Context-Aware Text Normalisation for Historical Dialects

Maria Sukhareva
Applied Computational Linguistics Lab (ACoLi)
Goethe University Frankfurt
sukhareva.maria@gmail.com

Abstract

Context-aware historical text normalisation is a severely under-researched area. To fill the gap we propose a context-aware normalisation approach that relies on the state-of-the-art methods in neural machine translation and transfer learning. We propose a multidialect normaliser with a context-aware reranking of the candidates. The reranker relies on a word-level n-gram language model that is applied to the five best normalisation candidates. The results are evaluated on the historical multidialect datasets of German, Spanish, Portuguese and Slovene. We show that incorporating dialectal information into the training leads to an accuracy improvement on all the datasets. The context-aware reranking gives further improvement over the baseline. For three out of six datasets, we reach a significantly higher accuracy than reported in the previous studies. The other three results are comparable with the current state-of-the-art. The code for the reranker is published as open-source[1].

1 Introduction

Historical languages rarely have enough resources to satisfy the needs of state-of-the-art supervised machine learning methods. Manual labeling of large amounts of training data is frequently out-of-the-question as it demands rare and expensive expertise of historical linguists. Historical dialects, however, are frequently grammatically and lexically similar to their modern high-resource descendants. Thus, being able to normalise the historical spelling to the modern standards can enable processing of historical dialects with NLP tools trained on modern languages.

Modern high-resource languages are well-standardized and have prescriptive grammar, spelling and lexica usage standards. Historical languages, on the contrary, did not have such standardization. While modern German also has a plenty of dialects, there exists its standardized variety i.e. Standard High German. Middle High German (MHG), on the contrary, does not have contemporary standardization. When we deal with a historical text, we deal with a historical dialect (e.g. Middle High German, Middle Low German) that text is written in. Even more, there is a plenty of variability within the dialects caused by the lack of orthographic standards, geographical differences, the exact time when a text was written etc. The normalisation task presupposes that there exists a standard to which the historical variants are normalised. As historical languages do not have such a standard, the closes possible goal for historical dialect normalisation would be their modern standardized descendants.

Throughout centuries, languages undergo systematic phonological changes such as the Great Vowel Shift in English, the first and second consonant shifts in Germanic dialects or the first and second palatalization in Slavonic languages. This can be captured systematically by machine learning algorithms and applied to unseen words. Thus, the current state-of-the-art approaches to the historical normalisation rely on statistical or neural machine translation methods and define the task as a problem of translating between characters or substrings (Mansfield et al., 2019) instead of words.

Historical texts pose an additional challenge of non-standardized orthography. Spelling variations can depend on the region, authorship and time of writing. These kinds of irregularities complicate the

[1] https://github.com/ktoetotam/contextaware-reranking

Proceedings of the 28th International Conference on Computational Linguistics, pages 1023–1036
Barcelona, Spain (Online), December 8-13, 2020

training of a normaliser as it fails to unambiguously align the modern word to its historical form. In this paper, we show that providing the dialectal context during the training disambiguates the alignment between the historical and modern words and, thus, leads to a significant improvement of the normaliser's performance.

Most of the research on historical text normalisation has focused on context-agnostic approaches that do not consider the immediate surrounding of the word. However, the context-agnostic approaches do not have an efficient method to deal with homographs, polysemy and homonyms (e.g. the Middle High German (MHG) word *in* can be normalised as a modern German preposition *in* or a pronoun *ihn*, an accusative form of the pronoun *er "he"*). As a context-agnostic neural normaliser cannot disambiguate between them, we hypothesise that all the valid normalisations are expected to appear among the top five normalisation results. We have tested this hypothesis and confirmed that a correct normalisation is found among the top five candidates for 96.22% to 98.83% of words depending on the dataset. We propose a reranking method by applying a word-based n-gram language model to the top five normalisation candidates. We show that reranking improves the word accuracy and delivers state-of-the-art performance.

2 Related work

The early work on text normalisation relied on rule-based methods (Pettersson et al. (2012); Bollmann et al. (2011)) and dictionary lookups (Rayson et al., 2005). Reynaert et al. (2012) compare the performance of two statistical normalisation tools on Portuguese and arrive at the conclusion that context-agnostic normalisation can only achieve limited results. Later, statistical machine translation (SMT) methods became the state-of-the-art. The normalisation problem was defined in the SMT context as a problem of translation between characters which maximizes the probabilities of $P(f|e) \times P(e)$ with $P(f|e)$ being the probability that the word f is the historical variant of the modern word e and $P(e)$ is the probability of observing the normalised form in the modern language. Pettersson et al. (2014); Jiampojamarn et al. (2007); Nakov and Tiedemann (2012); Schneider et al. (2017) have all successfully applied character-based SMT to various normalisation tasks.

