Can rules still beat neural networks?
The case of automatic normalisation for 18th-century Portuguese texts

Leonardo Zilio
FAU Erlangen-Nürnberg, Germany
leonardo.zilio@fau.de

Rafaela R. Lazzari
UFRGS, Brazil
rafaelalazzari@hotmail.com

Maria José B. Finatto
UFRGS, Brazil
mariafinatto@gmail.com

Abstract

In this paper, we tested whether fine-tuned neural machine translation (NMT) models can produce better results than a rule-based method for the task of normalising historical medical documents written in 18th-century Portuguese. We used a replacement glossary as basis for the rule-based method, and tested three NMT models against it in an in-domain setting and in two out-of-domain scenarios. In-domain results showed that the rule-based method was better than off-the-shelf NMT models, and it still surpassed one of the in-domain fine-tuned models. The fine-tuned models showed their efficacy on out-of-domain settings, where only one NMT model did not surpass the rule-based method in one scenario.

1 Introduction

Working with historical documents has proven time and time again to be an enormous challenge for Natural Language Processing (NLP) (cf. Quaresma and Finatto, 2020; Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022). While there is progress in the field, most of the tools developed for working with natural language have modern iterations of the language as focus, and only a few studies have been dedicated to computationally process historical documents as they are, and even fewer such studies exist for historical Portuguese.

To help alleviate the historical gap between historical and modern-era texts, researchers started resorting to normalising the writing of historical documents (Piotrowski, 2012; Bollmann and Søgaard, 2016; Bawden et al., 2022); that is, they started updating the spelling of historical texts based on modern-day orthographic rules. However, this normalisation work is mainly done manually and is thus very time consuming.

This study has the main objective of exploring ways of automatically normalising documents, so that less work has to be spent in converting the writing of historical documents into modern-day standards, and allowing for the mass-normalisation of larger corpora. Taking advantage of already existing normalised, available corpora, and of recently developed machine translation models, we analyse how three multilingual neural machine translation (NMT) models fare when compared to a rule-based normalisation model that uses a static glossary as main reference.

The main contributions of this paper are the following:

- The release of a dataset for fine-tuning and testing NMT on the task of normalising historical medical texts written in Portuguese.
- The release of scripts for automatic normalisation of historical documents[1]. These scripts are fairly simple to use and can also be applied in other tasks related to sequence-to-sequence translation.
- A comparison of three off-the-shelf NMT models and their fine-tuned version in the task of normalising historical texts, both in and out of domain.
- An error analysis that shows what might still pose problems for fine-tuned NMT models in this context.
- A support for the further analysis of historical medical documents, such as the

[1] These scripts and datasets can be found on the following repository: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

one carried out by Lazzari and Finatto (2023).

- The advancement of the project *Corpus Histórico da Linguagem da Medicina em Português do Século XVIII* [Historical Corpus of Medical Language in 18th-century Portuguese][2]

The remainder of the paper is organised as follows: Section 2 discusses other work related to the normalisation of historical texts; Section 3 briefly describes our historical corpus and the four methods used for automatically normalising historical sentences; Section 4 presents the results of the automatic normalisation experiments; in Section 5, we discuss some key issues detected when analysing what went wrong with the automatic normalisation; Section 6 describes an experiment with out-of-domain historical documents, to evaluate the robustness of in-domain fine-tuned and rule-based methods; Section 7 sums up and discusses some aspects of the experiments with in- and out-of-domain normalisation, and discusses future work.

2 Related work

Several studies have been dedicated to the normalisation of historical documents in various languages, including Portuguese. As for automatic normalisation, most of the approaches seem to have stopped before the advent of transformer models, which make this study unique in applying the most recently developed NMT models based on the transformers architecture.

Most studies involving NMT use an encoder-decoder, character-based architecture based on long-short term memory (LSTM) models (cf. Bollmann and Søgaard, 2016; Domingo and Nolla, 2018; Domingo and Casacuberta, 2019). While these studies make sense, by modelling the spelling normalisation problem as a character-based replacement, much similar to what rule-based systems have done, Tang et al. (2018) have already hinted that subword tokens can provide a better solution to character-based models. This would naturally lead to the use of transformer-based architecture. However, as far as we could verify, the study by Tang et al. (2018) is the only one testing transformers for this task to date, and Portuguese is not among the tested languages.

