Normalizing without Modernizing: Keeping Historical Wordforms of Middle French while Reducing Spelling Variants

Raphael Rubino **Johanna Gerlach** **Jonathan Mutal** **Pierrette Bouillon**
TIM/FTI, University of Geneva
1205 Geneva, Switzerland
firstname.lastname@unige.ch

Abstract

Conservation of historical documents benefits from computational methods by alleviating the manual labor related to digitization and modernization of textual content. Languages usually evolve over time and keeping historical wordforms is crucial for diachronic studies and digital humanities. However, spelling conventions did not necessarily exist when texts were originally written and orthographic variations are commonly observed depending on scribes and time periods. In this study, we propose to automatically normalize orthographic wordforms found in historical archives written in Middle French during the 16th century without fully modernizing textual content. We leverage pre-trained models in a low resource setting based on a manually curated parallel corpus and produce additional resources with artificial data generation approaches. Results show that causal language models and knowledge distillation improve over a strong baseline, thus validating the proposed methods.

1 Introduction

Normalizing orthographic variations of historical texts is a crucial task for digital humanities. It allows for both conservation and easy consultation of ancient documents. Archives conservation is a task conducted mostly manually by trained experts who could benefit from advances in automatic normalization and modernization of historical texts. Previous work has mainly focused on transforming the spelling of historical texts into their modern counterpart in order to apply computational tools (Bollmann, 2013; Pettersson et al., 2013; Sánchez-Martínez et al., 2013; Robertson and Goldwater, 2018). However, reducing orthographic variation while keeping historical spelling for a given era is the cornerstone of reliable diachronic studies. Yet, variations in wordforms lead to data scarcity and the lack of corpora containing spelling variants of historical texts is a serious impediment to supervised learning possibilities. Furthermore, spellings to retain among variants observed in historical texts may vary according to editorial guidelines, which is akin to a highly personalized natural language processing task.

Ledictz jour, vendredy 28 octobrix 1547, en l'Evesché
Ledit jour vendredi 28 octobris 1547 en l'Évêché
Said day Friday October 28 1547 in the bishop's house

L'on fasse respondre aut president de sadicte lectre
L'on fasse répondre au président de sadite lettre
We answer to the president about his letter

Ayme Richard, habitant et ferratier, filz de feu Thivent Richard, de Sonzier
Aimé Richard habitant et ferratier fils de feu Thivent Richard de Scionzier
Aimé Richard inhabitant and ironworker son of the late Thivent Richard of Scionzier

Figure 1: Segments sampled from our Middle French corpus in their original form (top, colored), normalized version (middle, in black, normalized words underlined) and English translation (bottom, *italic*). In the first sample, the bishop's house, translation of *Evesché* in this example, refers to the house inhabited by the previous bishop which was converted into a prison.

During the process of digitization and modernization of historical texts, researchers in digital humanities could benefit from the data produced during each steps of this process. In this work, we focus on the normalization without modernization of archives from the 16th century written in Middle French. To the best of our knowledge, there are no datasets of archival documents from this era and language containing non-normalized wordforms aligned to their normalized counterparts. We first manually normalize and align a set of historical texts with their spelling-normalized forms, keeping the syntactic and semantic content identical to the originally authored manuscripts. Second, we investigate how to automatically normalize these texts

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3394–3402

inspired by low-resource machine translation approaches. Empirical results show that two orthogonal solutions, namely sequence-level knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015; Kim and Rush, 2016) and language model transfer learning combined with back-translation (Marie and Fujita, 2021; Tonja et al., 2023), allow to produce reliable synthetic parallel corpora.

Our summarized contributions are: (i) we pave the way towards archival documents conservation with expert targeted normalization and spelling variants reduction, keeping historical wordforms without modernization, (ii) we show evidence that leveraging pre-trained models and synthetic data improves normalization performances over a strong baseline, (iii) we release a Middle French handcrafted parallel corpus aiming at orthographic normalization along with several fine-tuned models.[1]

The remainder of this paper is organized as follows. We briefly introduce the background work in Section 2. We describe the data production methods in Section 3, followed by our experiments and results in Section 4. Finally, conclusions and future work are presented in Section 5.

