An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre
Uppsala University
Uppsala, Sweden
`firstname.lastname@lingfil.uu.se`

Abstract

In this paper, we apply different NMT models to the problem of historical spelling normalization for five languages: English, German, Hungarian, Icelandic, and Swedish. The NMT models are at different levels, have different attention mechanisms, and different neural network architectures. Our results show that NMT models are much better than SMT models in terms of character error rate. The vanilla RNNs are competitive to GRUs/LSTMs in historical spelling normalization. Transformer models perform better only when provided with more training data. We also find that subword-level models with a small subword vocabulary are better than character-level models. In addition, we propose a hybrid method which further improves the performance of historical spelling normalization.

1 Introduction

With increasing access to digital historical text, the processing of these historical texts is attracting more and more interest. However, in contrast to modern text, historical text processing faces more challenges. First, for historical text, there is little annotated data for training a model, which leads to data sparsity issues when using statistical methods, similar to the situation for low-resource languages. Second, there are a lot of variations in historical texts from different time periods, not only in spelling but also in lexical semantics and syntax. Therefore, the NLP tools developed for modern text cannot be used for these historical texts directly. Spelling normalization is the task of mapping a historical spelling to its modern spelling. It is usually used as a preprocessing step before feeding the historical text into modern NLP tools (Pettersson et al., 2013b; Bollmann, 2013; Sánchez-Martínez et al., 2013), which leads to much better results compared to analyzing unnormalized historical texts.

There are some papers in which neural machine translation (NMT) models are employed for the spelling normalization task. Korchagina (2017) utilizes a character-level NMT model for medieval German texts. Bollmann et al. (2017) apply an attention-based NMT model to historical German texts. The evidence so far is too incomplete to draw any general conclusions about the utility of different NMT models for historical spelling normalization. We are interested in exploring how different properties of NMT models interact with different aspects of the spelling normalization problem and find some generalizations about the use of NMT models for this task.

In this paper, we apply different NMT models to the spelling normalization task for historical stages of five languages, English, German, Hungarian, Icelandic, and Swedish. We compare our result to those of Pettersson et al. (2014), which are obtained with statistical machine translation (SMT) models. We investigate whether NMT models outperform SMT models in general, and explore which properties of NMT models are suitable for spelling normalization. Compared to the conventional machine translation (MT) tasks, we train models on token pairs instead of sentence pairs. Token length is usually shorter than sentence length. After reviewing related work in Section 2, we give our hypotheses about utilizing NMT models for the spelling normalization task and select different NMT models based on our hypotheses in Section 3. The selected NMT models are at different levels (character-level, subword-level), have

different attention mechanisms (no attention, soft-attention, multi-head-attention), and different neural network architectures (vanilla recurrent neural networks (RNNs), gated recurrent units (GRUs), long short-term memory units (LSTMs), and self-attention). In Section 4, we describe the datasets and our detailed experimental settings. In Section 5, we give our results and analyze the performance of different NMT models. Our conclusions and future work are in Section 6.

To conclude, our main contributions can be summarized as follows:

- We evaluate different NMT models on historical spelling normalization in a multilingual setting.
- We find that NMT models are better than SMT models considering character error rate (CER).
- We show that vanilla RNNs are competitive to GRUs/LSTMs.
- We demonstrate that transformer models perform better when provided with more training data.
- We reveal that models with a small subword vocabulary are better than character-level models.

2 Related Work

2.1 Historical Spelling Normalization

Various methods have been employed for historical spelling normalization. Rayson et al. (2005) use a dictionary to map tokens to their modernized spellings, and many different edit-distance-based methods have been proposed to deal with spelling normalization (Bollmann et al., 2011; Pettersson et al., 2013a). In addition, character-level SMT models have been applied to spelling normalization, where models are trained on token pairs instead of sentence pairs (Pettersson et al., 2013b; Scherrer and Erjavec, 2013; Sánchez-Martínez et al., 2013). Each character of a token is viewed as a word of a sentence. The language models are trained on character N-grams instead of word N-grams. Pettersson et al. (2014) evaluate dictionary-based methods, edit distance-based methods, and SMT methods on five different historical languages. The results show that the character-level SMT model performs best on four out of five historical languages.