As neural machine translation has gained momentum in the past few years, the neural normalisers have also started producing state-of-the-art results. Tang et al. (2018) compares SMT-based methods to sequence-to-sequence models and concludes that NMT methods perform consistently better than SMT. Bollmann (2019) also shows that NMT methods perform consistently better than SMT under the condition that enough data is available.

Joint training and multitask learning have long been successfully used for various NLP applications. Bollmann and Søgaard (2016) introduced a multitask learning approach to historical German normalisation. They trained a bidirectional LSTM network on the whole Anselm corpus and retrained its prediction layer on each text separately which led to an improvement over the baseline. Bollmann (2018) conducted a detailed evaluation of various setups of joint and multitask training. Pairwise dataset combination showed that multitask learning on two datasets of different languages (e.g. German and Slovene) did not give a consistent improvement but the common trend was that closely related datasets benefit the most from the multitask training. The joint training setup with no indication to which language the dataset belongs by Bollmann (2018) had mostly a negative effect on the normalisation accuracy.

The general MT research gets more interested in context-aware normalisation. Voita et al. (2019) propose an decoder architecture that takes into consideration previously translated sentences while generating new texts. Zhang et al. (2018) show that integrating document level information helps to improve French and Chinese translation into English.

There is still not much work done in the area of context-aware normalisers. Mansfield et al. (2019) proposed to use sequence-to-sequence models to normalise full sentences for conversational systems. Jurish (2010) proposed to use hidden markov models to choose over the normalised candidates in a sentential context. Mitankin et al. (2014) used a language model for candidate correction of an unsupervised normaliser.

Dataset	Tokens			tag	Dataset	Tokens			tag
	Train	Dev	Test			Train	Dev	Test	
DE:					**ES:**				
Anselm	233,947	45,996	45,999	dea	16^{th}	6,668	785	843	ps16
Ridges	41,857	9,712	9,587	der	17^{th}	20,175	2,578	2,457	ps17
SL:					18^{th}	49,033	5,932	6,402	ps18
Bohoric	50,023	5,841	5,969	boh	19^{th}	21,444	2,355	2,777	ps19
Gaj	161,211	20,878	21,493	gaj	**PT:**				
					16^{th}	14,781	1,820	1,840	ps16
					17^{th}	47,680	5,987	5,607	ps17
					18^{th}	78,232	8,932	9,654	ps18
					19^{th}	81,833	10,012	9,977	ps19

Table 1: The historical multidialect dataset used for the experiments and the corresponding tags for the joint training.

3 Data

Historical languages did not have universally accepted prescriptive grammar or standardized orthography and, thus, spelling varies depending on the authorship, region or time when the document was written. Some historical languages could have even utilized various alphabets (e.g. Slovene) depending on the region or time. Spelling variations can also be found even within the same document.

For our experiments, we have used an openly available normalisation dataset[2] by Bollmann (2018). We selected the languages that have regional or diachronic dialects in the dataset. Thus, we used four historical languages (German, Slovene, Spanish and Portuguese).

Table 1 shows the statistics over these datasets. German is comprised of two datasets both written in the 16^{th} century with Anselm corpus being a collection of religious manuscripts and Ridges - a collection of scientific botanic articles with an abundance of specialized terms. The Slovene datasets are comprised of two collections: texts written before 19^{th} century in the Bohoric alphabet and texts written afterwards in the Gaj alphabet. Portuguese and Spanish datasets are a collection of letters written between 16^{th} and 19^{th} century.

The dataset includes the following languages: German, Hungarian, Icelandic, Portuguese, Slovene, Spanish, Swedish. The German, Portuguese, Slovene and Spanish datasets are not homogeneous and are comprised of several diachronic and geographical varieties of the corresponding languages that are normalised to its modern standard variation.

4 Multidialect text normalisation

Previous work has shown that multitask and joint training setups are particularly useful for closely related datasets. The historical texts are a perfect testbed for multitask and joint setups: The spelling varies depending on the dialect, region, time of writing or even the author. While many of the previous normalisation approaches do not incorporate this information into the training, we assume that the dialectal context can disambiguate homographs and help a normaliser to learn region- and time-specific diachronic spelling changes. In this section we show that providing a neural normaliser with dialectal context leads to significantly [3] better performance over all the datasets.

