with plT5-large was of poor quality (2). In the case of very long sentences, plt5-large tends to generate much shorter translations, omitting some parts of the original sentence, resulting in a lower adequacy score. Some of the plT5 translations contained repetitions of words or longer phrases, having a negative impact on the fluency score. The sample contained a sentence consisting solely of Latin words, which plT5 failed to correctly translate: not only did it not translate the sentence into Polish, but the output contained Latin words mixed with words of neither Polish nor Latin origin; at the same time, ChatGPT achieved a fair translation quality.

It is possible that the low performance of plT5-large results from insufficient size of the training corpus or over-training — the best checkpoint was selected based on the BLEU score, however it has been shown that the metric has flaws and using it for model selection might lead to suboptimal decisions [15].

6 Conclusions

This paper addresses the problem of intralingual diachronic translation for Polish. Archaic texts from the CBDU digital library were translated into contemporary Polish using ChatGPT. The resulting parallel corpus was used to train neural text-to-text models. The results of the automatic evaluation indicate that tuned models may produce translations more fluent than ChatGPT (on which they were trained). However, human-based evaluation has shown that ChatGPT translations are better than the translations of the best-performing trained model in terms of both adequacy and fluency. While the performance of the trained models is far from perfect, the quality of ChatGPT-produced translations was rated as good or very good in most cases. Although caution should be exercised, we believe that LLMs have a high potential for text-to-text annotation applications. The annotated corpus and the predictions of the models are available at https://github.com/ipipan/cbdu-idt.

Possible reproducibility issues are a major limitation of this work: as ChatGPT is nondeterministic, repeating the procedure described would likely yield different translations, even if the same prompt was used. Training may not have been conducted optimally due to reliance on BLEU values for model selection (see Section 5). Furthermore, The prompt used to translate the texts included the same set of examples for all cases; using an example selection technique could improve the quality of the translation [1,9].

References

1. Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., Ghazvininejad, M.: In-context examples selection for machine translation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8857–8873. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.findings-acl.564
2. Bawden, R., Poinhos, J., Kogkitsidou, E., Gambette, P., Sagot, B., Gabay, S.: Automatic normalisation of early Modern French. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 3354–3366. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.lrec-1.358/
3. Chrabrowa, A., Dragan, Ł., Grzegorczyk, K., Kajtoch, D., Koszowski, M., Mroczkowski, R., Rybak, P.: Evaluation of transfer learning for Polish with a text-to-text model. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 4374–4394. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.lrec-1.466
4. Ding, B., Qin, C., Liu, L., Chia, Y.K., Li, B., Joty, S., Bing, L.: Is GPT-3 a good data annotator? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11173–11195. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.626
5. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. arXiv:2303.15056 (2023)
6. Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., Woliński, M.: The Electronic Corpus of 17th-and 18th-century Polish Texts. Language Resources and Evaluation **56**(1), 309–332 (2022)
7. Gruszczyński, W., Ogrodniczuk, M.: Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski dotyczących z XVI, XVII i XVIII w. w nauce i dydaktyce. In: Materiały konferencji Polskie Biblioteki Cyfrowe 2010. pp. 23–27. Poznań, Poland (2010)
8. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E.: ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3309–3326. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.234

9. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How good are GPT models at machine translation? A comprehensive evaluation. arXiv:2302.09210 (2023)
10. Huang, F., Kwak, H., An, J.: Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023. ACM (2023). https://doi.org/10.1145/3543873.3587368
11. Jassem, K., Graliński, F., Obrębski, T., Wierzchoń, P.: Automatic Diachronic Normalization of Polish Texts. Investigationes Linguisticae **37**, 17–33 (2018). https://doi.org/10.14746/il.2017.37.2.
12. Jiao, W., Wang, W., Huang, J., Wang, X., Tu, Z.: Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745 (2023)
13. Karpinska, M., Iyyer, M.: Large language models effectively leverage document-level context for literary translation, but critical errors persist. arXiv:2304.03245 (2023)
14. Kieraś, W., Woliński, M.: Manually annotated corpus of Polish texts published between 1830 and 1918. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018), https://aclanthology.org/L18-1609
15. Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., Menezes, A.: To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: Proceedings of the Sixth Conference on Machine Translation. pp. 478–494. Association for Computational Linguistics, Online (2021), https://aclanthology.org/2021.wmt-1.57
16. Korchagina, N.: Normalizing Medieval German Texts: from rules to deep learning. In: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. pp. 12–17. Linköping University Electronic Press, Gothenburg (2017), https://aclanthology.org/W17-0504
17. Kuzman, T., Mozetic, I., Ljubešic, N.: ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. arXiv:2303.03953 (2023)
18. Laskar, M.T.R., Rahman, M., Jahan, I., Hoque, E., Huang, J.: CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia. arXiv:2305.06147 (2023)
19. Liu, D., Yang, K., Qu, Q., Lv, J.: Ancient–Modern Chinese Translation with a New Large Training Dataset. ACM Transactions on Asian and Low-Resource Language Information Processing **19**(1) (2019). https://doi.org/10.1145/3325887
20. Minixhofer, B., Pfeiffer, J., Vulić, I.: Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. arXiv:2305.18893 (2023)
21. Ogrodniczuk, M., Kryńska, K.: Evaluating Machine Translation of Latin Interjections in the Digital Library of Polish and Poland-related News Pamphlets. In: Tseng, Y.H., Katsurai, M., Nguyen, H.N. (eds.) From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries. ICADL 2022. pp. 430–439. No. 13636 in Lecture Notes in Computer Science, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-21756-2$_3$4
22. Park, C., Lee, C., Yang, Y., Lim, H.: Ancient Korean Neural Machine Translation. IEEE Access **8**, 116617–116625 (2020). https://doi.org/10.1109/ACCESS.2020.3004879
23. Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D.: Towards Making the Most of ChatGPT for Machine Translation. arXiv:2303.13780 (2023)

24. Post, M.: A Call for Clarity in Reporting BLEU Scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Belgium, Brussels (2018), https://www.aclweb.org/anthology/W18-6319
25. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A Neural Framework for MT Evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Online (2020), https://www.aclweb.org/anthology/2020.emnlp-main.213
26. Reiss, M.V.: Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. arXiv:2304.11085 (2023)
27. Shen, X., Chen, Z., Backes, M., Zhang, Y.: In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. arXiv:2304.08979 (2023)
28. Skórzewski, P., Jassem, K., Graliński, F.: Automated Normalization and Analysis of Historical Texts. In: Vetulani, Z., Paroubek, P., Kubis, M. (eds.) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2017. Lecture Notes in Computer Science, vol. 12598, pp. 73–86. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-66527-2$_6$
29. Törnberg, P.: ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv:2304.06588 (2023)
30. Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: Want To Reduce Labeling Cost? GPT-3 Can Help. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4195–4205. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.354
31. White, J.S., O'Connell, T.A., O'Mara, F.E.: The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas. Columbia, Maryland, USA (1994), https://aclanthology.org/1994.amta-1.25
32. Yang, Z., Chen, K.j., Chen, J.: Guwen-UNILM: Machine Translation Between Ancient and Modern Chinese Based on Pre-Trained Models. In: Wang, L., Feng, Y., Hong, Y., He, R. (eds.) Natural Language Processing and Chinese Computing. NLPCC 2021. Lecture Notes in Computer Science, vol. 13028, pp. 116–128. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-88480-2$_1$0
33. Zhang, Z., Li, W., Su, Q.: Automatic Translating Between Ancient Chinese and Contemporary Chinese with Limited Aligned Corpora. In: Tang, J., Kan, M.Y., Zhao, D., Li, S., Zan, H. (eds.) Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science, vol. 11839, pp. 157–167. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6$_1$3