In terms of languages, the focus of studies on automatic normalisation have been on European languages. Bollmann (2019) developed a large comparison of automatic normalisation methods for English, German, Hungarian, Icelandic, Portuguese, Slovene, Spanish, and Swedish. Studies with less languages involve the work of Domingo and Casacuberta (2019) for Slovene and Spanish, Bawden et al. (2022) for French, and Robertson (2017) for English, German, Icelandic, and Swedish. For Portuguese, we could only find the above-mentioned work of Bollmann (2019), who used a corpus of letters from the 15th to 19th century that was made available by the Post Scriptum project (CLUL, 2014).

More recently, researchers at the University of Évora started working with text normalisation. Cameron et al. (2023) propose a categorisation of variants, which can support the normalisation of historical Portuguese texts, and Olival et al. (2023) present and discuss the normalisation of six documents that belong to the *Parish Memories*. There are also some papers that use normalised versions of Portuguese documents for different NLP tasks, such as named entity recognition (Zilio et al., 2022) and textual complexity (Zilio et al., 2023).

Considering the work that has been done, this study is the first to present an automatic approach for normalising historical medical documents in Portuguese, and possibly the first to leverage existing multilingual, transformer-based NMT models for the normalisation task.

3 Methodology

In this section, we briefly describe the corpora that were used for in-domain fine-tuning of NMT models and for glossary extraction, and also for in- and out-of-domain testing. We also describe our baseline rule-based method and present the multilingual NMT models.

[2] For more information about this project, please visit the following website (in Portuguese): https://sites.google.com/view/projeto38597. The project website also contains more information about the historical medical corpus that we use in this study.

3.1 Corpora

In this study, we used a total of three corpora, all of them written in Portuguese in the 18th century: a historical medical corpus, which is the focus of this study and was used for fine-tuning and testing NMT models, and for extracting a glossary for the rule-based approach; a historical corpus of censual information collected by priests in different Portuguese parishes; a historical corpus of letters collected within the Post Scriptum project (CLUL, 2014). All corpora were semi-automatically aligned at the sentence level using OmegaT's aligner tool[3]. This process allowed the generation of TMX files, which were then used for further preprocessing the aligned texts for the different tasks.

Our historical medical corpus was originally transcribed from three books written in the 18th century: *Observaçoens Medicas Doutrinaes de Cem Casos Gravissimos* [Medical and Doctrinal Observations of a Hundred Severe Cases] (Semedo, 1707), *Postilla Religiosa, e Arte de Enfermeiros* [Religious Postil, and Art of Nurses] (de Sant-Iago, 1741) and *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health] (Mauran, 1794). Since we needed to manually normalise each of the texts used in this study, we only selected a few documents from each of the books, aiming at a balanced corpus.

Some documents from the *Parish Memories* corpus have recently undergone a normalisation process (Olival et al., 2023), so we took advantage of this fact and used this corpus as an out-of-domain test for our automatic normalisation systems. For this task, we used the six documents related to Vila Viçosa (a location in Portugal) that are currently available in normalised format[4]. Each document was written by a different author, and each refers to a parish in Vila Viçosa: Nossa Senhora das Ciladas, Nossa Senhora da Conceição, Pardais, Santa Ana de Bencatel, São Bartolomeu, and São Romão.

While the *Parish Memories* provide an out-of-domain test set, it is still a somewhat structured type of text, in which each paragraph contains very specific information about a census that was carried out in 1758 in Portugal. To provide an even less structured test to our automatic normalisation models, we resorted to a selection of handwritten letters from the Post Scriptum collection (CLUL, 2014). The full corpus from the 18th century contains 758 letters[5]. However, due to the semi-automatic nature of the sentence-alignment process, we randomly selected 10 letters from the corpus (taking care of selecting five from each of the two available subcorpora). Here is the list of letters that were used in this study: CARDS1082, CARDS1089, CARDS2108, CARDS2707, CARDS3148, PSCR0515, PSCR0613, PSCR1648, PSCR2526, and PSCR4643.