To incorporate the dialectal information into the normaliser, we follow a standard approach to multitask learning used in multilingual machine translation (Johnson et al., 2017): After splitting the input words into characters we modify the input by adding a tag that identifies the word's historical dialect. Table 1 provides the lists of the used tags. For training the normaliser we use a toolkit for neural sequence-to-sequence transduction *Sockeye* (Hieber et al., 2017). All the encoder-decoder transformer models had

[2]https://github.com/coastalcph/histnorm

[3]From here on "significantly" is used exclusively in statistical sense i.e. $p < 0.05$ as computed by McNemar's test

Language	S	J	J_t	J_{tb}
DE:				
Anselm	89.02	89.1	89.27*	**89.29**
Ridges	86.33	83.65	87.97	**89.01**
SL:				
Bohoric	92.2	92.55	92.79	**93.29**
Gaj	95.57	95.56	95.79*	**95.81**
ES:	-	93.79	**94.54**	94.12
PT:	-	94.47	94.96*	**95.02**

Table 2: Evaluation of the multidialect models. S - separate training; J - joint training, J_t - joint training with tags, J_{tb} - joint training with tags on a balanced dataset. *McNemars test at p < 0.05 as compared to the best result.

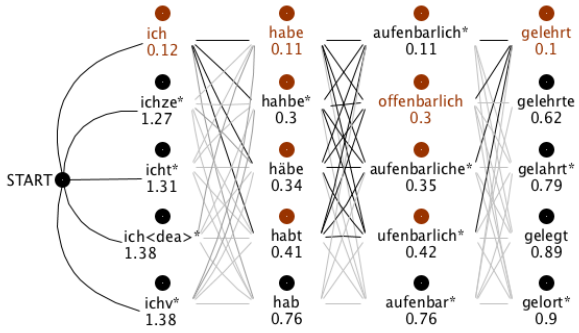

Figure 1: The beam search through the top five normalisation candidates. The scores are the negative log probabilities assigned by the normaliser. The words in red are the correct normalisations while the red dots mark all the words that are left after the applied threshold of $\leq -log(P(w_1)) + 0.4$.

six hidden layers and were trained on 50 epochs with 0.1 dropout[4].

During the training we observed that balancing the data can play a crucial role in the normaliser's performance. Thus, we ensured that all the dialects are equally represented in the training and in the validation by means of oversampling. This means that if a dialect has five times as few data as the dialect with the largest amount of data, each normalisation pair of the underrepresented dataset will appear five times in the training data.

Table 2 shows the evaluation results for the joint training. The metric that was used is normalisation accuracy i.e. the percentage of the wordforms that were normalised correctly. The largest improvement over the baseline is achieved on the Ridges corpus. The joint training without tags results are consistent with the results shown by Bollmann (2018) who reported a decrease in performance in this scenario. In our experiment, the word accuracy also dropped by almost 0.03. Nevertheless, providing the dialectal information through tags immediately gave an improvement of 0.16 and oversampling the Ridges corpus for data balancing resulted in a further improvement of 0.14. The accuracy consistently improved for the low resource dialect on the Slovene dataset as well: the accuracy on balanced joint training with tagged data is 0.11 higher than for a model trained and tested on the Slovene Gaj only. As the Spanish and Portuguese datasets contain very few data for each dialect, we omitted training single models and evaluated the performance of joint models only. The results have proven again that adding dialectal information significantly improves the performance. However, the balancing of the Spanish datasets had a negative effect on the accuracy.

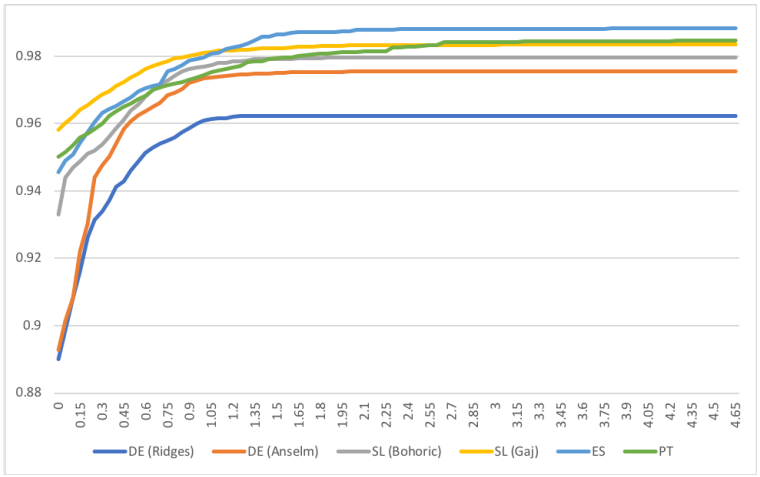

Figure 2: The vertical axis is the percentage of source words that have at least one correct normalisation among the candidates within the corresponding threshold $-log(P(w_1)) + d$ with d being the coordinates on the horizontal axis.