2 Background Work

This Section presents previous work on normalization and modernization of historical texts, followed by background work on synthetic data generation methods. Finally, we formalize our spelling normalization as machine translation approach.

2.1 Historical Text Normalization

The majority of previous work in normalizing historical text aims at modernizing the spelling of ancient documents, which usually contain inconsistent orthographic and syntactic variations. Various methods have been proposed to conduct this modernization task, including rule-based (Baron et al., 2009; Bollmann et al., 2011) and statistical approaches (Pettersson et al., 2013, 2014), and more recently using neural networks (Bollmann and Søgaard, 2016; Korchagina, 2017; Tang et al., 2018; Bawden et al., 2022). However, when normalization does not involve modernization, i.e. the task is to reduce spelling variations while keeping historical wordforms and syntactic structures, the lack of training data for supervised learning methods becomes a hurdle. Unsupervised approaches could potentially be applied to historical texts normalization but one requires large amounts of source and target corpora. It is thus crucial to produce corpora, manually or automatically, containing spelling variations as source and their normalized counterparts as target.

2.2 Synthetic Data Generation

Recently, methods towards producing artificial parallel training data for tasks such as machine translation have been explored, relying only on source or target texts. One of the most popular approaches is back-translation (Sennrich et al., 2016). It leverages large amounts of target-side monolingual data and translates them automatically into the source language. This leads to a parallel corpus where the source side, possibly noisy, does not impact the quality of the target side.

Due to the lack of large target corpora for the task of Middle French archives normalization, fine-tuning a generative model using a small amount of data appears to be an interesting and cost-efficient avenue to explore (Marie and Fujita, 2021; Tonja et al., 2023). Such a model is used to produce artificial target data which is then back-translated, leading to a parallel corpus usable for supervised learning of a normalization model. This assumes that a small amount of parallel data is already available to train a back-translation model.

An alternative technique to leveraging target side monolingual data is to make use of source documents and producing synthetic target text (Mittal et al., 2023) by relying on the forward-translation technique (Zhang and Zong, 2016; Bogoychev and Sennrich, 2019), also called self-training (He et al., 2019). If some target segments are available, segment-level knowledge distillation (Kim and Rush, 2016) was shown to improve a student model trained on synthetic data produced by a teacher model. We opt for segment-level knowledge self-distillation where the same model is the teacher and the student, the latter being trained on data produced by the former.

2.3 Normalization as Machine Translation

Assuming an available parallel dataset $\mathcal{D}$, its vocabulary noted $\mathcal{V}$ with $v_c \in \mathcal{V}$ ($0 \leq c \leq C$). Pairs of sequences compose the dataset with non-normalized source segments aligned to their normalized target counterparts, such as $\mathcal{D} = (X_1, Y_1), \ldots, (X_n, Y_n)$. The normalization task is formalized as the supervised neural ma-

[1] Data and models available at https://www.unige.ch/registresconseilge/en

Corpus	Segments	Source	Target
Manuscripts	59.9k	71.8k	–
Hand-crafted	1.6k	5.9k	4.6k
Distillation	59.9k	71.8k	58.9k
Generation	215.0k	472.4k	229.2k

Table 1: Number of parallel segments, source and target vocabulary sizes (k for thousands) for the hand-crafted and synthetic data produced in our study. The *Manuscripts* corpus is the manual transcription of the original RCs without normalization and is the source side of the *Distillation* corpus.

