Normalizing Medieval German Texts: from rules to deep learning

Natalia Korchagina
Institute of Computational Linguistics
University of Zurich
korchagina@cl.uzh.ch

Abstract

The application of NLP tools to historical texts is complicated by a high level of spelling variation. Different methods of historical text normalization have been proposed. In this comparative evaluation I test the following three approaches to text canonicalization on historical German texts from 15th–16th centuries: rule-based, statistical machine translation, and neural machine translation. Character based neural machine translation, not being previously tested for the task of normalization, showed the best results.

1 Introduction

Due to an increased interest in Digital Humanities, more and more heritage texts are becoming available in digital format. The ever growing amount of these text collections motivates researchers to use automatic methods for its processing. In many cases, automatic processing of historical corpora is complicated by a high level of spelling variation. Non-standardized orthography, resulting in inconsistent data, is a substantial obstacle to the application of the existing NLP tools. Normalization of historical texts, i.e., the mapping of historical word forms to their modern equivalents (see Figure 1), has proven to be an effective method of improving the quality of the automatic processing of historical corpora.

SOURCE: *Witter sy im nitt zu wissen .*
NORM.: Weiter sei ihm nicht zu wissen .

Figure 1: Sentence in historical German (SOURCE) and its modernised spelling (NORM.).

Various approaches to text normalization have been proposed. For instance, methods based on the Levenshtein edit distance algorithm and its variations are widely used for text canonicalization. Bollmann et al. (2011) described a technique performing automatic Levenshtein-based rule derivation from a word-aligned parallel corpus. Pettersson et al. (2013a) presented a different string similarity approach, using context-sensitive, weighted edit distance calculations combined with compound splitting. Another approach, applying character-based statistical machine translation (SMT) is documented in (Pettersson et al., 2013b; Scherrer and Erjavec, 2013; Sánchez-Martínez et al., 2013). Pettersson et al. (2014) conducted a comparative evaluation of the following three normalization approaches: filtering, Levenshtein-based and SMT-based, to show that the latter generally outperformed the former two methods. Bollmann and Søgaard (2016) reported that a deep neural network architecture improves the normalization of historical texts, compared to both baseline using conditional random fields and Norma tool (Bollmann, 2012). Deep learning methods are known to work best with large amounts of data, and yet the authors witnessed an improvement with only a few thousand tokens of training material.

Considering the above mentioned successful applications of both character-based SMT and neural networks for normalization of historical texts, I explore the suitability of character-based neural machine translation for this task. Costa-Jussà and Fonollosa (2016), and Lee et al. (2016) presented character-based neural MT systems improving machine translation. Moreover, compared to the deep learning architecture described in (Bollmann and Søgaard, 2016), a neural MT system does not require an explicit character alignment, which makes the normalization setup easier.

This paper reports the results of a comparative evaluation of normalization methods applied to Early New High German texts (1450–1550).

For this assessment I tested the following normalization methods: edit-based, statistical machine translation, and neural machine translation. The first two approaches were previously tested on German texts from the same period, but the application of neural MT to text normalization has not yet been documented. Section 2 introduces the data used for the experiments. In Section 3, I will describe the normalization methods. Section 4 will present evaluation results. Finally, in Section 5 I will summarize the outcome of the comparative evaluation and give some possible direction for future work.

2 Historical Text Corpora

This study is part of a larger project funded by the Swiss Law Sources Foundation, where I use historical legal texts[1] (i.e., decrees, regulations, court transcripts) kindly provided by the Foundation as material for my research. Therefore, I am particularly interested in finding the best performing method for normalizing these historical texts. The Collection of the Swiss Law Sources is multilingual and contains texts issued on Swiss territory from the early Middle Ages up to 1798. In my research project I work with texts written between 1450 and 1550, which corresponds to the Early New High German period. Available in digital format as critical editions of the primary sources (i.e., manuscripts), they do not contain any linguistic annotation or normalized forms. For this case study, we manually normalized a subset of the corpus, 2500 historical-modern word pairs. This dataset will be referred to as baseline in this paper.