With the development of deep learning, various neural networks have been applied to many tasks. In recent years, NMT models (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015) have outperformed SMT models (Koehn et al., 2003) distinctly in various translation tasks. We hypothesize that NMT models also perform better than SMT models for the historical spelling normalization task. Bollmann and Søgaard (2016) view the spelling normalization as a character-level sequence labeling task, and utilize a bi-directional LSTM for this task, which is better than a conditional random field (CRF) model. They also use additional data with similar but not the same historical spelling for a multi-task learning model, and gain further improvement. Korchagina (2017) applies a character-level NMT model to medieval German text, and finds that the NMT models can only outperform the SMT models with a larger training set. Bollmann et al. (2017) test attention-based NMT models, and multi-task learning models which learn to normalize and pronounce with a grapheme-to-phoneme dictionary, on spelling normalization. Both of them achieve good performance. They hypothesize that the reason why the combination of these two models does not gain more improvement is that the multi-task learning has already learned the attention patterns.

2.2 Neural Machine Translation

In vanilla NMT models, the source sentence is encoded into a fixed-size vector by the encoder. Then, this vector is fed into a decoder. The decoder generates the target sentence word by word conditioned on the fixed-size vector and the generated target words (Kalchbrenner and Blunsom, 2013). Various RNN architectures are usually used as encoders and decoders. Cho et al. (2014) find that the vanilla RNN-based NMT models perform poorly in translating long sentences, which means that vanilla RNNs have problems with long-distance dependencies. To deal with these problems, Cho et al. (2014) propose GRUs, while Sutskever et al. (2014) use LSTMs (Hochreiter and Schmidhuber, 1997) to replace the vanilla RNNs. However, any two tokens in RNNs still have a linear distance. Thus, Vaswani et al. (2017) replace RNNs with self-attention networks which connect any two tokens in a sentence directly.

Due to the expensive computation of NMT models, the vocabulary size is usually very limited, which causes a lot of out-of-vocabulary (OOV) words. Character-level models (Ling et al., 2015; Costa-jussà

and Fonollosa, 2016; Chung et al., 2016) and subword-level models (Sennrich et al., 2016; Wu et al., 2016) are widely used to deal with OOV problems. These two kinds of models need additional segmentation compared to word-level models. For character-level models, we just need to separate each character by space. But we need more complicated segmentation methods for subword-level models. Sennrich et al. (2016) utilize character n-grams and a byte pair encoding (BPE) algorithm (Gage, 1994) for segmentation. Wu et al. (2016) apply the wordpiece model (Schuster and Nakajima, 2012) to segmentation. Based on their experiments, subword-level models outperform word-level and character-level models.

Attention-based NMT models have outperformed all the other architectures in NMT in recent years. Many improved attention-based models have been proposed. Bahdanau et al. (2015) propose an attention-based model which can automatically search for source words that their hidden states are relevant to predicting a target word during decoding. Source words which have a higher correlation with the predicting target word will be assigned a higher weight. Most of the attention-based models use this kind of attention, which is called soft-attention in Xu et al. (2015). Vaswani et al. (2017) propose a model named *Transformer*, with multi-layer and multi-head attention mechanism which is more fine-grained.

3 NMT Models

When we apply NMT models to the historical spelling normalization task, the first research question is which NMT model is most suitable for this task. In this section, we first give four hypotheses about NMT models for spelling normalization, based on the data features of historical spellings and the features of NMT models. Then, we list 8 different NMT models to consider for the spelling normalization task.