5 Context-aware reranking

The weakness of context-agnostic text normalisation is its inability to differentiate between the homographs, homonyms and polysemy. Example 1 from the Anselm normalisation dataset shows a MHG sentence with a homograph *das* that can be either normalised as a modern German (DE) demonstrative pronoun *das* (*"this"*) or a relative pronoun *dass* (*"that"*). Context-agnostic normalisation can only produce the exact same normalised form for both words. Yet, the correct normalisation can usually be still found in the top five candidates. If the normaliser has observed multiple homographs in the source training data for which correct normalised forms differ, the normaliser's uncertainty about the output is very high which leads to it assigning nearly equal probabilities to both normalisations.

(1) **das** irhorte eyn iude **das** myn lybes kynt...
das erhörte ein jude **dass** mein liebes kind...
A jew heard it that my dear child...

Thus, we propose a context-aware reranking of the normalisation output. We borrow the idea from the statistical machine translation that rescores the translation probabilities with the probabilities of observing a n-gram in a large monolingual corpus.

For the context-aware reranking, all the multidialect normalisation models were trained as described in the previous section (sec. 4) and balanced accordingly with the only exception of Spanish for which the best performing model was trained on an unbalanced dataset. Furthermore, we train an n-gram model on the monolingual target corpus and apply a beam search to the top five normalisation results.

5.1 Language model

A five-gram language model is trained on the preceding and succeeding context of the word. The n-gram probabilities are computed in a forward and in a backward pass. In case of unknown words we use the α-smoothing with the α being the probability of observing a unigram only once in the corpus divided by two. This smoothing strongly penalizes unseen words. For each dataset, the language model is trained on the training part of the corresponding normalisation datasets and a bible. We have tried to train the language model on a larger amount of data i.e. adding more bible translations when available as well as news corpora, however, this did not give any improvement over the results. This can be explained by the fact that the word order on the normalised dataset follows the rules of the historical dialect that is different from the patterns observed in the modern data originally written in the target language.

[4]See appendix for the exact command

Not all the datasets kept the sentence segmentation. In particular, the Anselm dataset does not have sentence segmentation and, thus, we did not add any start and end symbols and just learnt the n-grams over the whole corpus.

5.2 Beam search

We implement a beam search over the top five normalisation candidates produced by the normalisers. The beam's width is set to 10. If a dataset does not have the sentence segmentation, the language model is applied to the sequences of ten words. The ten word threshold is motivated by several reasons. First of all, it is computationally efficient. We tried longer sequences and did not see any improvement of the accuracy. As the sentence length depends on the genre, sentence definition and can even vary depending on the author, we chose the lowest boundary of the average sentence length in German as defined by Best (2002). Figure 1 shows the beam search graph where all the five top normalisation results are interconnected. Yet, the context-agnostic multidialect normalisation assigns the correct normalised form in the absolute majority of cases (see Tab. 2). Computing the beam search over all the five top normalisation candidates is computationally expensive and can lead to introduction of avoidable errors. We have made an observation that in the cases when the normaliser does not rank the best translation as top, the negative log probability of the correct translations is close to the negative log probability of the top-ranked normalisation.

The beam search applied to the Figure 1 reranks the normalisation results of the following phrase:

(2) ich habe offenbarlich gelehrt
I have openly taught
MHG: ich habe vffenbarlich gelahrt..

The normaliser has ranked the results by the negative log probability with the lowest being the best result. The more uncertain the normaliser is about the correct normalisation, the closer the scores are to each other. For example, the negative log probability difference between the frequent word *ich "I"*, which is correctly normalised, and the second best candidate *ichze*, that happened to not be a valid German word, is relatively large. On the contrary, the score gap between the incorrectly first ranked normalisation of the MHG word "vffenbarlich" and the consequent top four candidates is much more narrow. In general, such a narrow gap between the top normalisation candidates is a strong indication that the top ranked normalisation might be incorrect. Therefore, to save computational efforts and prevent error propagation, we introduce a threshold for the normalisation results to be considered by the beam search. On Figure 1, the red knots mark all the graph points that are below or equal to $-log(P(w_1)) + d$ where w_1 is the first ranked normalisation candidate and $d = 0.4$ is the threshold's window. The dark lines are the paths that the beam search computes for this threshold window.

The language model is applied in the following way: We compute n-grams up to the 5^{th} order. However, given that our monolingual corpora only consist of a bible and the target-side of the training data, five-grams are very sparse and context sensitive. In order to resolve this, we apply interpolation to lower n-grams sensitivity i.e. we are also computing all the lower order n-grams and average over their probabilities. Let us define a n-gram w_1^n as a string of $w_1, ..., w_n$. Thus, a notation for a unigram would be w_1^1, for a bigram w_1^2 etc. The probability of $P(w_1^n)$ is then:

$$P(w_1^n) = \frac{1}{n} \sum_{i=1}^{n} P(w_n | w_{n-1}^{n-i})$$

Thus, we can compute the probability of a sequence $w_1, ..., w_m$ where m is the length of the sentence s or a phrasal unit.