The baseline dataset being considerably small, I also augmented it with other historical German data, to observe, if the amount of training data influences normalization results.

First, I added the data from the database of historical terms of the Swiss Law Sources Foundation. The German part of this database covers the period from 1220 to 1798. The database contains historical terms situated at the end of each printed volume of the Foundation, as well as modern keywords, corresponding to the source terms. I extracted 16,857 historical-modern pairs for normalization experiments. This corpus, due to its provenance, i.e., dictionary of terms, mostly contains nouns. In the next sections, I will refer to this dataset as LemmData.

Another corpus to augment the training set, is a manually annotated subset of the GerManC corpus (Scheible et al., 2011), containing 50,310 historical-modern word pairs belonging to the time period 1650–1800 (Early Modern German), and to the following eight genres: drama, newspapers, sermons, personal letters, narrative prose, scholarly, scientific and legal text.

The additional datasets, LemmData and GerManC, are quite different from the baseline. The LemmData corpus is closer to the baseline geographically, being produced on the Swiss territory, but it covers a much larger temporal span. GerManC is the largest corpus of the three, but it belongs to a much later period and was produced mainly on the German territory. Given the areal diversity of historical German, the regional provenance of GerManC contributes to its difference from the baseline. Nevertheless, by now, it is the only publicly available corpus of historical German containing manually produced normalizations. To measure the spelling variance present in the three datasets, I calculated the average string distance. For the baseline corpus, LemmData, and GerManC it corresponds to 0.91, 2.36, and 0.32, respectively. The biggest amount of spelling variation is thus present in the LemmData corpus. This can be explained by the following two facts. First, some of its lexicon belongs to the earliest period of the three texts (13th century). Furthermore, in contrast to the other two datasets consisting of regular texts, the LemmData corpus is based on a dictionary of terms. It mostly contains nouns, and does not include any punctuation marks.

The datasets' details are summarized in Table 1.

3 Normalization Methods

3.1 The Norma tool

The Norma tool[2] was developed for (semi-)automatic normalization of historical corpora. It was originally created for canonicalization of Early New High German texts, but can be trained on any data. The tool comes with three external modules, "normalizers", each implementing a normalization method. These modules can be used either separately or combined. The normalizers provide normalization candidates. Depending on how the candidate's confidence score compares to a pre-defined

[1] https://www.ssrq-sds-fds.ch/online/

[2] https://github.com/comphist/norma

Corpus	Period	Pairs	Region	Genres	Content	Av. LD
baseline	1463-1538	2500	CH: Bern	legal texts	text	0.91
LemmData	1220-1798	16,857	CH: all German speaking Swiss cantons	legal texts	dictionary	2.36
GerManC	1650-1800	50,310	DE: North, West Central East Central West Upper East Upper	drama newspapers sermons personal letters narratives scientific texts legal texts	text	0.32

Table 1: Corpora used in this case study.

threshold, Norma decides, whether this candidate is acceptable.

The three normalizers are: *Mapper*, *RuleBased*, and *Weighted Levenshtein Distance*. *Mapper* uses a simple wordlist mapping method. The *RuleBased* normalizer uses context-aware rules automatically derived from aligned training data, to rewrite sequences of the input characters. More details on this approach can be found in (Bollmann et al., 2011). The *Weighted Levenshtein Distance* normalizer finds a candidate with the lowest weighted Levenshtein distance score.

Since the mapping method is conceptually simple, I will not be using it in this case study. For the evaluation, I tested the remaining two normalizers separately and combined, to find out the combination where the *RuleBased* normalizer followed by *Weighted Levenshtein Distance* works best. This setup will be referred to as *Norma* in further sections.

3.2 Statistical Machine Translation

As a second method for this case study, I used character-level statistical machine translation. It differs from word-level machine translation in that it aligns characters occurring in token pairs, instead of aligning words. As a result, translation models contain phrases consisting of character sequences instead of word sentences. Language models, in their turn, are trained on character n-grams instead of word n-grams.

For the SMT experiments, I used the Moses toolkit[3] with settings as described in (Pettersson et al., 2013b).