3.1 Hypotheses

Hypothesis 1   The performance gap between vanilla RNNs and GRUs/LSTMs is small. In contrast to conventional NMT models, the historical and modern token pairs are our training data instead of parallel sentence pairs. In our experiments, the average token length varies from 4 to 6, which means that we build the model on much shorter sequences. The long-distance problem will be alleviated. It should be noted that we compare the gap to the gap in NMT (Bahdanau et al., 2015)[1].

Hypothesis 2   The gap between NMT models with attention and without attention is also small. Since the average token length is only around five, additionally paying attention to all the tokens in the source sentence may be unnecessary. Thus, we hypothesize that the decoder in the vanilla Encoder-Decoder model can predict most of the targets correctly with only one fixed-size vector from the encoder, even without any attention mechanisms. It should be mentioned that we compare the gap to the gap in NMT (Britz et al., 2017)[2].

Hypothesis 3   Transformer models perform better than soft-attention-based models. Transformer models have more advanced self-attention networks and more fine-grained multi-head attention mechanisms compared to RNN-based models with soft-attention. Thus, transformer models have better performance in conventional translation tasks. We hypothesize that it is the same in the spelling normalization task.

Hypothesis 4   Subword-level NMT models perform better than character-level NMT models. Character-level and subword-level models are proposed to deal with the problem of out-of-vocabulary words mainly, and subword-level NMT models usually outperform character-level models. As we only have small sets of token pairs, it is better to use character-level or subword-level NMT models rather than word-level models.

3.2 Models

To test our hypotheses proposed in the previous section, we will explore 8 different NMT models for the spelling normalization task. The NMT models vary in attention mechanism, neural network architecture, and token granularity. Table 1 gives a more detailed overview.

---

[1]The BLEU (Papineni et al., 2002) scores are 15.73, and 21.83.
[2]The BLEU scores are 17.82, and 26.75.

Name	Level	Attention	Architecture
NoAtt-RNN	character	no	RNN
NoAtt-GRU			GRU
NoAtt-LSTM			LSTM
Att-RNN		soft	RNN
Att-GRU			GRU
Att-LSTM			LSTM
Transformer		multi-head	Self-attention
BPE-Soft	subword	soft	LSTM

Table 1: NMT models for the spelling normalization task. *RNN* means vanilla RNNs.

Since we have hypothesized that different RNN architectures have slight differences, all the subword-level models with soft-attention are trained on LSTMs.

4   Experimental Setup

4.1   Data

All the datasets[3] are exactly the same as the parallel datasets for the SMT models in Pettersson et al. (2014), which are described in Table 2. Data details are shown in Table 3. The datasets consist of a list of token pairs, which have one historical spelling and the corresponding modernized spelling. Note that the same modern spelling may occur with different historical spellings. Moreover, some historical words may be extinct, and people have to use a spelling with a similar meaning but different lexemes as its normalization. Some illustrative English examples are given in Table 4. If a historical spelling is identical to its modern spellings, we call it an *unchanged spelling*. Otherwise, it is called a *changed spelling*. In different languages, the number of unchanged spellings is different.

Language	Time period	Origin
English	1386–1698	*Innsbruck Corpus of English Letters*, a subset of the *Innsbruck Computer Archive if Machine-Readable English Texts* (Markus, 1999)
German	1659–1780	*GerManC* corpus (Scheible et al., 2011)
Hungarian	1440–1541	*Hungarian Generative Diachronic Syntax* project (Simon, 2014)
Icelandic	1150–2008	*Icelandic Parsed Historical Corpus* (Rögnvaldsson et al., 2012)
Swedish	1527–1812	*Gender and Work corpus* (GaW) (Fiebranz et al., 2011)

Table 2: Origin and time periods of the datasets.

