# Revisiting NMT for Normalization of Early English Letters

**Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann and Eetu Mäkelä**
Department of Digital Humanities
University of Helsinki
firstname.lastname@helsinki.fi

## Abstract

This paper studies the use of NMT (neural machine translation) as a normalization method for an early English letter corpus. The corpus has previously been normalized so that only less frequent deviant forms are left out without normalization. This paper discusses different methods for improving the normalization of these deviant forms by using different approaches. Adding features to the training data is found to be unhelpful, but using a lexicographical resource to filter the top candidates produced by the NMT model together with lemmatization improves results.

## 1 Introduction

Natural language processing of historical data is not a trivial task. A great deal of NLP tools and resources work out of the box with modern data, whereas they can be of little use with historical data. Lack of a written standard in the early days, and the fact that the language has changed over the centuries require addressing in order to achieve higher-level NLP tasks.

The end goal of our project is to identify neologisms and study their spread in the CEEC (Corpora of Early English Correspondence) (Nevalainen et al., 1998–2006), a letter corpus consisting of texts starting from the 15th century ranging all the way to the 19th century. In order to achieve a higher recall in neologisms, the corpus needs to be normalized to present-day spelling.

A regular-expression based study of neologisms (Säily et al., In press) in the same corpus suggested the use of the Oxford English Dictionary (OED, n.d.) as a viable way of detecting neologism candidates. Words occurring in the corpus before the earliest attestation in the OED would thus be considered potential neologism candidates. However, in order to achieve this, the words in the corpus need to be mappable to the OED, in other words, normalized to their modern spelling. As we are dealing with historical data, the fact that a neologism exists in the OED is a way of ensuring that the new word has become established in the language.

A previous study in automatic normalization of the CEEC comparing different methods (Hämäläinen et al., 2018) suggested NMT (neural machine translation) as the single most effective method. This discovery is the motivation for us to continue this work and focus only on the NMT approach, expanding on what was proposed in the earlier work by using different training and post-processing methods.

In this paper, we will present different NMT models and evaluate their effectiveness in normalizing the CEEC. As a result of the previous study, all the easily normalizable historical forms have been filtered out and we will focus solely on the historical spellings that are difficult to normalize with existing methods.

## 2 Related Work

Using character level machine translation for normalization of historical text is not a new idea. Research in this vein has existed already before the dawn of neural machine translation (NMT), during the era of statistical machine translation (SMT).

Pettersson et al. (2013) present an SMT approach for normalizing historical text as part of a pipeline where NLP tools for the modern variant of the language are then used to do tagging and parsing. The normalization is conducted on a character level. They do alignment of the parallel data on both word and character level.

SMT has also been used in normalization of contemporary dialectal language to the standardized normative form (Samardzic et al., 2015). They test normalization with word-by-word trans-



*Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 71–75
Minneapolis, MN, USA, June 7, 2019. 

lation and character level SMT. The character level SMT improves the normalization of unseen and ambiguous words.

Korchagina (2017) proposes an NMT based normalization for medieval German. It is supposedly one of the first attempts to use NMT for historical normalization. The study reports NMT outperforming the existing rule-based and SMT methods.

A recent study by Tang et al. (2018) compared different NMT models for historical text normalization in five different languages. They report that NMT outperforms SMT in four of the five languages. In terms of performance, vanilla RNNs are comparable to LSTMs and GRUs, and also the difference between attention and no attention is small.

## 3 The Corpus

We use the CEEC as our corpus. It consists of written letters from the 15th all the way to the 19th century. The letters have been digitized by hand by editors who have wanted to maintain the linguistic form as close to the original as possible. This means that while our data is free of OCR errors, words are spelled in their historical forms.

The corpus has been annotated with social metadata. This means that for each author in the corpus we can get various kinds of social information such as the rank and gender of the author, time of birth and death and so on. The corpus also records additional information on a per letter basis, such as the year the letter was written, the relationship between the sender and the recipient, and so on.

## 4 The NMT Approach

We use OpenNMT[1] (Klein et al., 2017) to train the NMT models discussed in this paper. The models are trained on a character level. This means that the model is supplied with parallel lists of historical spellings and their modern counterparts, where the words have been split into individual characters separated by white spaces.

The training is done for pairs of words, i.e. the normalization is to be conducted without a context. The NMT model would then treat individual characters as though they were words in a sentence and "translate" them into the corresponding modernized spelling.