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i^{i+n})$$

Not only preceding but also succeeding context is important for estimating a probability of observing a word and, therefore, we also compute a backward pass in the same manner as the forward pass. The

Lang	Acc
DE_a	97.54
DE_r	96.22
SL_b	97.5
SL_g	98.08
ES	98.83
PT	98.49

Table 3: The percentage of the historical words in the test sets for which a correct normalisation can be found within the five top candidates proposed by the normaliser.

final probability of a sentence s is bidirectional and is computed as:

$$P(s) = P(w_1, ..., w_m) \times P(w_m, ..., w_1)$$

The $P(w_m, ..., w_1)$ is just a probability of the backward pass computed as the probability of the forward pass that is described above.

6 Evaluation

In order to facilitate the beam search and to exclude error propagation, we use a threshold over the negative log probability assigned by the normaliser. Figure 3 shows the percentage of the correct normalisations found among the top five results filtered within a given threshold window. The threshold is defined as $-log(P(w_1)) + d$ i.e. the sum of the negative log probability assigned to the first-ranked candidate and a variable d that determines the threshold's window (the horizontal axis of Figure 3). The vertical axis shows how many correct normalisations are found within the given window i.e. if at least one candidate is correct. The graph shows that most of the correct normalisations can be found within a narrow window of $d < 0.75$ and a steep surge is observed at the d between 0.2 and 0.3. This is an expected observation as neural normalises typically assign closer scores to uncertain outcomes. Having made this observation, we propose a context-aware reranking method for the normalisation results.

All the language models were trained on the target side of the normalisation datasets and on a bible acquired from the OPUS parallel corpus[5]. We downloaded the PT, ES, DE and SL bibles aligned to an English bible so that all the bibles have the same amount of verses. We have attempted to train larger language models on news corpora but they did not perform better than the bible-based LMs. This is likely due to the fact that normalised historical texts have different syntax and lexica than news corpora. This assumption is also confirmed by the fact that we achieve a very significant improvement over the multidialect and state-of-the-art (SOTA) baselines for the German Anselm dataset which is a collection of religious texts. The improvement was much lower for the botanical texts of the German Ridges corpus. Thus we conclude that having an in-domain monolingual corpus is extremely beneficial for this approach.

We studied the effect of the threshold window by applying the language model to rerank the normalisation results through beam search as described in the previous chapter. Figure 1 shows the relations between the normaliser's accuracy and the window size. For all the datasets except for Spanish, the accuracy significantly improved over the baseline starting from the window set to 0.15 and started to deteriorate at 0.3. The best accuracy for the Spanish, Portuguese, Slovene and German Ridges datasets was achieved with the window value of 0.2. The best results for the German Anselm dataset was achieved with the window set to 0.25. This is also the only dataset for which the accuracy for the threshold of 0.2 is significantly lower than for the reranked results over the window of 0.2. Table 7 shows the detailed evaluation of the language model with the two windows and its comparison with the SOTA[6].

For the SOTA comparison, we used the results reported by bollmann-2019-large on the same datasets. The normalisers that achieve the best results on the dataset were: a neural normaliser by tang-etal-2018-evaluation and a character-based normaliser enhanced with a language model trained on a large amount

[5]http://opus.nlpl.eu/

[6]the full evaluation is found in Appendix

	DE_r	DE_a	SL_b	SL_g	ES	PT
SOTA	88.22‡	89.64‡	93.3†	**96.01**†	**95.02**†	95.18†*
MD	89.29	89.01	93.29	95.81*	94.54	95.02
LM ($d < 0.2$)	**89.84**	90.95	**94.11**	95.85*	94.81	**95.26**
LM ($d < 0.25$)	89.64*	**91.81**	94.01*	95.76*	94.88*	95.25*

Table 4: The historical multidialect dataset used for the experiments and the corresponding tags for the joint training and comparison with SOTA: † denotes cSMTiser+LM and ‡ denotes NMT (Tang et al., 2018). The * means that the results are not significantly different from the best result.