Language	Training	Development	Test	Unchanged	Token	Char	Max	Avg
English	148,852	16,461	17,791	75.8	22,302	102	22	4.16
German	39,887	5,418	5,005	84.4	11,521	100	27	4.74
Hungarian	137,669	17,181	17,214	17.1	69,624	128	27	5.91
Icelandic	52,440	6,443	6,384	50.5	14,845	89	16	4.14
Swedish	28,327	2,590	33,544	64.6	11,129	92	36	4.55

Table 3: Statistics of the datasets. The figures in *Training*, *Development*, and *Test* are the numbers of token pairs. The *Unchanged* (%) means the rate of unchanged spellings in the test set. *Token* and *Char* show the token and the character vocabulary sizes in the training set. *Max* and *Avg* show the max length and average length of token in the training set. All counts are based on case-sensitive data.

---

[3] http://stp.lingfil.uu.se/histcorp/tools.html

Historical	citee	gyve	gyf	late
Modern	city	give	give	late

Table 4: Token pair examples in English.

From Table 3, we can see that English and Hungarian have more training data, around 140,000 token pairs, while Swedish only has about 28,000 token pairs. Swedish has the largest test set with more than 33,000 token pairs. The unchanged rate also differs a lot. There are 84.4% and 75.8% historical spellings that are identical to their modern spellings in German and English, respectively. However, the unchanged rate is only 17.1% in Hungarian. In addition, Hungarian has the largest token vocabulary and character vocabulary[4]. The longest token in Icelandic is only 16, but the longest token in Swedish is 36. The average token length of Hungarian is 5.91, which is the longest in all five languages. This is because Hungarian is an agglutinative language.

4.2 Experimental Settings

Different architectures are hard to compare fairly because many factors affect performance. We aim to create a level playing field for the comparison by training with the same toolkit, Marian (Junczys-Dowmunt et al., 2018). Since there is no implementation of models without attention in Marian, we modify the decoder part to enable Marian to train models without attention.[5] We assume that the case of letters is useful for predicting the modern spellings. Thus, the letters in the training set and the tuning set are case-sensitive. The historical spellings in the test set which are the inputs of the model are also case-sensitive. However, to keep consistency with the baseline, we lowercase all the predicted modern spellings during evaluation.

For character-level models, all the characters are added into the vocabulary, even if they only appear once. For subword-level models, we utilize the BPE method in Sennrich et al. (2016) to generate subword units. We try different BPE vocabulary sizes, varying between 100, 200, 300, 500, 1,000 and 5,000.

The vanilla RNN chooses the "tanh" RNN cell. We enable "mini-bach-fit" which automatically choose the mini-batch size for the given "workspace" size, and the "workspace" is set to 7500. We use *Adam* (Kingma and Ba, 2015) as the optimizer. The learning rate is set to 0.0003, but we set the warmup steps to 16,000, which means that the learning rate increases linearly before 16,000 steps. A model checkpoint is saved every 500 updates. The evaluation metrics on the development set are cross-entropy and perplexity. We set the early stopping patience to 8 checkpoints. All the neural networks have 6 layers. The size of embeddings is 512. We tie the target embeddings and the output embeddings in the output layer. We use the checkpoint that achieves the best perplexity to generate the normalizations. We set the beam size to 5 during decoding.

5 Results

The baseline from Pettersson et al. (2014) has very high word accuracy and low CER scores in all five languages. The results in the baseline are obtained using character-level SMT models except for Icelandic, where the combination of a Levenshtein-based method and a dictionary-based method achieved the best results. We use word accuracy and CER to evaluate the predictions. For the historical spelling normalization task, word accuracy is a very important evaluation metric. Moreover, word accuracy is the only evaluation metric in Bollmann and Søgaard (2016) and Bollmann et al. (2017). However, CER is a good supplement to word accuracy. It is more fine-grained and evaluates the character-level normalizations. In our experiments, we use Levenshtein distance to compute CER. Table 5 gives the detailed results of different models in five languages.

---

[4]The character vocabulary includes both alphabetic and non-alphabetic characters.