[1] Version 0.2.1 of opennmt-py

### 4.1 The Parallel Data

We use different sources of historical-modern English parallel data. These include the normalized words from the CEEC, the historical forms provided in the OED and the historical lemmas in the Middle English Dictionary (MED, n.d.) that have been linked to the OED lemmas with modern spelling. This parallel data of 183505 words is the same as compiled and used in Hämäläinen et al. (2018).

For testing the accuracy of the models we prepare by hand gold standards by taking sets of 100 words of the previously non-normalized words in the CEEC. The accuracy is tested as an exact match to the gold standard. We prepare one generic test set and four century specific test sets of the 15th, 16th, 17th and 18th century words. Each of these five gold-annotated test sets consists of 100 words normalized by a linguist knowledgeable in historical English. The reason why we choose to prepare our own gold standard is that we are interested in the applicability of our approach in the study of the CEEC corpus as a step in our neologism identification pipeline.

### 4.2 Different NMT models

The previous work (Hämäläinen et al., 2018) on the normalization of the CEEC corpus used the default settings of OpenNMT. This means that the encoder is a simple recurrent neural network (RNN), there are two layers both in the encoder and the decoder and the attention model is the general global attention presented by Luong et al. (2015).

In this section we train the model with different parameters to see their effect on the accuracy of the model. The accuracy is evaluated and reported over a concatenated test set of all the five different gold standards.

At first, we change one parameter at a time and compare the results to the default settings. We try two different encoder types, bi-directional recurrent neural networks (BRNNs) and mean, which is an encoder applying mean pooling. BRNN uses two independent encoders to encode the sequence reversed and without reversal. The default RNN, in contrast, only encodes the sequence normally without reversing it.

In addition to the default attention model, we also try out the MLP (multi-layer perceptron) model proposed by Bahdanau et al. (2014). We

change the number of layers used by the encoder and decoder and run the training with four and six layers for both encoding and decoding.

| | default | mlp | mean | brnn | 4 layers | 6 layers |
|---|---|---|---|---|---|---|
| acc. | *35.6%* | *36.6%* | *13%* | *39.8%* | *37.2%* | *36.6%* |

Table 1: Accuracy of each method

Table 1 shows the accuracy of the model trained with the different parameters. BRNNs seem to produce the best results, while the MLP attention model and additional layers can be beneficial over the default attention and number of layers. Next, we will try out different combinations with the BRNN encoder to see whether we can increase the overall accuracy.

| | brnn | brnn +mlp | brnn +4 layers | brnn+mlp +4 layers |
|---|---|---|---|---|
| acc. | *39.8%* | *36%* | *35.8%* | *38.2%* |

Table 2: Accuracy of BRNN models

We can see in Table 2 that the BRNN with the default attention and the default number of layers works better than the other combinations. This means that for our future models, we will pick the BRNN encoder with default settings.

### 4.3 Additional Information

The previous study (Hämäläinen et al., 2018) showed that using information about the centuries of the historical forms in training the NMT and SMT models was not beneficial. However, there might still be other additional information that could potentially boost the performance of the NMT model. In this part, we show the results of models trained with different additional data.

In addition to the century, the CEEC comes with social metadata on both the letters and the authors. We use the sender ID, sender rank, relationship code and recipient rank as additional information for the model. The sender ID is used to uniquely identify different senders in the CEEC, the ranks indicate the person's social status at the time of the letter (such as nobility or upper gentry) and the relationship code indicates whether the sender and recipient were friends, had a formal relationship and so on.

The social information is included in the parallel data in such a way that for each historical form, the social metadata is added if the form has appeared in the CEEC. If the form has not appeared in the CEEC, generic placeholders are added instead of real values. The metadata is appended as a list separated by white spaces to the beginning of each historical form.

| | 15th | 16th | 17th | 18th | generic |
|---|---|---|---|---|---|
| *eSpeak IPA with graphemes* | 22% | 25% | 31% | 14% | 20% |
| *Only eSpeak IPA* | 43% | 35% | 52% | 20% | 36% |
| *Metaphone* | 22% | 23% | 25% | 12% | 23% |
| *Bigram* | 16% | 9% | 11% | 3% | 9% |
| *No feature* | 45% | 35% | 48% | 25% | 42% |

Table 3: Results with additional information

When reading the historical letters, what is helpful for a human reader in understanding the historical forms is reading them out loud. Because of this discovery, we add pronunciation information to the parallel data. We add an estimation of pronunciation to the beginning of each historical form as an individual token. This estimation is done by the Metaphone algorithm (Philips, 1990). Metaphone produces an approximation of the pronunciation of a word, not an exact phonetic representation, which could be useful for the NMT model.