3.3 Neural Machine Translation

The recently proposed approach to machine translation, neural MT (Bahdanau et al., 2014; Sutskever et al., 2014; Luong et al., 2015; Cho et al., 2014) obtained state-of-the-art results for various language pairs. Neural MT systems are generally implemented as an encoder-decoder architecture. The encoder reads the source sentence and encodes it into a sequence of hidden states, whereas the decoder generates a corresponding translation based on the encoded sequence of hidden states.

I did not find any reports on the application of neural MT to the task of historical text normalization, but the comparative study by Sennrich (2016) proved that a fully character-level neural MT model outperformed a fully subword model at transliterating unknown names. This task is similar to normalization. The fully character-level neural MT approach in these experiments which I followed in mine, is described in (Lee et al., 2016).

This method maps a source character sequence to a target character sequence without explicit segmentation. Due to the fact that this model has no explicitly hard-coded knowledge of word boundaries, it is possible to use sentence-aligned data for training and testing. Nevertheless, since part of my data, i.e., LemmData is not a set of sentences, but a set of historical-modern pairs, I use tokenized, word-aligned datasets for neural MT experiments as well.

The source code implementing the models described by Lee et al. (2016) is publicly available[4].

[3] `http://www.statmt.org/moses/`

[4] `https://github.com/nyu-dl/dl4mt-c2c`

4 Evaluation

Given the small size of the manually normalized baseline (2500 historical-modern word pairs), I applied 10-fold cross-validation to evaluate the performance of the three normalization methods. First, the experiments were conducted on the baseline, with 2000 pairs (2250 for Norma) of training data, 250 pairs in development set (for SMT and neural MT), and 250 pairs in the test set. Then, the training set was augmented with LemmData and GerManC data, while using both development and test sets in their initial size. Table 2 shows the evaluation results.

The neural MT system trained on the baseline combined with LemmData and GerManC (69,167 tokens) showed the best accuracy score, 0.81. It is followed by SMT results, 0.79, trained on 18,857 tokens of the baseline augmented with LemmData.

To estimate the average variability in the output between the folds of test data, I calculated the standard deviation of the accuracy for each system (SD_{acc} in Table 2). This measure demonstrates how close or far away the data is from the mean (average accuracy, ACC in Table 2). It approximates the mean distance between each fold and the arithmetic mean. The majority of the data (68.2% assuming that the distribution is normal) would be located between one standard deviation above and below the mean. For instance, given the average accuracy 0.75 of the Norma baseline system, the standard deviation 0.03 means that the accuracy scores for the majority of the folds vary from 0.72 to 0.78. The standard deviation between different systems changes slightly, from 0.02 to 0.04.

It is interesting to observe, how the systems respond to the augmentation of the training set (see Figure 2). While the performance of the rule-based system, Norma, remains rather stable, it changes by the other two systems. The SMT system first reacts positively to the increase of the training data with LemmData. This data is similar to the baseline in its regional provenance, though is very varied with respect to the covered time periods (see Table 1). When the training set was further augmented with GerManC, belonging to a later period of time, it resulted in a performance decrease. On the other hand, the performance of the neural MT system steadily increased with each addition of data. This observation corresponds to the one made in (Bollmann and Søgaard, 2016) where the normalization accuracy increased with a deep learning normalization method and remained stable or decreased with other methods, including Norma.

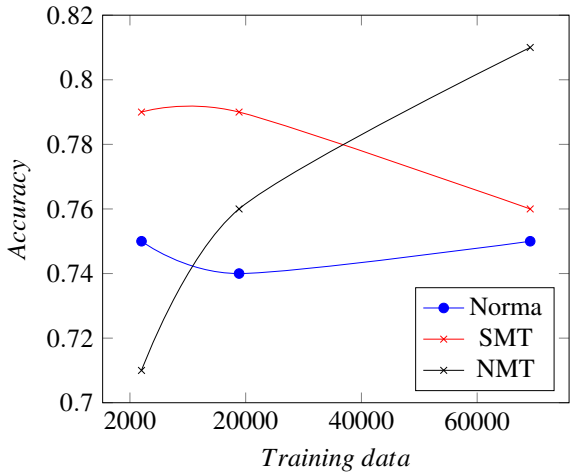

Figure 2: Word accuracy averaged over 10 folds for different sizes of the training set.