chine translation paradigm which aims at finding parameters θ of the conditional distribution $p(Y|X;\theta)$, maximizing the log-likelihood $\mathcal{L}_\theta = \sum_{(X_i,Y_i)\in D} \log p(Y_i|X_i;\theta)$. Training the neural network consists in minimizing the cross-entropy loss based on an input source sequence X_i, its corresponding target reference Y_i and the model output $\hat{Y}_i$. More formally, for every one-hot encoded token $y_j \in \mathbb{R}^C$ forming the sequence $Y_i = y_1, \ldots, y_{|Y_i|}$ $(1 \leq j \leq |Y_i|)$, the token-level loss is computed following (eq. 1), where $\hat{y}_j \in \mathbb{R}^C$ is the model output (logits) and y_j is the gold reference. By averaging token-level losses, we obtain the sequence-level loss following (eq. 2):

$$l_j = -\log \frac{\exp(\hat{y}_{j,y_j})}{\sum_{c=1}^{C} \exp(\hat{y}_{j,c})} \cdot y_j \quad (1)$$

$$\mathcal{L}_{(X_i,Y_i,\hat{Y}_i,\theta)} = \sum_{j=1}^{|Y_i|} \frac{1}{\sum_{j=1}^{|Y_i|} y_j} l_j \quad (2)$$

3 Parallel Data Production

Supervised training of an automatic normalization model is possible with a small amount of hand-crafted parallel data which we will describe in Section 3.1. We then present the two automatic data generation methods in Section 3.2 and Section 3.3. A set of non-normalized source documents were manually transcribed from the original manuscripts but do not have their normalized target counterparts. This non-parallel corpus was used in our distillation experiments. Details about the data produced, along with the number of segments and vocabulary sizes, are reported in Table 1.

3.1 Manual Normalization

The historical documents in our study are sourced from the publicly available Geneva Council Registers.[2] More precisely, we focus on the Geneva Council Registers in Calvin's time. They were written in Middle French from the 16th century including sections with mixed languages in Middle French and Latin, as illustrated in Figure 1. These manuscripts were written between 1536 and 1550. Their transcription requires the expert knowledge of historians and paleographers due to the vast number of patronymic, toponymic, geographical and generally era-related specialized vocabulary employed. The transcription of these documents was conducted manually over several years. Text normalization aiming at reducing spelling variants is still an ongoing work and will benefit from the use of computational tools.

To build a gold normalization dataset, we collected a small amount of source segments spanning over four years, from 1546 to 1549, which were manually normalized by experts according to the requirements and standards defined by the editors.[3] The normalization guidelines used to manually produce this hand-crafted dataset are described in Appendix A. The resulting curated and aligned pairs of segments were used as training and testing corpora to fine-tune and evaluate pre-trained encoder–decoder models, leading to our baseline normalization system. In addition, we fine-tuned the same pre-trained models with our parallel corpus in the reversed direction (normalized into non-normalized) to obtain models used for back-translation in the data production method described in Section 3.3. All the normalization models trained in our study follow the learning objective presented in (eq. 2).

3.2 Sequence-level Knowledge Distillation

Using the source documents manually transcribed from the original manuscripts (the corpus noted *Manuscripts* in Table 1) and the normalization system trained on our hand-crafted data, we produced a synthetic parallel corpus using the forward translation technique (Bogoychev and Sennrich, 2019). The source segments of our hand-crafted parallel corpus were thus manually and automatically normalized, leading to two sets of aligned target segments. In this particular scenario, we were able to perform knowledge self-distillation training, opting for the sequence-level approach

[2] https://ge.ch/arvaegconsult/ws/consaeg/public/FICHE/AEGSearch

[3] A segment contains sequences of various lengths, from a single token up to several sentences.

