A Comparison of Character-based Neural Machine Translations Techniques Applied to Spelling Normalization*

Miguel Domingo and Francisco Casacuberta

PRHLT Research Center
Universitat Politècnica de València
midobal@prhlt.upv.es, fcn@prhlt.upv.es

Abstract. The lack of spelling conventions and the natural evolution of human language create a linguistic barrier inherent in historical documents. This barrier has always been a concern for scholars in humanities. In order to tackle this problem, spelling normalization aims to adapt a document's orthography to modern standards. In this work, we evaluate several character-based neural machine translation normalization approaches—using modern documents to enrich the neural models. We evaluated these approaches on several datasets from different languages and time periods, reaching the conclusion that each approach is better suited for a different set of documents.

1 Introduction

Due to the lack of spelling conventions and the nature of human language, orthography in historical texts changes depending on the author and time period. For instance, as Laing [22] pointed out, the data in *LALME* (Linguistic Atlas of Late Medieval English) indicate 45 different forms recorded for the pronoun *it*, 64 for the pronoun *she* and more than 500 for the preposition *through*. This linguistic variation has always been a concern for scholars in humanities [3].

Since historical documents are an important part of our cultural heritage, interest in their effective natural language processing is on the rise [3]. However, the aforementioned linguistic problems suppose an additional challenge. In order to solve these problems, spelling normalization aims to achieve an orthography consistency by adapting a document's spelling to modern standards. Fig. 1 shows an example of normalizing the spelling of a text.

In this work, we evaluate several normalization approaches based on different character-based neural machine translation (NMT) techniques. The rest of this document is structured as follows: Section 2 introduces the related work. Then, in Section 3 we present the different character-based NMT techniques

* Author version of the paper published in *Proceedings of the International Conference on Pattern Recognition. International Workshop on Pattern Recognition for Cultural Heritage, 2021*. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-68787-8_24.

¿Cómo est**ay**s, Ro**z**inante, tan delgado?	¿Cómo est**á**i**s**, Ro**c**inante, tan delgado?
Porque nunca se come, y se trabaja.	Porque nunca se come, y se trabaja.
Pues ¿qué es de la ce**u**ada y de la paja?	Pues ¿qué es de la ce**b**ada y de la paja?
No me de**x**a mi amo ni **v**n bocado.	No me de**j**a mi amo ni **u**n bocado.

Fig. 1: Example of adapting a document's spelling to modern standards. Characters that need to be adapted are denoted in **red**. Its modern versions are denoted in **teal**. Example extracted from *El Quijote* [13].

and normalization approaches. Section 4 describes the experiments conducted in order to assess our proposal. The results of those experiments are presented and discussed in Section 5. Finally, in Section 6, conclusions are drawn.

2 Related Work

Some approaches to spelling normalization include creating an interactive tool that includes spell checking techniques to assist the user in detecting spelling variations [2]. A combination of a weighted finite-state transducer, combined with a modern lexicon, a phonological transcriber and a set of rules [31]. A combination of a list of historical words, a list of modern words and character-based statistical machine translation (SMT) [36]. A multi-task learning approach using a deep bi-LSTM applied at a character level [4]. The application of a token/segment-level character-based SMT approach to normalize historical and user-created words [26]. The use of rule-based MT, character-based SMT (CB-SMT) and character-based NMT (CBNMT) [21]. Domingo and Casacuberta [10] evaluated word-based and character-based MT approaches, finding character-based to be more suitable for this task and that SMT systems outperformed NMT systems. Tang et al. [42], however, compared many different neural architectures and reported that the NMT models are much better than SMT models in terms of CER. Hämäläinen et al. [16] evaluated SMT, NMT, an edit-distance approach, and a rule-based finite state transducer, and advocated for a combination of these approaches to make use of their individual strengths. Finally, Domingo and Casacuberta [11] proposed a method for enriching neural system using modern documents.

Character-based MT strikes to be a solution in MT to reduce the training vocabulary by dividing words into a sequence of characters, and treating each character as if it were a basic unit. Although it was already being researched in SMT [43,27], its interest has increased with NMT. Some approaches to CB-NMT consist in using hierarchical NMT [23], a character level decoder [7], a character level encoder [9] or, for alphabets in which words are composed by fewer characters, by constructing an NMT system that takes advantage of that alphabet [8].

3 Normalization Approaches

In this section, we present the different normalization approaches under study—which are based in several CBNMT techniques—and the CBSMT approach which is used as an additional baseline.