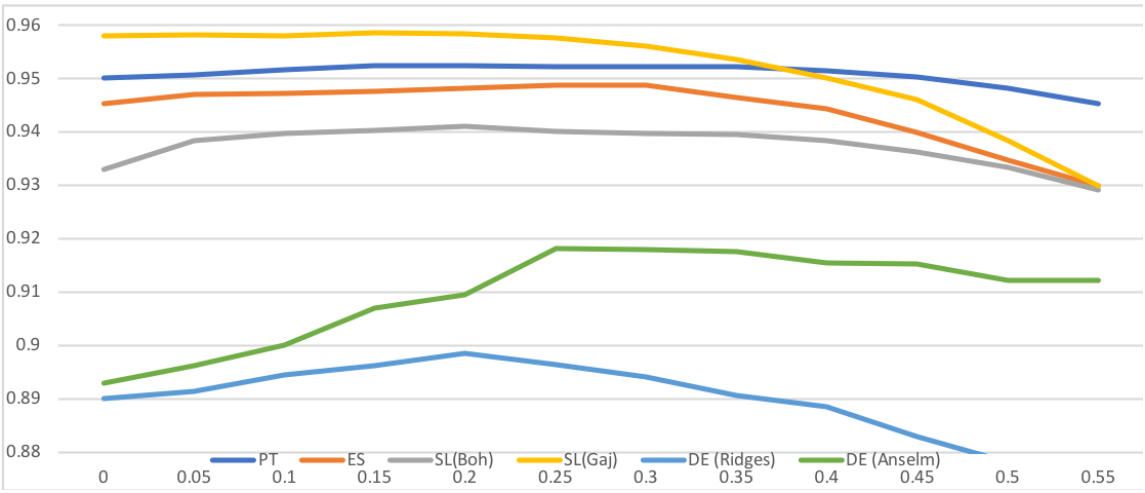

Figure 3: The normalisation accuracy (the vertical axis) changes depending on the threshold's window (the horizontal axis). The larger the window, the more candidates are considered.

of monolingual data by ljubevsic2016normalising, bollmann-2019-large. Further on, we will also refer to them as NMT (Tang et al., 2018) and $cSMTiser_{+LM}$ respectively.

For three datasets we achieved a significant ($p < 0.05$) improvement over the reported state-of-the-art. Also for all the datasets except for Gaj Slovene applying the context-aware reranking leads to a significant improvement over the multidialect normaliser. The greatest improvement over the SOTA baseline is observed for the German Anslem dataset with $+2.2$ to the accuracy. The balanced multidialect joint training also resulted in the significant improvement over the SOTA ($+1.07$) for the German Ridges dataset and the application of the reranking further significantly improved the accuracy by $+0.55$. The context-aware reranking also led to a significant accuracy improvement as compared to SOTA and to the multidialect normaliser for Slovene Bohoric. There is no significant difference between the performance of the models on the Gaj Slovene and Portuguese as compared to the SOTA. The reranking, however, has significantly improved the accuracy for the Portuguese as compared to the multidialect baseline. The Spanish dataset was the most challenging for the multidialect normaliser and while the reranking significantly improved the results as compared to the multidialect setup, the final accuracy score was still insignificantly below the SOTA.

7 Qualitative evaluation of the results

The evaluation showed that the largest improvement was achieved on the German datasets and Slovene Bohoric while the performance on the Slovene Gaj, Spanish and Portuguese stayed the same as SOTA. The German texts are older, the spelling is more chaotic and more distant from the modern standard. Table 5 shows that the German datasets and Slovene Bohoric have significantly smaller identity matching score. Non-standardized spelling in German Anselm is particularly problematic: The historical part of the dataset has 3.76 times as many wordforms as its modern counterpart. Some words have tens of historical spelling variations e.g. *kreuz 'cross'* has 79 wordforms, *zu 'to'* 46, *gott 'god'* 11 etc. This

Lang	¬match	ratio
DE_a	69.37%	3.76
DE_r	55.72%	1.35
SL_b	56.2%	1.3
SL_g	13.92%	1.12
ES	26.58%	1.36
PT	34.38%	1.66

Table 5: ¬match column shows the percentage of the historical and modern wordforms that are not identical in the datasets. The ratio column is the $\frac{|w_h|}{|w_m|}$ where $|w_h|$ and $|w_m|$ are the counts of unique historical and modern wordforms correspondingly

ambiguity increases the uncertainty of the context-agnostic normaliser and increases the likelihood that the correct normalisation appears in the top results but is not top ranked. This explains why a correct normalisation was found in 97.54% of cases among the top five normalisation results but only 89% are ranked as top (see table 3 and 2). We assume that this property will hold for all the dialects that have low orthographic standardisation. While it can correlate with the diachronic distance: the older the dialect, the less standardized it is, it is not necessarily the case and we believe that it will also hold for modern dialects if they do not have standardized spelling. Thus, we propose a general recommendation to use the context-aware reranking of top normalisation results for normalisation tasks of data with non-standardized orthography. Reranking corrects such errors by choosing words that were observed in the modern language within the given context. We believe that this explains why we saw the best results on the most complicated datasets. We believe that this is an important observation that shows that integrating contextual information is particularly relevant for texts that have many spelling variations. It is also confirmed by the results on Slovene Gaj, Spanish and Portuguese datasets that are very close to the spelling of its modern varieties, it is also more standardized and has less variations. Thus, the performance of the context-agnostic normaliser is already very good as it is not challenged by the data sparsity and ambiguity seen in the other three datasets and, therefore, there is hardly any room for improvement in the first place.