The accuracy and character error rate scores of the three normalization systems compared in the best performing configurations does not differ much: from Norma's 0.75/0.14 to neural MT's 0.81/0.08. To estimate how different the output of the systems actually is, I conducted a quantitative analysis of the output (see Table 3). First, I compared how similar is the output of the systems, i.e., how often the systems agree on a certain normalization. The lowest, 70%, is the agreement between the three systems, and the highest, 80%, between the SMT and the neural MT systems. In addition, based on the amount of the commonly incorrect cases, I calculated the percentage of the "error agreement", i.e., how often the systems produced the same erroneous normalization. The pair SMT/neural MT leads with 51% of error similarity. Thus, the output produced by SMT and neural MT systems is the most similar. It can be explained by the statistical nature of both systems, in contrast to the rule-based Norma.

Table 4 presents contrastive examples of the output, where one system produced the correct normalization, and the other two failed.

5 Conclusion

I presented a comparative evaluation of the approaches to spelling normalization in historical texts, tested on Early New High German data (1450-1550). I tested the following three meth-

Training data	Pairs	Norma			SMT			NMT		
		ACC	CER	SD_{acc}	ACC	CER	SD_{acc}	ACC	CER	SD_{acc}
baseline	2000	0.75	0.14	0.03	0.79	0.08	0.03	0.71	0.17	0.04
baseline+LemmData	18,857	0.74	0.14	0.03	0.79	0.08	0.03	0.76	0.11	0.04
baseline+LemmData+GerManC	69,167	0.75	0.13	0.02	0.76	0.10	0.04	0.81	0.08	0.03

Table 2: Averaged evaluation results, i.e., accuracy (ACC) and character error rate (CER) over 10 folds.

Systems	Agreement	Common incorrect normalizations
Norma & SMT & NMT	70%	46%
Norma & SMT	76%	44%
Norma & NMT	75%	35%
SMT & NMT	80%	51%

Table 3: Analysis of the output: total amount of cases the systems agreed upon (Agreement) and amount of cases where the systems produced the same incorrect normalization, calculated based on the number of common incorrect cases.

SOURCE	**Norma**	**SMT**	**NMT**	REF
meyen	maien	mein	mai	mai
ander	ander	andere	ander	andere
sturen	steuern	sturen	steueren	steuern

Table 4: Normalization examples. Correct normalizations are highlighted.

ods: rule-based, character-level statistical machine translation, and character-level neural machine translation. In this case study, neural MT outperformed the other two methods. In contrast to the rule-based method and SMT, it also benefited most from the augmentation of the training set.

Considering the success of the applied neural method, future work may consist in testing other deep learning methods. For instance, I used only one of the systems presented in (Lee et al., 2016), the fully character-based one. The other described a system performing neural machine translation with subword units.

Another direction for future work could consist in adding more training data to observe, if the performance of the neural MT system would continue to improve.

More effort could also be invested into the SMT method. The SMT system did not profit from the augmentation of the training set, due to its period and domain differences from the baseline. This is similar to the problem of the out-of-domain data in phrase-based machine translation. Out-of-domain data introduces ambiguity to the translation model, resulting in the translation choices irrelevant for the test set. Translation model domain adaptation approach was proposed by Sennrich (2012) to deal with the out-of-domain data. This method can potentially improve the results of the SMT experiments with additional training sets.

Acknowledgments

The author would like to thank Dr. Pascale Sutter and Rebekka Plüss for normalizing the source data used in the experiments as baseline. This work is supported by the Swiss Law Sources Foundation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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 Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pages 34–42.

Marcel Bollmann. 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Marta R. Costa-Jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. *CoRR*, abs/1603.00810.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013a. Normalization of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics*.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013b. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA*.

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.

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. *CoRR*, abs/1306.3692.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 124–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. In *BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, August.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 539–549.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. *CoRR*, abs/1612.04629.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.