Given a source sentence $\mathbf{x}$, MT aims to find the most likely translation $\hat{\mathbf{y}}$ [5]:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \; Pr(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

3.1 Character-based SMT

CBSMT computes Eq. (1) at a character level, using models that rely on a log-linear combination of different models [29]: namely, phrase-based alignment models, reordering models and language models; among others [46,20].

Since CBSMT approaches are still part of the state of the art for some tasks [21,42,16], we used a CBSMT approach as an additional baseline—we considered as baseline the quality of the original document with respect to its ground truth version, in which the spelling has already been normalized. To that end, considering the document's language as the source language and its normalized version as the target language, we split words into character and applied conventional SMT.

3.2 Character-based NMT

CBNMT models Eq. (1) with a neural network. Its most frequent architecture is based on an encoder-decoder (although others architectures are possible), featuring recurrent networks [1,40], convolutional networks [14] or attention mechanisms [44]. At the encoding state, the source sentence is projected into a distributed representation. Then, at the decoding step, the decoder generates its most likely translation—word by word—using a beam search method [40]. The model parameters are typically estimated jointly on large parallel corpora, via stochastic gradient descent [34,35]. Finally, at decoding time, the system obtains the most likely translation by means of a beam search method.

CBNMT works at a character level: words are split into a sequence of characters and each character is treated as a basic unit. There are different approaches to CBNMT, some of which combine character level strategies with sub-word level strategies. In this work, we made our normalization approaches using the following CBNMT techniques:

- **CBNMT**: This technique uses a character level strategy. Words from both the source and the target are split into characters.
- **SubChar**: This technique combines a sub-word level and a character level strategies. Source words are split into sub-words and target words into characters.

– **CharSub**: This technique combines a character level and a sub-word level strategy. Source words are split into characters and target words into sub-words.

For working at a sub-word level, we use Byte Pair Encoding [37]. This algorithm is a standard in NMT. Based on the intuition that various word classes are translatable via smaller units than words, this technique aims at encoding rare and unknown words as sequences of sub-words units.

Normalization approaches For each CBNMT technique (see Section 3.2), we propose a different normalization approach. Considering the document's language as the source language and its normalized version as the target language, each approach follows a CBNMT strategy. Source and target words are split into either characters or sub-words (depending of the technique) and, then, conventional NMT is applied to train the normalization system.

Additionally, considering how the scarce availability of parallel training data is a frequent problem when working with historical documents [4]—specially for NMT approaches, which need an abundant quantity of parallel training data. Thus, we propose additional normalization approaches (one for each CBNMT) based on Domingo and Casacuberta [11]'s proposal for enriching normalization models using modern documents to generate synthetic data with which to increase the training data. To achieve this, we follow these steps:

1. We train a CBSMT system—since SMT is less affected by the problem of scarce availability of training data— using the normalized version of the training dataset as source and the original version as target, and following the *cbnmt* technique (i.e., splitting all words into characters).
2. We use this system to translate the modern documents, obtaining a new version of the documents which, hopefully, is able to capture the same orthography inconsistencies that the original documents have. This new version, together with the original modern document, conform a synthetic parallel data which can be used as additional training data.
3. We combine the synthetic data with the training dataset, replicating several times the training dataset in order to match the size of the synthetic data and avoid overfitting [6].
4. We use the resulting dataset to train the enriched CBNMT normalization system.

4 Experiments

In this section, we describe the experimental conditions arranged in order to assess our proposal: MT systems, corpora and evaluation metrics.

4.1 Systems

NMT systems were built using `OpenNMT-py` [18]. We used long short-term memory units [15], with all model dimensions set to 512. We trained the system using Adam [17] with a fixed learning rate of 0.0002 [45] and a batch size of 60. We applied label smoothing of 0.1 [41]. At the inference time, we used a beam search with a beam size of 6.

The CBSMT systems used for enriching approaches were trained with `Moses` [19]. Following the standard procedure, we used `SRILM` [39] to estimate a 5-gram language model—smoothed with the improved KneserNey method—and optimized the weights of the log-lineal model with MERT [28].

As baseline, we considered the quality of the original document with respect to its ground truth version, in which the spelling has already been normalized. Additionally, taking into account that CBSMT approaches are still part of the state of the art for some tasks [21,42,16], we used a CBSMT approach as an additional baseline. This approach uses the CBSMT models trained for the enriched CBNMT approaches.

4.2 Corpora

In order to asses our proposal, we made use of the following corpora:

Entremeses y Comedias [13]: A 17^{th} century Spanish collection of comedies by Miguel de Cervantes. It is composed of 16 plays, 8 of which have a very short length.

Quijote [13]: The 17^{th} century Spanish two-volumes novel by Miguel de Cervantes.