[5]The modification, the NMT model settings, and the code are available in `https://github.com/tanggongbo/normalization-NMT`

	English		German		Hungarian		Icelandic		Swedish	
	Acc	CER	Acc	CER	Acc	CER	Acc	CER	Acc	CER
Baseline	94.3	0.07	96.6	0.04	80.1	0.21	84.6	0.19	**92.9**	0.07
NoAtt-RNN	94.73	0.02	94.89	0.02	90.99	0.03	86.73	0.05	91.44	0.03
NoAtt-GRU	94.79	0.02	94.85	0.02	91.03	0.03	86.98	0.05	91.34	0.03
NoAtt-LSTM	94.61	0.02	95.78	0.02	90.91	0.03	86.61	0.05	91.29	0.03
Att-RNN	94.69	0.02	94.23	0.02	91.69	0.02	**87.59**	0.04	91.56	0.03
Att-GRU	94.80	0.02	94.83	0.02	91.68	0.02	87.17	0.05	91.68	0.03
Att-LSTM	94.85	0.02	96.00	0.02	91.57	0.03	86.83	0.05	91.72	0.03
Transformer	**95.16**	0.02	95.22	0.02	**92.14**	0.02	86.45	0.05	88.99	0.05
BPE-Soft	95.02	0.02	**96.64**	0.01	91.96	0.03	87.19	0.03	91.21	0.03

Table 5: Evaluation results in word accuracy (Acc, %) and CER. The best results in each language have background color. Many identical values in CER are different, but the difference is irrelevant in Chi-square test.

5.1 Word Accuracy

Table 5 shows that NMT models outperform SMT models in four out of five languages, except for Swedish, when we use word accuracy as the evaluation metric. Compared to the other four languages, we get a huge absolute improvement of 12.04% in Hungarian, improving the word accuracy from 80.1% to 92.14%. We get 0.04%, 0.86%, and 2.99% absolute improvement in German, English, and Icelandic, respectively. Our best NMT result in Swedish is still a little lower than the baseline in word accuracy. We attribute the reason to the dataset size, because Swedish has the smallest training set.

We divide the incorrectly normalized spellings into three groups by checking the normalizations of the test set automatically:

1. *Change*: modern spelling is identical to historical spelling, but the model normalized the historical spelling to another spelling.
2. *Copy*: modern spelling is different from historical spelling, but the model copied the historical spelling as the normalization.
3. *Other*: other types of error.

	English	German	Hungarian	Icelandic	Swedish
Change	22.3	28.5	6.1	33.8	25.0
Copy	22.7	41.7	6.1	20.8	23.6
Other	55	29.8	87.8	45.4	51.4

Table 6: Error distributions (%).

Table 6 gives us the error distributions of the best model in each language. The *Change* and *Copy* errors only account for 12.2% in Hungarian which is reasonable, because the changed rate in Hungarian is only 17.1%. The other four languages still have a lot of *Change* and *Copy* errors. This finding reveals that it is a little bit difficult for the NMT model trained on the data that mixed with changed and unchanged spellings to normalize unchanged spellings. However, there are only very few unchanged translations in the MT task.

Therefore, we explore a hybrid method, combining the NMT-based method and the dictionary-based method. More specifically, we first extract a list of unchanged spellings from the training set. During the evaluation, if a word is in this list, we simply copy it as its normalization. If it is not in the list, we feed it to the NMT models. The results in Table 7 show that this hybrid method improves the accuracy further. In particular, the improvements on Icelandic are around 5%.

	English		German		Hungarian		Icelandic		Swedish	
	Acc	△	Acc	△	Acc	△	Acc	△	Acc	△
NoAtt-RNN	95.92	1.19	95.78	0.90	91.81	0.82	91.92	5.18	91.83	0.39
NoAtt-GRU	95.93	1.14	95.44	0.60	91.87	0.84	91.70	4.71	91.73	0.39
NoAtt-LSTM	95.81	1.20	96.42	0.64	91.75	0.83	91.78	5.17	91.69	0.41
Att-RNN	95.90	1.21	94.93	0.70	92.47	0.77	92.54	4.95	91.94	0.38
Att-GRU	95.99	1.19	95.48	0.66	92.49	0.82	92.25	5.08	92.04	0.36
Att-LSTM	96.02	1.17	96.44	0.44	92.36	0.78	91.76	4.93	92.08	0.36
Transformer	96.33	1.17	95.70	0.48	92.94	0.80	91.60	5.15	89.48	0.49
BPE-Soft	96.19	1.18	96.96	0.32	92.74	0.78	92.14	4.95	91.56	0.35

Table 7: The results of combining the NMT-based method and the dictionary-based method. "△" denotes the absolute improvement on accuracy (%) compared to the NMT-based method.