A very important caveat needs to be presented here: our historical corpus of medical documents was normalised having modern Brazilian Portuguese as reference, while the other two corpora were normalised having European Portuguese as reference. As such, for instance, while in our corpus we normalised words like “cousa” to “coisa” [thing], this was not done in the other two corpora, as “cousa” can still be found in European Portuguese. This certainly had an impact in the results of the experiments and should be kept in mind when observing the results that we present in this study.

Table 1 presents the data information for each corpus. As can be seen in the table, our historical medical corpus has a total of 5,584 types, while the version with modernised spelling has 5,341 types. This gives us an idea of how much spelling variation there was in the original corpus: we have 1.05 type for each type in the normalised corpus. This variation is larger in both other corpora, and a possible reason for this is that they are both based on handwritten documents by several different authors, while our medical corpus was published in printed form and was the work of three authors. The medical corpus also clearly

[3] OmegaT is an open-source tool used for computer-assisted translation. It is available at: https://omegat.org/.

[4] The original texts are available on CIDEHUS's website (https://www.cidehusdigital.uevora.pt/portugal1758), while the normalised versions are provided as annex in Olival et al. (2023).

[5] All files can be freely downloaded from the Post Scriptum website: http://teitok.clul.ul.pt/postscriptum/index.php?action=downloads.

distinguishes itself from the others by the amount of tokens per sentence, with around 54 T/S against ~39 and ~17 for the Parish Memories and the Post Scriptum, respectively. The medical corpus is marked by a constant use of semi-colons, where in a modern writing probably a full stop would be used. The much smaller sentence size in the Post Scriptum corpus is mostly due to the genre, but the normalisation probably also contributed to this: many of the original letters have little to no punctuation, and the normalisers added punctuation, including full stops, in the normalisation process, which might have led to a more modern use of punctuation.

The medical corpus was further split into train, development (dev) and test sets, for fine-tuning NMT models. Table 2 shows the number of tokens, types and sentences in each split, considering original and normalised versions of the corpus. An important detail in the design of the splits is that the texts used in the train and dev sets were different from the ones used in the test set. The train and dev sets were a random selection of sentences (90% for train and 10% for dev) from these texts:

- ***Aviso***: chapters 2, 8, and 13, all from the second part of the book.
- ***Observaçoens***: observations 42, 88, and 92.
- ***Postilla***: chapters 17, 22, 29, 30, 32, 33, 34, 40, 41, 42, 43, 44, 46, 47, 48, and 58, all from the second part of the book[6].

For the test set, we used one text from *Aviso* (chapter 5, also from the second part of the book) and from *Observaçoens* (observation 92), and two chapters from the *Postilla* (chapters 1 and 7, also from the second part of the book).

3.2 Rule-based method

The rule-based normalisation method was planned as a baseline for the automatic normalisation process. We used a glossary of aligned original and normalised words that was automatically extracted from the combined train and development corpus.

To extract this glossary, we first had to use a word-level aligner, to identify the pairs of historical-normalised words that actually underwent any change. For this, we used SimAlign (Sabet et al., 2020), along with the recently released Albertina model (PT-PT) (Rodrigues et al., 2023), and we carried out a semi-automatic alignment, in which instances of no alignment or of many-to-one alignments were validated manually. However, there might still be a few one-to-one wrong alignments in the dataset.

From this word-aligned dataset, we observed that 1,228 types in the original texts had a different spelling when compared to their normalised counterparts. This indicates that almost a third (31.46%) of the types needed to be normalised, reinforcing the importance of automatising the normalisation process.

The word-aligned dataset was used as input for the glossary. We also manually removed the entry "as" = "às", because "as" might simply be the plural form of the feminine definite article "a" [the$_{feminine}$], and not the merge of preposition "a" and the plural form of the feminine definite article, as it was represented in the automatically extracted glossary. After this cleaning, the resulting glossary was then used as a replacement dictionary.

The first step in the process for the rule-based normalisation was to tokenise each sentence with NLTK's[7] word tokeniser. Then each token was checked against the glossary to verify if any replacement was needed. Phrases longer than one token were processed separately in a similar way. If a word or phrase was present in the historical text, then it was replaced with its normalised form. The rules also ensured that punctuation was correctly rendered in the output (for instance, by removing space between a word a comma, which is very common in historical documents).