In addition to the Metaphone approximation, we use eSpeak NG[2] to produce an IPA transcription of the historical forms. For the transcription, we use British English as the language variant, as the letters in our corpus are mainly from different parts of England. We use the transcription to train two different models, one where the transcription is appended character by character to the beginning of the historical form, and another where we substitute the transcription for the historical form.

The final alteration in the training data we try in this section is that instead of providing more information, we try to train the model with character bigrams rather than the unigrams used in all the other models.

The results for the different approaches discussed in this section are shown in Table 3. As we can see, only the eSpeak produced IPA, when it no longer includes the original written form, comes close to using the character unigrams from the parallel data. Training with just the IPA transcription outperforms the character approach only in the 17th century.

[2]https://github.com/espeak-ng/espeak-ng/

### 4.4 Picking Normalization Candidate

Looking at the results of the NMT model, we can see that more often than not, when the normalization is not correct, the resulting word form is not a word of the English language. Therefore, it makes sense to explore whether the model can reach a correct normalization if instead of considering the best normalization candidate produced by the NMT model, we look at multiple top candidates.

During the translation step, we make the NMT model output 10 best candidates. We go through these candidates starting from the best one and compare them against the OED. If the produced modern form exists in the OED or exists in the OED after lemmatization with Spacy (Honnibal and Montani, 2017)[3], we pick the form as the final normalization. In other words, we use a dictionary to pick the best normalization candidate that exists in the English language.

| | 15th | 16th | 17th | 18th | generic |
|---|---|---|---|---|---|
| OED +Lemma | *49%* | *42%* | *51%* | *19%* | *43%* |
| Lemma | *45%* | *35%* | *48%* | *25%* | *42%* |

Table 4: Results with picking the best candidate with OED

Table 4 shows the results when we pick the first candidate that is found in the OED and when we only use the top candidate for the BRNN model. We can see improvement on all the test sets except for the 18th century.

| | 15th | 16th | 17th | 18th | generic |
|---|---|---|---|---|---|
| OED +Lemma | *69%* | *78%* | *71%* | *50%* | *61%* |
| Lemma | *61%* | *67%* | *63%* | *45%* | *53%* |

Table 5: Results with OED and lemmatization

If we lemmatize both the input of the NMT model and the correct modernized form in the gold standard with Spacy before the evaluation, we can assess the overall accuracy of OED mapping with the normalization strategies. The results shown in Table 5 indicate a performance boost in the mapping task, however this type of normalization does not match the actual inflectional forms. Nevertheless, in our case, lemmatization is possible as we are ultimately interested in mapping words to the OED rather than their exact form in a sentence.

[3]With model en_core_web_md

## 5 Conclusions

Improving the NMT model for normalization is a difficult task. A different sequence-to-sequence model can improve the results to a degree, but the gains are not big. Adding more features, no matter how useful they might sound intuitively, does not add any performance boost. At least that is the case for the corpus used in this study, as the great deal of social variety and the time-span of multiple centuries represented in the CEEC are reflected in the non-standard spelling.

Using a lexicographical resource and a good lemmatizer, as simplistic as they are, are a good way to improve the normalization results. However, as getting even more performance gains for the NMT model seems tricky, probably the best direction for the future is to improve on the method for picking the contextually most suitable normalization out of the results of multiple different normalization methods as originally explored in Hämäläinen et al. (2018). Thus, the small improvement of this paper can be brought back to the original setting as one of the normalization methods.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*.

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann, and Eetu Mäkelä. 2018. Normalizing early English letters to Present-day English spelling. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 87–96.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*.

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.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

MED. n.d. Middle English Dictionary. University of Michigan. Https://quod.lib.umich.edu/m/med/.

Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Samuli Kaislaniemi, Mikko Laitinen, Tanja Säily, and Anni Sairio. 1998–2006. CEEC, Corpora of Early English Correspondence. Department of Modern Languages, University of Helsinki. Http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/.

OED. n.d. OED Online. Oxford University Press. Http://www.oed.com/.

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; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, 087, pages 54–69. Linköping University Electronic Press.

Lawrence Philips. 1990. Hanging on the Metaphone. *Computer Language*, 7(12).

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of swiss german. In *Proceedings of the 7th Language and Technology Conference*.

Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. In press. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition*.

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.