5.2 CER

With the CER measure, we calculate the number of correctly normalized characters, without considering the word level. CER is similar to the BLEU score in MT, and we evaluate at sub-sequence-level rather than the overall accuracy. When we use CER as the evaluation metric, NMT models get the best results for all five languages, even though some models achieve lower accuracy than the baseline. This result is different from the result of Korchagina (2017). In her paper, if the SMT models are better than the NMT models in word accuracy, these SMT models are better than the NMT models in CER as well. We assume that this may be due to different neural network architectures: they use CNNs while we use RNNs and self-attention networks.

	English	German	Hungarian	Icelandic	Swedish
Changed	1.45	1.07	2.58	1.41	1.32
Incorrect	1.81	1.64	1.78	1.64	1.54

Table 8: The average edit distance of the changed spellings in test set and the average edit distance of the incorrectly normalized changed spellings.

Table 8 shows the edit distance of spellings. For the incorrectly normalized changed spellings, the average edit distance is smaller than 2. In other words, we just need less than two edits to translate an incorrectly normalized spelling into the correct one. In the incorrect normalizations, Swedish has the shortest average edit distance 1.54, and English has the longest average edit distance 1.81.

Intuitively, if a spelling has smaller edit distance, it is easier for the model to normalize this spelling correctly. That is to say, the average edit distance of incorrectly normalized spellings will be larger compared to the average edit distance before normalization. However, Hungarian is the exception in Table 8, which indicates that spellings with longer edit distance are more likely to be normalized close to modern spellings in Hungarian. For example, the edit distance between "mōdanac" and "mondák" is 6, yet the model can normalize it correctly. Although the model normalized "mėgbètèǵeitnç" into "megbetegíteniúk", which is not identical to the modern spelling "megbetegítenék", the edit distance nevertheless decreased from 9 to 2. We hypothesize that this could be due to the fact that Hungarian belongs to a different language family than the other four languages.

Table 9 gives some incorrectly normalized examples from the development set. Most of the edit distances of spellings are longer than 1. In addition to *Change* and *Copy* errors, some historical spellings are quite different from their modern spelling, such as "wett" in English. For the historical word "wett", it is extinct, people just mapped a semantic related modern word to it. "know" has no relations with "wett" in spelling and pronunciation. Characters with different accents also cause mistakes easily. For example, "vetém" in Hungarian and "sér" in Icelandic.

	English	German	Hungarian	Icelandic	Swedish
Historical	alys	julius	vètē	uopn	sielffuer
Normalized	alis	jiues	vetem	opnu	själver
Modern	alice	julius	vetém	vopn	själv
Historical	wett	cohærentz	haila	sier	herrskaper
Normalized	wit	cohaerenz	hajola	sjer	herrskaper
Modern	know	kohärenz	hajla	sér	herrskapen

Table 9: Some incorrectly normalized examples from the development set.

5.3 NMT versus SMT

In the conventional MT tasks, NMT models usually outperform SMT models. The first reason is that the dense embeddings in NMT are powerful representations. The second reason is that NMT models usually consider a larger context compared to SMT models. This is the same in historical spelling normalization. In our experiments, the most obvious example is Hungarian. The absolute improvement is 12.04% in word accuracy. Compared to other languages, Hungarian has the largest token and character vocabularies and the highest changed rate. It also has the longest average token length. Thus, NMT models can represent these larger vocabularies better than SMT models. NMT models are also better at capturing the context information when generating the normalization. For example, the NMT models can normalize a 14-character spelling "aldozatt'oknak" into "áldoza**tuk**nak" correctly, while the SMT models normalize it into "áldoza**tok**nak". In the training set, 'tok' is much more frequent than 'tuk'. Since SMT models are more focused on a local context, the SMT models choose 'tok' rather than 'tuk'.