3.3 Neural machine translation models

We used three multilingual neural machine translation (NMT) models:

- **opus-mt-tc-big-itc-itc (OPUS)**[8]: this

[6] The chapters in the *Postilla* are smaller, so we had to select more chapters than in the other two books in order to balance the dataset.

[7] NLTK's website: https://www.nltk.org/.

[8] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-itc-itc.

	Medical		Parish Memories		Post Scriptum	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	24504	24815	9561	9661	2547	2549
Types	5584	5341	2381	2027	950	852
Type Ratio	-	1.05	-	1.17	-	1.12
Sentences	453	453	244	244	147*	147
T/S	54.09	54.78	39.18	39.59	17.33	17.34

Table 1: Corpus information. Type Ratio = division of types in the original by types in the normalised corpus; T/S = tokens per sentence.
* The number of sentences in the PS original corpus was based on the normalised version, as there are very few or no instances of punctuation in some of the original letters.

	Train		Dev		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	18047	18286	2038	2067	4419	4462
Types	3386	3213	826	803	1372	1325
Type Ratio	-	1.05	-	1.03	-	1.04
Sentences	342	342	38	38	73	73
T/S	52.77	53.47	53.63	54.39	60.53	61.12

Table 2: Information about the individual data splits. Type Ratio = division of types in the original by types in the normalised corpus; T/S = tokens per sentence.

model was originally trained in the scope of the OPUS-MT project (Tiedemann and Thottingal, 2020; Tiedemann, 2020). It comprises 19 languages from the Italic family, including Portuguese, and it was trained with all possible language combinations. This model is by far the smallest, as the final folder of the fine-tuned model has a size of only around 2.3GB, while the other two have a size of almost 7GB each.

- **mbart-large-50-many-to-many-mmt (mBart)**[9]: this model was originally developed by Tang et al. (2020) and comprises 50 languages, including Portuguese, trained in a many-to-many fashion, *i.e.* all possible language pairs are included in the training set.

- **nllb-200-distilled-600M (NLLB)**[10]: the NLLB paper (Team, 2022) caused much stir in the machine translation community, as it offers a huge combination of languages, including low-resource languages. This model builds on the idea of leveraging higher resourced languages for the automatic translation of low resourced ones. Because it involves so many languages, it is also a less focused model, and while it works in advancing the machine translation state of the art for some low resourced languages, it might not perform as brilliantly for highly resourced ones, such as Portuguese.

These models were first tested as they are provided by their developers, to set some baselines for the models themselves, and then they were also used in a fine-tuning pipeline, where our training and development datasets were used to adapt these models to our normalisation task. For fine-tuning, we used the standard Transformers library, as provided by Huggingface[11] (Wolf et al., 2020). All models were fine-tuned with the same parameters, except for batch size, as the larger models were simply too large for our single graphics card NVidia RTX 4090 (with 24GB RAM) to handle: learning rate of 2e-5, weight decay of 0.01, and 100 epochs; batch size was 16 for OPUS, 6 for

[9] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt.

[10] https://huggingface.co/facebook/nllb-200-distilled-600M.

[11] https://github.com/huggingface/transformers.

NLLB and 4 for mBart. All other parameters were left as default. At the end of the fine-tuning process, the best model was selected based on BLEU scores (Papineni et al., 2002), as evaluated in the default implementation of SacreBLEU (Post, 2018) in Huggingface's Evaluate library[12].

None of these models differentiate between Brazilian and European Portuguese, the two varieties that we are working with, so we simply used the tags "por" (European Portuguese) and "pob" (Brazilian Portuguese) for reference, but none of the models were actually trained to differentiate between the two. As such, we actually fine-tuned the models to translate from Portuguese into Portuguese, but using a dataset that was specifically curated for this normalisation task.

3.4 Evaluation

We evaluated all models using the BLEU score metric (Papineni et al., 2002). Several papers criticise the use of BLEU scores, including the paper that proposes SacreBLEU (Post, 2018), which is the implementation that we used via the Evaluate library from Huggingface, as explained in the previous subsection. BLEU is a metric that compares the number of n-grams in the target text with reference text(s), and produces a score from 0 to 100. Because any size of n-gram can be used, it is a metric that has to be well-detailed in the methodology to be reproducible, something that SacreBLEU addresses very well.