In order to understand why adding the German side of WMT corpus[7] to the data does not help to improve the performance, we analysed frequent corrections that were made by the language model on the German datasets. A qualitative analysis by a German speaker showed that the vocabulary and syntax of the normalised side of the training data contains many archaisms and set expressions and syntax follows the rules of MHG rather than modern German. For example, *liebes "dear"* is normalized by the context-agnostic normaliser as *leibes "bodily"* in *liebes kind "dear child"*. The phrase occurred 0 times in WMT and 190 times in German Anselm training set. Another common correction in German Anselm is *sandte 'sent'* changed to *"sankt" saint*. The correction occurred 116 times in the testset. In the WMT corpus "sankt" occurs exclusively as a part of geographic locations and never in the same context as in the German Anselm corpus (names of saints).

Not only historical spelling variations, but also modern homographs pose a challenge that context-agnostic normalisation cannot overcome. The top ranked incorrectly normalised wordforms in the German Anselm (see Table 3 in Appendix), that were corrected by the reranker, are all valid German words and are incorrect only in the given context. Thus, the context-aware reranking deals exactly with the errors that go beyond the capabilities of context-agnostic approaches. For example, MHG "das" can be normalised as "das" an article "the" and "dass" as a relative pronoun "that". The reranker changed "das" to "dass" 576 times. Appendix contains the full list of top corrections made by the model. We believe that adding further out-of-domain modern data are unlikely to improve the results while adding in-domain data can be helpful. However, the true in-domain data would be word by word normalised historical texts which are extremely hard to acquire. The next possible approximation of in-domain data are religious texts that contain a lot of archaisms.

[7]https://www.statmt.org/wmt15/translation-task.html

The German Ridges corpus poses an additional challenge of abundance of botanical terminology: We have calculated that 90% of words, that were assigned the incorrect wordform by the reranker, occur only once in the test set and are mainly botanical terms that are not found in the training set. As these terms also are neither found in the WMT corpora, it again explains why adding more modern data did not result in any improvement.

We have examined the effect of the window in which the candidates are considered for the beam search and observed that the model is very sensitive to the error propagation. In particular, allowing a large window or considering all the top five normalisation results irrespective of their score decrease the accuracy. This mostly happens when the lower ranked candidates are valid words that are frequently observed in this contexts (e.g. "spoke" and "speaks"). The normaliser would assign a much lower probability to a lower ranked normalisation but the language model would have an approximately equal n-gram counts for the both of them. Thus, it is important to determine the correct window. The general observation is that if the model is trained on in-domain data, the window can be slightly larger than if the model is trained on out-of-domain data. This is probably due to the fact that it better covers domain-specific lexica.

The only dataset for which the context-aware reranking did not result in significant improvement over the multidialect model is the Gaj Slovene dataset. Figure 2 gives a clue why this is the case: the Slovene dataset plateaus faster than any other dataset and does not have a steep surge near the window of 0.3 like the German datasets and Slovene Bohoric. Thus, the possible improvement can only be rather limited and enlarging the window would cause the error propagation as described above as too many correct first-ranked normalisations would be unnecessary reranked by the language model which would eliminate the benefit of its usage in the first place.

8 Conclusion

We have presented a reranking method for a multidialect normaliser. We have shown that neural historical normalisers perform better when trained on the data with explicitly provided dialect information. The context-aware reranking of the normalisation results implemented by means of a statistical n-gram model has resulted in improvement over the baseline and performs comparable or better than the state-of-the-art results reported in previous studies. We will publish all the code used in this study as open-source.

Thus, future work will focus on exploring more elaborated methods of context-aware reranking i.e. comparison with neural language models.

References

KH Best. 2002. Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft*, 7:7–33.

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, December. The COLING 2016 Organizing Committee.

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

Marcel Bollmann. 2018. Normalization of historical texts with neural network models. *Bochumer Linguistische Arbeitsberichte*, 22. Revised and updated version of PhD thesis.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*, December.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Bryan Jurish. 2010. More than words: Using token context to improve canonicalization of historical German. *JLCL*, 25:23–39, 01.