introduced by Kim and Rush (2016). Formally, based on the hand-crafted parallel corpus $\mathcal{D}$ and the sequence-level distilled corpus $\mathcal{D}_{KD}$, training of the encoder–decoder neural model is conducted by linearly interpolating the loss function presented in (eq. 2) with the loss calculated using the model output $\hat{Y}_i^{'} = \hat{y}_1^{'}, \ldots, \hat{y}_{|Y_i|}^{'}$ (with $1 \leq j \leq |\hat{Y}_i^{'}|$ and $\hat{y}_j^{'} \in \mathbb{R}^C$) as target reference (eq. 3):

$$\mathcal{L}_{(X_i, Y_i, \hat{Y}_i, \hat{Y}_i^{'}\theta)} = (1-\alpha)\, \mathcal{L}_{(X_i, Y_i, \hat{Y}_i, \theta)} + \alpha\, \mathcal{L}_{(X_i, \hat{Y}_i^{'}, \hat{Y}_i, \theta)} \quad (3)$$

where α is set to 0.5 following the empirical observation made by Kim and Rush (2016). The linear interpolation of losses is possible for segments pairs where both a target hand-crafted reference and a distilled target sequence are available. For segments in the distilled corpus $\mathcal{D}_{KD}$ which are not included in the hand-crafted corpus $\mathcal{D}$, a single loss is computed based on (eq. 2), replacing the reference Y_i by the distilled target $\hat{Y}_i^{'}$. To train our model noted *Distillation*, both datasets are used, namely $\mathcal{D}$ and $\mathcal{D}_{KD}$, the source side of $\mathcal{D}$ being included in the source side of $\mathcal{D}_{KD}$. The main intuition behind interpolating losses is to perform a smooth supervised continued training by making use of the network output given the D_{KD} source corpus. We assume that we could limit hand-crafted data over-fitting with this technique, while allowing the encoder to process a larger amount of source in-domain data.

3.3 Transfer Learning and Back-translation

Generative, decoder-only, neural causal language models allow to produce data based on a set of seed tokens, prompts or instructions. It was shown in previous work that a small amount of relevant target data could be enough to steer a pre-trained model towards a specific domain or genre through fine-tuning, allowing to produce artificial data of interest for a given task (Marie and Fujita, 2021). For archival documents normalization, we leveraged the target side of our hand-crafted parallel data to fine-tune a generative causal language model using the cross-entropy objective function with mean reduction. We then designed a set of seed sequences[4] based on the target side of the hand-crafted corpus and used it as inputs (i.e. prompts) to the model for text generation. No prompting template was employed. Sequences of tokens were fed to the generative model and the produced artificial text was then considered as target data. Finally, we used the back-translation model trained on the hand-crafted parallel corpus to generate the source-side of the synthetic parallel corpus. The resulting dataset was used as training material for our second model trained on synthetic data (noted *Generation*).

[4]The seed sequences, or prompts, contain between 8 and 12 tokens, as preliminary experiments showed more generation stability within this length range

4 Historical Documents Normalization

Models We trained individual models based on the three parallel datasets described in Section 3, namely *Hand-crafted*, *Distillation* and *Generation*. All normalization models were based on the pre-trained model *M2M100* with 418M parameters (Fan et al., 2021). The causal language model used to generate target synthetic data from prompts was *Bloomz* (Muennighoff et al., 2022) with 560M parameters.

The training procedure for normalization models using synthetic datasets followed a two-step process. First, continued training (Gururangan et al., 2020) with mixed synthetic and hand-crafted corpora. During this first step, we applied either sequence-level knowledge distillation or synthetic data generation followed by back-translation, leading to two models. Second, fine-tuning using the hand-crafted parallel corpus only, applied to each model resulting from the first step. We considered three baselines: pre-trained model without fine-tuning, copy of the source side of the test set (identity function), and a previously released model for early Modern French normalization (*ModFr*) (Bawden et al., 2022).