Another downside of BLEU is that it bases the correctness of a target text on the basis of reference texts. These references may or may not be good target texts themselves, and they do not necessarily invalidate other alternative, equally correct translation options for a given source text. As such, a low BLEU score might be just a reflex of different translation choices in the reference texts. While this issue can be mitigated by using several reference texts for each test sentence, several references are not always available. In our case, however, most of the time, there is no alternative correct option for a given token in the normalisation pipeline. Most historical words can only be normalised to one single form in the modern spelling paradigm, so the issue of multiple references will rarely apply, making BLEU a perfectly sound choice for evaluating a spelling normalisation task. The choice of SacreBLEU also ensures that any researcher can use the exact same format of BLEU when trying to reproduce this study, as we used the default parameters of the metric.

Model	SacreBLEU
Baseline models	
OPUS*	47.57
mBart	30.73
NLLB	40.64
Rule-based model	
Replacement glossary	83.26
Fine-tuned models	
OPUS	75.05
mBart	88.20
NLLB	83.65

Table 3: SacreBLEU scores based on our test set.
* We prepended >>pob<< to the source text, as required by the system. Without prepending the language ID, the model translated the source text into Spanish, and it achieved a BLEU score of 7.77. Prepending >>pob<<actually made the fine-tuned system perform around 2 BLEU points worse on both test sets, so we did not prepend >>pob<< for the fine-tuned model.

4 Results of the automatic normalisation

As it can be seen in Table 3, the baseline models (without fine-tuning) perform very poorly on our data, with the highest BLEU score being achieved by OPUS at 47.57. Meanwhile, our simple rule-based system achieved 83.26 in the BLEU scale. Surprisingly, even after fine-tuning, the rule-based system remained very competitive, and was still better than the OPUS model by more than 8 points and was only barely surpassed by the NLLB model, giving us an initial answer to our main question in the title of this paper. However, mBart showed a great improvement with fine-tuning (an increase of more than 54 points) and achieved the highest score, almost five points higher than the second-best model.

Although mBart was able to beat the rule-based model with some margin, the results seem to show that a well-curated glossary

[12] https://github.com/huggingface/evaluate.

could actually be a better option for automatic normalisation, depending on the needs of the researchers and on the equipment available, as fine-tuning mBart is not a computationally cheap task: it requires a computer with a high-end graphics card, even for a small train and development dataset such as ours. The NMT models are also much slower at the inference phase (*i.e.*, when they are producing the normalised text): while the rule-based method is almost instantaneous for our test set, the NMT models take a few minutes on a good GPU, and up to an hour on a 12th Gen Intel Core i7-1260P CPU.

5 Error analysis

To better understand what types of errors were prevalent in the fine-tuned NMT models, we conducted an error analysis, focusing on sentences that had a low BLEU score[13]. We analysed the sentences checking for missing translations, overtranslations, hallucinations, and any common pattern that we could identify that helped bring down the scores.

In the OPUS model, we noticed that several normalised sentences missed portions of the source text, and we also noticed that the system was producing several hallucinations. One example of hallucination from OPUS is the following:

> cuja verdadeira descarga se deve fazer por via de purga, & naõ de sangria, & por estas razões as purguei com felicidade [14]

↓

> mas sim com os quais se descarregou a verdadeira felicidade, e purgaram as sangrias, e por isso só por causa de que os reis purgaram purgas [15]

The NLLB model had much less salient issues, as they were more focused on single tokens, and involved miss-normalisations or lack of normalisation, and the substitution of historical words with synonyms. A similar error pattern was observed for mBart, but it presented only a few cases of replacement with a synonym. These usually single-token errors included lack of or non-removal of diacritics in most cases; this involved the model simply not changing the word in the source text. In the NLLB model, we also observed a few hallucinations, mostly just short repetitions of words, such as “poreis poreis poreis” [(you will) put put put] and “sumas sumas” [(that you) disappear disappear]. For mBart, one curious case was the replacement of Outubro [October] with Novembro [November] in a segment, but the rest were mostly small issues.

6 Robustness test: use in out-of-domain historical texts

As the models that we developed and fine-tuned were focused on specialised historical medical language, we wanted to check how much information had also been gained for normalisation on out-of-domain texts. This was an experiment in “knowledge” transfer, where we try to observe how much of the information that was gathered from medical texts can be transferred to the normalisation of texts from other domains. For this, we tested our models on the normalised texts from the *Parish Memories* corpus and on normalised letters from the Post Scriptum dataset, as we described in Section 3.