Bohorič [25]: A collection of 18^{th} century Slovene texts written in the old Bohorič alphabet.

Gaj [25]: A collection of 19^{th} century Slovene texts written in the Gaj alphabet.

Table 1 shows the corpora statistics. As we can see, the size of the corpora is small. Thus, the need of profiting from modern documents to increase the training data. To that respect, we selected half a million sentences from OpenSubtitles [24]—a collection of movie subtitles in different languages—to use them as monolingual data to enrich the neural systems.

4.3 Metrics

We made use of the following well-known metrics in order to compare our different strategies:

Character Error Rate (CER): number of character edit operations (insertion, substitution and deletion), normalized by the number of characters in the final translation.

		Entremeses y Comedias	Quijote	Bohorič	Gaj
Train	$\|S\|$	35.6K	48.0K	3.6K	13.0K
	$\|T\|$	250.0/244.0K	436.0/428.0K	61.2/61.0K	198.2/197.6K
	$\|V\|$	19.0/18.0K	24.4/23.3K	14.3/10.9K	34.5/30.7K
	$\|W\|$	52.4K	97.5K	33.0K	32.7K
Development	$\|S\|$	2.0K	2.0K	447	1.6K
	$\|T\|$	13.7/13.6K	19.0/18.0K	7.1/7.1K	25.7/25.6K
	$\|V\|$	3.0/3.0K	3.2/3.2K	2.9/2.5K	8.2/7.7K
	$\|W\|$	1.9K	4.5K	3.8K	4.5K
Test	$\|S\|$	2.0K	2.0K	448	1.6K
	$\|T\|$	15.0/13.3K	18.0/18.0K	7.3/7.3K	26.3/26.2K
	$\|V\|$	2.7/2.6K	3.2/3.2K	3.0/2.6K	8.4/8.0K
	$\|W\|$	3.3K	3.8K	3.8K	4.8K
Modern documents	$\|S\|$	500.0K	500.0K	500.0K	500.0K
	$\|T\|$	3.5M	3.5M	3.0M	3.0M
	$\|V\|$	67.3K	67.3K	84.7K	84.7K

Table 1: Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens, $|V|$ for size of the vocabulary and $|W|$ for the number of words whose spelling does not match modern standards. M denotes millions and K thousand. *Modern documents* is the monolingual data used to enrich the neural systems.

Translation Error Rate (TER) [38]: number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.

BiLingual Evaluation Understudy (BLEU) [30]: geometric average of the modified n-gram precision, multiplied by a brevity factor.

In order to ensure consistency with BLEU scores, we used `sacreBLEU` [32]. Additionally, in order to determine whether two systems presented statistically significant differences, we applied approximate randomization tests [33] with $10,000$ repetitions and using a p-value of 0.05.

5 Results

Table 2 presents the results of our experimental session. As baseline, we assessed the spelling differences of the original documents with respect to their normalized version. Additionally, since CBSMT approaches are still part of the state of the art for some tasks [21,42,16], we used a CBSMT approach as a second baseline.

With a few exceptions, all approaches successfully improved the baseline according to all metrics. Those exceptions include all approaches using *Bohorič* and *Gaj*—this behavior was already noticed by Domingo and Casacuberta [10] and is most likely related to the small size of the corpora and the nature of Slovene—and the SubChar approaches with *Entremeses y Comedias*.

From all the proposed approaches, CBNMT yielded the best results in all cases except with *El Quijote*, for which the best results were achieved using the SubChar approach. While this last approach yielded improvements for *El Quijote*, results were slightly worse for *Gaj*, considerably worse for *Bohorič* and

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
Baseline	8.1	28.0	47.0	7.9	19.5	59.4	21.7	49.0	18.0	3.5	12.3	72.6
CBSMT	**1.3**	**4.4**	**91.7**	2.5	**3.0**†	**94.4**†	**2.4**	**8.7**	**80.4**	**1.4**	**5.1**	**88.3**
CBNMT	1.7†	12.0	82.7	2.7	4.3†	93.3‡	29.4	39.5	48.7	31.5‡	36.9	53.1
SubChar	23.3	32.8	54.1	**2.2**†	3.7	93.8‡	36.7	47.7	39.4	32.7	37.3	52.4
CharSub	5.8	18.2	75.2	3.7	5.8	89.8	67.9	83.8	5.3	37.2	48.1	36.3
Enriched CBNMT	1.7†	13.3	79.4†	**2.2**†	4.0†	93.2‡	28.6	38.3	49.5	30.5	35.4†	54.9†
Enriched SubChar	37.8	35.8	59.3	**2.3**†	**3.3**†	**94.9**†	29.5	36.9	51.5	31.5‡	35.9†	54.3†
Enriched CharSub	3.8	15.2	78.9†	**2.3**†	4.1†	93.0‡	27.5	39.6	47.2	29.4	37.2	52.3