Courtney Mansfield, Ming Sun, Yuzong Liu, Ankur Gandhe, and Björn Hoffmeister. 2019. Neural text normalization with subword units. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 190–196, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Petar Mitankin, Stefan Gerdjikov, and Stoyan Mihov. 2014. An approach to unsupervised historical text normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 29–34. ACM.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 301–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Rule-based normalisation of historical text – a diachronic study. In *LThist 2012–First International Workshop on Language Technology for Historical Text (s), 11th Conference on Natural Language Processing (KONVENS 2012), September 19-21, 2012, Vienna, Austria*, pages 333–341. Österreichische Gesellschaft für Artificial Intelligence (ÖGAI).

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, Gothenburg, Sweden, April. Association for Computational Linguistics.

Paul Rayson, Dawn Archer, and Nicholas Smith. 2005. VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. *Corpus Linguistics 2005*.

Martin Reynaert, Iris Hendrickx, and Rita Marquilhas. 2012. Historical spelling normalization. a comparison of two statistical methods: TICCL and VARD2. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pages 87–98.

Gerold Schneider, Eva Pettersson, and Michael Percillier. 2017. Comparing rule-based and smt-based spelling normalisation for english historical texts. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 40–46.

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, August. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October-November. Association for Computational Linguistics.

A Appendix

```
--num-embed 256
--max-num-epochs 50
--checkpoint-interval 500
--encoder transformer
--decoder transformer
--num-layers 6
--layer-normalization
--transformer-model-size 256
--transformer-feed-forward-num-hidden 512
--transformer-dropout-prepost 0.1
```

Figure 4: Sockeye configurations.

LM (d)	DE_r	DE_a	SL_b	SL_g	ES	PT
0	89.00	89.29	93.29	95.81	94.54	95.02
0.05	89.14	89.63	93.83	95.82	94.70	95.07
0.1	89.45	90.00	93.98	95.81	94.72	95.16
0.15	89.62	90.69	94.03	95.87	94.76	95.26
0.2	89.84	90.95	94.11	95.85	94.81	95.25
0.25	89.64	91.82	94.01	95.76	94.88	95.23
0.3	89.40	91.79	93.98	95.61	94.88	95.22
0.35	89.07	91.76	93.95	95.35	94.65	95.23
0.4	88.85	91.54	93.83	95.02	94.43	95.16
0.45	88.29	91.53	93.63	94.60	94.00	95.04
0.5	87.81	91.22	93.34	93.84	93.47	94.81
0.55	87.28	91.22	92.91	92.99	92.99	94.53

Table 6: The full threshold window evaluation.

LM (d)	DE_r	DE_a	SL_b	SL_g	ES	PT
0	89.01	89.29	93.31	95.81	94.55	95.02
0.05	89.88	90.14	94.39	96.02	94.91	95.15
0.1	90.85	90.83	94.69	96.20	95.08	95.36
0.15	91.60	92.21	94.87	96.41	95.42	95.59
0.2	92.61	93.03	95.10	96.54	95.74	95.71
0.25	93.13	94.39	95.20	96.70	96.05	95.83
0.3	93.38	94.76	95.38	96.86	96.31	96.00
0.35	93.71	95.02	95.61	96.95	96.43	96.23
0.4	94.12	95.41	95.87	97.11	96.53	96.37
0.5	94.61	96.06	96.38	97.37	96.78	96.60
0.6	95.13	96.36	96.79	97.61	97.06	96.83
0.7	95.41	96.62	97.17	97.78	97.17	97.08
0.8	95.60	96.92	97.42	97.93	97.63	97.18
0.9	95.87	97.21	97.62	98.02	97.87	97.31
1	96.10	97.35	97.68	98.10	97.97	97.43
1.25	96.22	97.46	97.85	98.19	98.31	97.72
1.5	96.22	97.50	97.93	98.24	98.65	97.93
1.75	96.22	97.53	97.94	98.28	98.71	98.05
2	96.22	97.54	97.96	98.32	98.73	98.13
2.5	96.22	97.54	97.96	98.32	98.81	98.31
3	96.22	97.54	97.96	98.34	98.81	98.42
3.5	96.22	97.54	97.96	98.34	98.81	98.44
4	96.22	97.54	97.96	98.34	98.83	98.44
all	96.22	97.54	97.96	98.34	98.83	98.47

Table 7: The number of correct normalisations found within the corresponding threshold window.

normaliser	Reranked	Gold	Translation	count
dass	das	das	the	576
waı̈e	wer	wer	who	97
sandte	sankt	sankt	saint	116
herr	er	er	lord	25
htte	hat	hat	had	19
hat	htte	htte	would have	19
sag	sage	sage	(I) say	17
lieb	lief	lief	ran	13
tot	tod	death	death	12
herr	her	her	from	12

Table 8: Top ten most frequent corrections by the reranker in German Anslem.