Table 4 presents the results for all the models, including the non-fine-tuned ones, as a comparison for how much improvement was brought about by the fine-tuning procedure, and for how difficult the task was in relation to the normalisation task in our medical corpus. We can clearly see that the Post Scriptum dataset was much harder to normalise. Some of the originally transcribed texts do not have punctuation and have many abbreviations, which are usually extended in the normalised version. This made it much more difficult for all models to achieve a good normalisation, as they were not fine-tuned to add punctuation or to extend abbreviations. In

[13] For our reference in the error analysis, we computed the sentence-level BLEU score separately from the one presented in Table 3. We considered low BLEU scores the ones that deviated by at least one standard deviation from the model’s mean in the test set. This means that those considered as bad sentences in one model could actually be better than some “good” sentences in another model.

[14] Free translation: *whose true elimination should be done via purge, and not bleeding, and because of this I happily purged them.*

[15] The text does not make much sense, so we tried to keep a more literal translation: *but actually with those that true happiness eliminated itself, and purged bleedings, and that’s why only because of that the kings purged purges.*

the *Parish Memories*, with the exception of mBart, which had an almost 6-point worse BLEU score, the results of the baseline models did not vary too much from the results in the medical dataset. In both out-of-domain datasets, the rule-based method scored more than 7 BLEU points higher than OPUS, the best baseline NMT model.

When looking at the fine-tuning improvement, we see that, except for OPUS on the Post Scriptum dataset, all NMT models performed above the rule-based method. As expected, all of them performed worse than on the in-domain dataset, but the results in the *Parish Memories* were still much better than the ones achieved by their non-fine-tuned baselines, with improvements ranging from around 13 BLEU points for OPUS up to ~33 points for mBart. In the Post Scriptum dataset, the improvements were more modest, ranging from ~7 BLEU points for OPUS up to ~29 points for mBart. In this out-of-domain test, we also see that the fine-tuned NLLB model seems to really be able to draw on its information about 200 languages for keeping it robust, as it achieved the best score on both datasets, clearing more than 3 BLEU points from mBart.

Model	SacreBLEU	
	Parish Memories	Post Scriptum
Baseline models		
OPUS*	45.72	27.08
mBart	24.94	6.79
NLLB	39.55	20.80
Rule-based model		
Replacement glossary	53.70	34.56
Fine-tuned models		
OPUS	58.77	34.05
mBart	58.10	36.01
NLLB	61.41	39.34

Table 4: SacreBLEU evaluation scores on out-of-domain corpora.
* We prepended >>pob<< to the source text, as previously explained on Table 3.

7 Final remarks

In this paper, we set out the task of testing neural machine translation (NMT) models for automatically normalising historical medical documents. We compared fine-tuning methods with a rule-based implementation of a replacement method mainly based on a glossary, and the results showed that the rule-based method was indeed a strong baseline for the NMT models. It surpassed the non-fine-tuned NMT models in all scenarios, scoring up to ~52 BLEU score points higher in the in-domain test.

After fine-tuning the NMT models, as expected, all models improved over their baseline versions, but only mBart was clearly superior to the rule-based method. OPUS still scored ~8 BLEU points lower, and NLLB was only marginally superior (less than one point). As such, as a preliminary answer to the question in the title of this paper, we can say that rule-based systems can still be superior to neural-network-based methods in some scenarios, and they are certainly much less complicated to implement and less power-consuming.

The fine-tuning advantage of the NMT models was, however, clearly shown in the out-of-domain test, where all models scored at least 4 points higher than the rule-based method when tested on the *Parish Memories* dataset, and only OPUS was not able to beat the rule-based method on the Post Scriptum dataset, showing that the fine-tuned models are better able to transfer the information gathered from one domain to another. Still, when the task was too far off, as in the case of the handwritten letters of the Post Scriptum dataset, a post-editor with the task of normalising texts would probably be better served by a glossary replacement method. Such method at least would be less intrusive, as most errors would be in the form of non-changed input, rather than an erroneously changed input (such as the hallucinations produced by NMT). However, a detailed post-editing task would need to be developed to better test this hypothesis.