Table 2: Experimental results. Baseline system corresponds to considering the original document as the document to which the spelling has been normalized to match modern standards. All results are significantly different between all systems except those denoted with † and ‡ (respectively). Best results are denoted in **bold**.

significantly worse for *Entremeses y Comedias*. The CharSub approach yielded the worst results in all cases. These results, however, behave differently for each task: while they are only slightly worse than CBNMT's result for *El Quijote*, they are significantly worse for *Bohorič*. This shows that not all approaches are equally suited for each task.

Profiting from modern documents to enrich the neural systems improved results in all cases, except for a few exceptions in which they were not significantly different. In the cases of *Bohorič* and *Gaj*, however, these improvements were still worse than the baseline. None the less, results demonstrate how profiting from modern documents successfully improve the neural systems. In a future work, we shall investigate further methods for profiting from these documents.

All in all, except for *El Quijote*—for which the enriched SubChar approached yielded results as good as or better than the CBSMT approach—the CBSMT approach yielded the best results in all cases and according to all metrics. These results are coherent with other results reported in the literature [42,16,11].

5.1 In-depth comparison

In this section, we study the behavior of each normalization approach when normalizing a sentence from each dataset.

Fig. 2 shows an example from *Entremeses y Comedias*. In this case, the normalization only affects two characters. The CBSMT approach is able to successfully normalized those characters. However, it introduces an error (it normalizes the word *Salid* as *Salí*).

Both the CBNMT and the enriched CBNMT approaches behave as the CBSMT approach: they successfully normalized the words *O* and *moço*, but introduce an error normalizing the word *Salid*.

The SubChar approach successfully normalizes the word *moço* but makes a great mistake normalizing the word *O*. Additionally, it fails at normalizing the word *Salid*—it makes the same mistake as the previous approaches–and adds

Original: ¡O mal logrado moço! Salid fuera;
Normalized: ¡Oh mal logrado mozo! Salid fuera;

CBSMT: ¡Oh mal logrado mozo! Salí fuera;

CBNMT: ¡Oh mal logrado mozo! Salí fuera;
Enriched CBNMT: ¡Oh mal logrado mozo! Salí fuera;

SubChar: gueso mal logrado mozo Salí guesto fuera;
Enriched SubChar: ¡Oh mal logrado mozo! ␣␣␣␣␣ ␣␣␣␣␣␣

CharSub: ¡Oh mal logrado mozo! allí fuera;
Enriched CharSub: ¡Oh mal logrado mozo! Salí fuera;

Fig. 2: Example of modernizing a sentence from *Entremeses y Comedias* with all the different approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

a new word between *Salid* and *fuera*. Moreover, both this extra word and the error normalizing *O* correspond to made-up words. This phenomenon has been observed on other tasks [12] and it is most likely due to an incorrect segmentation of a word via the sub-word algorithm used in this approach (see Section 3.2). The enriched version of this approach solves this problem. However, part of the sentence (*Salí fuera;*) is gone. This is a known miss-behavior of neural systems in MT.

Finally, the CharSub approach is able to successfully normalize the words *O* and *moço* but fails at normalizing *Salid*—confusing that word with *allí*. Its enriched version improves that error, but it still is not able to make the correct normalization.

In the example from *El Quijote* (see Fig. 3), there are four characters that need to be normalized. In this case, the CBSMT, CBNMT and Enriched SubChar approaches are able to successfully normalized the whole sentence. The other approaches, however, fail to normalize the word *se* with all of them leaving the original word unnormalized. It is worth noting how, despite that the enriched CBNMT approach offered results equal or better than the CBNMT approached, in this case its normalized is slightly worse.

In the example from *Bohorič* (see Fig. 4), ten characters from four words need to be normalized. As with the previous dataset, the CBSMT approach successfully normalizes the sentence.

Both CBNMT approaches successfully normalized three of the words and make a mistake with two of the characters of one word: one of the, which did not exist in the original word, is still missing and the other one is left unnormalized.