However, in terms of accuracy, it is still hard for NMT models to exceed SMT models in Swedish. We also find that the performance of NMT models is quite close to the baseline in German which has the second smallest training dataset. We hypothesize that the size of training data is crucial for NMT models to exceed SMT models.

As there is much more test data in Swedish compared to other languages, we test our hypothesis by moving some token pairs from test sets to training sets and development sets. More specifically, we create two new datasets, in which 27,000 and 30,000 token pairs are moved from the beginning of the test set to the training set and the development set, respectively. Both the datasets and results are described in Table 10.

Training	Development	Test	Att-RNN	Att-GRU	Att-LSTM	Transformer
28,327	2,590	33,544	91.56	91.68	91.72	88.99
51,237	6,590	6,544	95.45	94.79	94.97	95.18
57,637	3,190	3,544	96.02	95.77	95.65	95.77

Table 10: The accuracy of different models in Swedish with different dataset settings.

Table 10 shows that the NMT models achieve much higher accuracy with more training data. This result indicates that the performance of NMT models is highly related to the size of training set.

5.4 Different NMT Models

Hypothesis 1 is that the performance gap between vanilla RNNs and GRUs/LSTMs will not be huge. The results in Table 5 reveal that the vanilla RNNs are competitive to the GRUs/LSTMs in this task. *Att-RNN* even performs better than *Att-GRU/LSTM* in Icelandic. However, *Att-RNN* is clearly worse than *Att-LSTM* in German. These results support our Hypothesis 1 well.

Hypothesis 2 states that NMT models with and without attention will not differ a lot. The models with attention are slightly better than models without attention in our experiments, which is in line with the results in Bollmann et al. (2017). However, the gap is quite small. Thus, it fits our Hypothesis 2.

Hypothesis 3 is that transformer models are better than soft-attention-based models. From Table 5,

we can see that *Transformer*, with self-attention and multi-head attention, achieves higher word accuracy in English and Hungarian compared to soft-attention-based models. It is interesting that English and Hungarian have much more training data compared to the other three languages. This result reveals that transformer models need more data to exceed RNN-based models.

Hypothesis 4, finally, states that subword-level models are better than character-level models. Our experimental results for four languages (except Swedish) show that subword-level models are superior to character-level models when the BPE vocabulary is small. In subword-level models, the vocabulary includes all the characters and learned subword units. We try different BPE vocabulary sizes. All the subword-level models are trained on LSTMs. Table 11 gives the results with different BPE sizes. Many historical spellings only have several instances in the training set. The NMT model cannot translate the token well at the token level. Moreover, there is also a data sparsity problem for the subwords when we set a larger BPE vocabulary. We assume that BPE maybe cannot learn rare subword units very well, because of the data sparsity. That is why subword-level models perform better in the conventional MT tasks, which have a much larger training set. We find that the subword-level models perform worse than character-level models when the BPE vocabulary is larger than 300 in all five languages.

BPE-size	English	German	Hungarian	Icelandic	Swedish
0	94.85	96	91.57	86.83	**91.72**
100	**95.02**	**96.64**	91.87	**87.19**	91.21
200	94.91	96.28	91.81	86.89	91.13
300	94.69	96.5	**91.96**	86.76	90.84
500	94.54	96.42	91.52	86.51	90.57
1,000	94.52	96.18	91.44	86.29	89.67
5,000	93.71	95.06	89.43	84.87	85.47

Table 11: Accuracy (%) with different BPE vocabulary sizes. "0" represents the character-level models.