In terms of fine-tuning improvement over the non-fine-tuned baselines, mBart was the model that had the best result in all scenarios. On the opposite side, OPUS was the model that showed the least improvement in all tasks. The OPUS model we used was specifically trained on Italic languages, which gave it the best result in all baseline tests. However its fine-tuned version was inferior to the rule-based method both in and out of domain, only being able to beat the rule-based method (and marginally also mBart) when tested on the *Parish Memories*.

It was interesting to see that, in the out-of-domain task, NLLB was superior to mBart in both datasets, probably due to the larger linguistic scope of the model. It is still not yet fully clear if this is caused by mBart being perhaps more prone to overfitting, and OPUS (and also NLLB) being then less prone to overfitting, or if the out-of-domain tasks rely more on the breadth of linguistic information that was used in the original training of the models. These are all questions that we plan to investigate in the future, as further tests are needed to verify them.

The work on this paper also sets out a methodology for replicating the work using other corpora, covering other time periods, other domains, and even other languages. With the scripts that are now available on Github[16], it is also possible to train models for the task of translating (instead of normalising) historical documents into modern languages using a very similar methodology as we presented in this paper, so one further future task is to create data for testing these models in a diachronic intralingual translation setting.

Acknowledgements

We would like to thank the Chair of Computational Corpus Linguistics at FAU Erlangen-Nürnberg and the PPG-LETRAS at UFRGS for support, and also the following Brazilian Institutions: PROPESQ - UFRGS – CNPq, for a PIBIC undergraduate research grant; CNPq, for funding a Productivity Research Grant (06/2019, 308926/2019-6) and a research project (26/2021 – PDE – 200051/2023-7); and FAPERGS-CAPES, for funding a research project (06/2018 – INTERNAC., 19/2551-0000718-3).

References

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional lstms and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139.

Helena Cameron, Fernanda Olival, Renata Vieira, and Joaquim Santos. 2022. Named entity annotation of an 18th-century transcribed corpus: problems and challenges. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 18–25. CEUR.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistórico, Rio de Janeiro, ISSN*, pages 2359–6910.

CLUL. 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*.

Fr. Diogo de Sant-Iago. 1741. *Postilla religiosa, e arte de enfermeiros: guarnecida com eruditos conceitos de diversos authores, facundos, moraes, e escriturarios*. Officina de Miguel Manescal da Costa, Lisboa, Portugal.

Miguel Domingo and Francisco Casacuberta. 2019. Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents. In *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 59–69. Springer.

Miguel Domingo and Francisco Casacuberta Nolla. 2018. Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 129–137. European Association for Machine Translation.

[16] Please check our repository: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

Rafaela Radünz Lazzari and Maria José Bocorny Finatto. 2023. Exame do vocabulário médico no português no século xviii: contribuições da lexicometria para o desenho de um dicionário histórico. *Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)*, 7(1):102–123.

G. Mauran. 1794. *Aviso a' Gente do Mar sobre a sua Saude*. R. Typ. de João Antonio da Silva, Lisboa, Portugal. Translated from the French original edition and extended with some notes by Bernardo José de Carvalho.

Fernanda Olival, Helena Freire Cameron, Fátima Farrica, and Renata Vieira. 2023. As Memórias Paroquiais (1758) do atual concelho de Vila Viçosa. *Callipole*, 29:85–128.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Paulo Quaresma and Maria José Bocorny Finatto. 2020. Information extraction from historical texts: a case study. In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 49–56. CEUR.

Alexander Robertson. 2017. *Automatic Normalisation of Historical Text*. Ph.D. thesis, School of Informatics, University of Edinburgh.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertina PT. *arXiv preprint arXiv:2305.06721*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

João Curvo Semedo. 1707. *Observaçoens Medicas e Doutrinaes de Cem Casos Gravissimos*. Officina de Antonio Pedrozo Galram, Lisboa, Portugal.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Leonardo Zilio, Maria Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the XVIII century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 1–10. CEUR.

Leonardo Zilio, Maria José B. Finatto, Renata Vieira, and Paulo Quaresma. 2023. A natural language processing approach to complexity assessment of 18th-century health literature. *Domínios de Lingu@gem*, 17.