Training and Evaluation Due to the small size of our hand-crafted parallel corpus and the need to divide it into training, validation and test sets, all our experiments were based on a 5-fold cross-validation setup (train, validation and test splits represent approx. 63%, 17% and 20% respectively). We monitored the models performance during training based on the validation set with the BLEU metric (Papineni et al., 2002) and the best performing model for each fold was kept to normalize the test set. Finally, the 5-fold combined outputs were evaluated using four additional metrics, namely chrF (Popović, 2015), Translation Edit Rate (TER) (Snover et al., 2006), Word Error Rate

Model	BLEU ↑	chrF ↑	TER ↓	WER ↓	Acc ↑
Baselines					
M2M100	23.0	57.1	54.0	66.2	1.4
Copy	25.1	66.2	43.7	41.6	13.9
ModFr	32.9	71.4	37.7	37.6	13.6
Fine-tuned Models (based on M2M100)					
Hand-crafted	77.7	89.9	13.7	6.8	50.0
Distillation	78.7†	90.5†	12.8†	6.4	50.8
Generation	82.8†	93.0†	9.3†	5.1	52.7

Table 2: Test results averaged over 5 folds. Baselines are the identity function (*Copy*) and non fine-tuned models. Fine-tuned models use the hand-crafted corpus. *Distillation* and *Generation* are trained on synthetic data. Scores with † indicate statistically significant difference compared to previous rows ($p < 0.01$ with the approximate randomization test).

(WER) and segment-level accuracy.[5]

The five metrics used in our evaluation measure the performances of automatic normalization models at various granularities. More precisely, BLEU measures the n-gram precision (with $0 < n < 5$), chrF is the F-score at the character level, TER measures the translation edit rate at the word-level including shifts, WER is the word error rate without shifts and accuracy is the number of exact matching segments. The latter metric is interesting as it indicates how many segments are exactly matching the reference, thus reducing the manual revision requiring human experts. Results obtained on the test set are presented in Table 2.

Results The automatic normalization results show that existing pre-trained models do not outperform a naive baseline consisting in copying the source corpus in terms of segment-level accuracy. However, using the small hand-crafted parallel corpus to fine-tune *M2M100* leads to a 36.1pts increase in accuracy compared to the copy baseline, as well as a drop of 30.8pts of WER compared to the *ModFr* model. Significant improvements are further achieved by using both synthetic data production methods, with the best performing approach being the target data generation combined with back-translation (*Generation*).

This model outperforms the knowledge distillation model (*Distillation*) by 1.9pts accuracy and 1.3pts WER. We hypothesize that the *Generation* model is outperforming the *Distillation* model because it was trained on a larger corpus, but further experiments are required to support this assumption. The generation and back-translation approach is more computationally expensive, as it requires training a total of three models. We hypothesize that this approach has a larger potential of improvement in terms of normalization performances due to the wide arrays of prompt engineering possibilities (Liu et al., 2023). Again, further experiments will be conducted as future work to verify this hypothesis. Finally, all the pre-trained models used in our experiments are based on their smallest released version in terms of number of parameters. Thus, we assume that better performances could be reached with larger models.

Comparing the results obtained on the validation and test sets (cf. Appendix B and Table 2 respectively), the hand-crafted only setup shows a decrease of 3.6pts BLEU and 1.1pts accuracy between validation and testing, while the distillation setup shows 1.2pts BLEU drop and 0.2pts accuracy improvement between validation and testing. These results could confirm our initial intuition, where distillation limits over-fitting towards the small hand-crafted corpus, but more experiments should be conducted to draw solid conclusions. Overall, on the test set, the model trained with the interpolated loss outperforms the model trained on hand-crafted pairs only.

5 Conclusion

This paper paved the way towards archival documents conservation with expert targeted normalization and spelling variants reduction while preserving historical wordforms of 16th century Middle French. Compared to previous work in this field, the proposed methods do not modernize the historical spelling and syntactic structure but rather reduce the orthographic variability observed in originally authored texts. Our approaches were inspired by low-resource machine translation and leveraged pre-trained models to achieve significant performance gains compared to a strong baseline. Furthermore, our approaches are orthogonal and will benefit from collecting additional relevant corpora for fine-tuning and prompt engineering, which is one of our goals in future work. Another research direction considered as future work stems from the large amount of publicly available modern French data and pre-trained LLMs which could be further leveraged for our task.