Historical languages are considered as low-resource languages. Hence the result of Hypothesis 4 can be interpreted to mean that subword-level models with a small subword vocabulary can further improve the performance compared to character-level models in low-resource languages.

6 Conclusions and Future Work

In this paper, we explore different NMT models for the historical spelling normalization task in five languages, English, German, Hungarian, Icelandic, and Swedish. We propose four hypotheses on NMT models, which are the general questions to ask when applying NMT models to the historical spelling normalization task.

We find that the performance gap between vanilla RNNs and GRUs/LSTMs is very small, vanilla RNNs are even competitive to GRUs/LSTMs in Hungarian and Icelandic. We demonstrate that the gap between NMT models with or without attention is also slight. We show that the subword-level models with a small subword vocabulary are better than character-level models. However, subword-level models with a larger vocabulary suffer from data sparsity.

When we use word accuracy as the evaluation metric, NMT models can get better results for four languages compared to SMT models. However, all NMT models perform better than SMT models for all five languages when we use CER as the evaluation metric. In addition, the size of the training set is crucial to NMT models. Particularly, transformer models are superior to RNN-based models only when provided with more training data. These findings could contribute to the development of general NMT systems, especially for low-resource languages. Since NMT models are more likely to generate incorrect normalizations of unchanged spellings, we propose a hybrid method using both NMT-based methods and dictionary-based method which improves the performance further.

In the future, we could 1) explore some hard-attention-based models, 2) introduce phoneme knowledge into NMT models, and 3) use sentence pairs for spelling normalization. Compared to soft attention,

hard attention (Xu et al., 2015) only pays attention to one or several specified source word annotations. Aharoni and Goldberg (2017) employ hard monotonic attention for a morphological inflection generation task. The variation between historical spelling and modern spelling is usually monotonic, which is similar to morphological inflection. Thus, hard attention should work well in historical spelling normalization as well.

Many words have changed their spellings, but they keep the same pronunciation. Thus, Bollmann et al. (2017) use an additional grapheme-to-phoneme dictionary in a multi-task learning setting. We can add the phonetic dictionaries as additional training data to improve the performance.

In addition to token-pair-based normalization, Ljubešić et al. (2016) use segment pairs with context information to do spelling normalization. NMT models are powerful in using context information. Thus, training the NMT models on sentence pairs is likely to improve the spelling normalization task further, which introduces more context information.

Acknowledgments

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA.

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.

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, Joachim Bingel, and Anders Søgaard. 2017. Learning attention for historical text normalization by learning to pronounce. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 332–344, Vancouver, Canada. Association for Computational Linguistics.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany. Association for Computational Linguistics.

Rosemarie Fiebranz, Erik Lindberg, Jonas Lindström, and Maria Ågren. 2011. Making verbs count: the research project 'gender and work'and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. 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*, number 133, pages 12–17, Gothenburg, Sweden. Linköping University Electronic Press.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing*, pages 146–155, Varanasi, India. Association for Computational Linguistics.

Manfred Markus. 1999. *Manual of ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)*. Leopold-Franzens-Universitat Innsbruck.

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

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013a. Normalisation of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 163–179, Oslo, Norway. Association for 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 2013*, pages 54–69, Oslo, Norway. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Paul Rayson, Dawn Archer, and Nicholas Smith. 2005. Vard versus word: A comparison of the UCREL variant detector and modern spell checkers on english historical corpora. In *Proceedings of the Corpus Linguistics 2005*, Birmingham, UK.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The icelandic parsed historical corpus (icepahc). In *Proceedings of the 8th International Conference on Language Resources and Evaluations*, pages 1977–1984, Istanbul, Turkey, May. European Language Resources Association.

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*.

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*, pages 124–128, Portland, Oregon, USA, June. Association for Computational Linguistics.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical slovene words with character-based smt. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 58–62, Sofia, Bulgaria. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto, Japan. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Eszter Simon. 2014. Corpus building from old hungarian codices. In *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 224–236. Oxford University Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., Montréal, Canada.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, Lille, France. PMLR.