[5]BLEU, chrF and TER implemented in SacreBLEU (Post, 2018), signatures: nrefs:1, case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1
case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1

Limitations

We recognize several limitations of this work.

First, the experiments were conducted on a variant of the Middle French language from 1536 to 1550. Middle French has evolved over time, from the 14th to the 17th century, and our work is considering a relatively narrow time frame in the history of this language.

Second, only a few pre-trained language models were tested during our preliminary experiments relatively to the large number of models currently publicly available. Some of these models were pre-trained on Modern or Early Modern French language, while other models were trained jointly on several languages, including languages relevant to our work such as Latin. Therefore, the models selected in our study may not be representative of all publicly released pre-trained models in terms of languages, number of parameters, training objectives nor architectures.

Third, the hand-crafted corpus produced in our work is relatively small in terms of number of tokens and vocabulary size compared to commonly used corpora in natural language processing experiments. This is mainly due to the high cost of producing such dataset for which the expertise of historians and paleographers is required, while following strict editorial guidelines.

Finally, we have not tried reducing the training data of the Generation approach to match the amount of data of the Distillation approach, thus we cannot draw conclusions on which approach is better.

Ethical Considerations

The dataset hand-crafted in our study is based on publicly available archives from the 16th century (non-license, public domain). We reviewed the content of the documents selected for manual normalization and we believe that this resource represents accurate historical events. However, some textual elements of this corpus could be considered as toxic and harmful, or disrespectful of the privacy of the people and places mentioned in these archives. We thus made sure that all data used in our work and to be released as part of our parallel datasets are in the public domain and already freely available. Consequently, no increased risks or harm is caused by our dataset. Instead, it serves as a resource for historical studies and digital humanities.

The fine-tuned models to be released with our work are based on publicly released and licensed pre-trained models (MIT License). We respect the permissions to use, modify and distribute the models. We release the fine-tuned models under the MIT License.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments.

This work was partially funded by the FNS (Fonds national suisse), SSH 2022 grant n. 215733, for the project entitled *Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (15 45-1550)*, acronym *RCnum*.

All experiments were conducted on the University of Geneva computing cluster HPC *Baobab* and *Yggdrasil*.

References

Alistair Baron, Paul Rayson, and DE Archer. 2009. Automatic standardization of spelling for historical text mining. *Proceedings of Digital Humanities 2009*.

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.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Marcel Bollmann. 2013. Pos tagging for historical texts with sparse training data. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 11–18.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42, Hissar, Bulgaria. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 535–541.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Natalia Korchagina. 2017. Normalizing medieval german texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 12–17.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Benjamin Marie and Atsushi Fujita. 2021. Synthesizing monolingual data for neural machine translation. *arXiv preprint arXiv:2101.12462*.

Sarthak Mittal, Oleksii Hrinchuk, and Oleksii Kuchaiev. 2023. Leveraging synthetic targets for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9365–9379, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th workshop on language technology for cultural heritage, social sciences, and humanities (LaTeCH)*, pages 32–41.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An smt approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, volume 18, pages 54–69.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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.

Alexander Robertson and Sharon Goldwater. 2018. Evaluating historical text normalization systems: How well do they generalize? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 720–725, New Orleans, Louisiana. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Isabel Martínez-Sempere, Xavier Ivars-Ribes, and Rafael C Carrasco. 2013. An open diachronic corpus of historical spanish: annotation criteria and automatic modernisation of spelling. *arXiv preprint arXiv:1306.3692*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

A Appendix: Normalization Guidelines

The normalization guidelines were defined by the historian in charge of manually normalizing RC content. This person is an expert in 16th century Middle French, in the Geneva region and in the political landscape in Calvin's time. The normalization applied to the source textual content is focused on local orthographic and grammatical elements while leaving syntactic structures unchanged. The guidelines were the following:

- First characters are uppercased at the start of sentences, but also for patroyms and toponyms.
- Limit the use of ponctuation marks:
  - semicolons in lemmas only to separate different items,
  - commas before decisions, e.g. (regarding) ordered/stopped/solved,
  - periods at the end of sentences.
- Use of diacritical marks (apostrophes) except for cases where *que* is followed by a vowel which in fact are *qui*, e.g. *sont survenues quelques lettres que attouchaient à Gen ève* (in English: a few letters about Geneva appeared).
- Extended emphasis and accentuation based on modern usage
- Gender and number of past participle agreement, e.g. *de celui qui les a baillé* becomes *de celui qui les a baillés*, *sus la supplication qui a présenté* becomes *sus la supplication qui a présentée*, except when there is a doubt such as *lui soit baillé trois écus* **not** to be corrected in *lui soient baillés trois écus* because it is an ambi guous case: *trois écus* could be the object or the subject.
- Verb agreement, e.g. *ordonné que lesdits six écus lui soit délivrés* becomes *ordonné que lesdits six écus lui soient délivrés* (in English: ordered that the said six écus be deliver ed to him)
- Modernisation of patronyms, first names and toponyms.
- Correction of genders according to modern usage, e.g. *la dimanche* (in English: the Sunday) becomes *le dimanche*, *la reste* (in English: the rest) becomes *le reste*.
- Singular feminine possessive determiner replacement, e.g. *ma* (my), *ta* (your), *sa* (his, her, their), for nouns starting with a vowel or with a silent *h*, by the ma sculine forms *mon*, *ton*, *son*. For instance, *à sa humble requête* becomes *à son humble requête* (in English: to his/her/their humble request).

Model	BLEU ↑	chrF ↑	Acc ↑
Hand-crafted	81.3	91.5	51.1
Synthetic Data			
Distillation	79.9	91.1	50.6
Generation	80.4	91.9	39.3
Synthetic + Hand-crafted Data			
Distillation	80.2	91.3	50.4
Generation	83.5	93.4	52.4

Table 3: Normalization results on the validation set averaged over 5 folds. Models under *Synthetic Data* were not finally fine-tuned on the hand-crafted data.

B Appendix: Ablation Study

Results obtained on the validation set with various models trained during our study are presented in Table 3.

C Appendix: Training Procedure

All pre-trained models used in our work are checkpoints released with *HuggingFace Transformers* (Wolf et al., 2020). All models were trained on single Nvidia RTX A5000 and 3090 GPUs with 24GB memory. Training was conducted for a maximum of 100k steps with early stopping based on the BLEU scores obtained during validation. All our code was implemented in *Pytorch* (Paszke et al., 2019) and we used the *AdamW* optimizer (Loshchilov and Hutter, 2017). We used batch sizes between 4 and 16 segments depending on training and testing phases. Hyper-parameter search was focused on the learning-rate, with values ranging between $5e^{-5}$ and $5e^{-7}$. All other hyper-parameters were kept as default according to the configuration files released with the checkpoints. The final batch size was set to 4 segments and the best learning-rates for each model and each fold (1 to 5) are the following:

- *Hand-crafted*
  1. $5e^{-6}$
  2. $2e^{-6}$
  3. $5e^{-6}$
  4. $2e^{-6}$
  5. $2e^{-6}$
- *Distillation* (continued training – fine-tuning)
  1. $5e^{-6}$ – $2e^{-6}$
  2. $5e^{-6}$ – $1e^{-6}$
  3. $1e^{-5}$ – $2e^{-6}$
  4. $5e^{-6}$ – $2e^{-6}$
  5. $1e^{-5}$ – $2e^{-6}$
- *Generation* (continued training – fine-tuning)
  1. $1e^{-5}$ – $1e^{-6}$
  2. $5e^{-6}$ – $2e^{-6}$
  3. $8e^{-6}$ – $1e^{-6}$
  4. $5e^{-6}$ – $2e^{-6}$
  5. $8e^{-6}$ – $2e^{-6}$