The SubChar approach behaves similarly to the CBNMT approaches—it makes the same mistakes normalizing the word *svédili*—but makes additional

Original: “Para esso se yo vn buen remedio”, dixo el del Bosque;
Normalized: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CBSMT: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CBNMT: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;
Enriched CBNMT: “Para es␣o se yo un buen remedio”, dijo el del Bosque;

SubChar: “Para es␣o se yo un buen remedio”, dijo el del Bosque;
Enriched SubChar: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CharSub: “Para es␣o se yo un buen remedio”, dijo el del Bosque;
Enriched CharSub: “Para es␣o se yo un buen remedio”, dijo el del Bosque;

Fig. 3: Example of modernizing a sentence from *El Quijote* with all the different approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

Original: vadljajo ali lófajo, de bi svédili, kdo jim je kriv te nefrezhe.
Normalized: vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč␣e.

CBSMT: vadljajo ali losajo, da bi izvedeli, kdo jim je kriv te nesreč␣e.

CBNMT: vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.
Enriched CBNMT: vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.

SubChar: vadol ali lozoja, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.
Enriched SubChar: vadljajo ali losajo, da bi ␣zvedili, kdo jim je kriv te nesreč␣e.

CharSub: ugaali ddobra, da bi jim je v va držala.
Enriched CharSub: valjo ali jokajo, da bi ␣zvedili, kdo jim je kri te nesreč␣e.

Fig. 4: Example of modernizing a sentence from *Bohorič* with all the different approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

mistakes normalizing the words *vadljajo* and *lófajo*. The enriched version of this approach does not suffer from this additional mistakes, but it is still unable to normalize the word *svédili* correctly.

Finally, the CharSub approach suffers from a combination of the two phenomenon mentioned in the example from *Entremeses y Comedias*: the generation of made-up words and the disappearance of part of the sentence. The enriched

version solves these problems, but behaves similarly to the SubChar approach (making different mistakes in the normalization of the words *vadljajo* and *lófajo*).

Original: mislili so povsod, de nihče iz zlate vasí beračevati ne more.
Normalized: mislili so povsod, da nihče iz zlate vasi berači␣␣ti ne more.

CBSMT: mislili so povsod, da nihče iz zlate vasi bračevati ne more.

CBNMT: mislili so povsod, da nihče iz zlate vasi bračevati ne more.
Enriched CBNMT: mislili so povsod, da nihče iz zlate vasi bračevati ne more.

SubChar: mislili so povsod, da nihče iz zlate vasi berača␣␣te ne more.
Enriched SubChar: mislili so povsod, da nihče iz zlate vasi beračevati ne more.

CharSub: mislili so povsod, da nihče iz zlate vasi varovati ne more.
Enriched CharSub: mislili so povsod, da nihče iz zlate vasi beračevati ne more.

Fig. 5: Example of modernizing a sentence from *Gaj* with all the different approaches. ␣ denotes a character that has been removed as part of its normalization. Unnormalized characters that should have been normalized and wrongly normalized characters are denoted in red. Characters which were successfully normalized are denoted in teal.

The last example comes from *Gaj* (see Fig. 5). In this case, five characters (two of which need to be removed) from three words are affected by the normalization. All the normalization approaches successfully normalized all words except *beračevati*. The SubChar approach correctly normalizes two out of three characters but changes a character that did not have to be modified. The CharSub approach replaces the word with *varovati* (which has the same suffix but a different meaning). Finally, the rest of the approaches leave the word unnormalized.

In general, the examples show how the CBSMT approach makes less mistake normalizing and how enriching the neural models using synthetic data from modern documents improve the normalizations generated by each approach.

6 Conclusions and future work

In this work, we evaluated different CBNMT normalization approaches, some of which their neural models were enriched using modern documents. We tested our proposal in different datasets, and reached the conclusion that not all approaches are equally suited for each task.

Additionally, while these approaches successfully improved the baseline—except for a few exceptions—CBSMT systems yielded the best results for three our of the four tasks. We believe that this is mostly due to the scarce availability of parallel training data when working with historical documents [4].

As a future work, we would like to further research the use of modern documents to enrich the neural systems. In this work, we used a previously known method in order to assess the different CBNMT approaches under the same circumstances. We should further investigate new methods such as using a data selection approach to find the most suitable data for each corpus.

Acknowledgments

The research leading to these results has received funding from the European Union through *Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER)* from Comunitat Valenciana(2014–2020) under project IDIFEDER/2018/025; from Ministerio de Economía y Competitividad under project PGC2018-096212-B-C31; and from Generalitat Valenciana (GVA) under project PROMETEO/2019/121.We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2015), *arXiv:1409.0473*
2. Baron, A., Rayson, P.: VARD2: A tool for dealing with spelling variation in historical corpora. Postgraduate conference in corpus linguistics (2008)
3. Bollmann, M.: Normalization of Historical Texts with Neural Network Models. Ph.D. thesis, Sprachwissenschaftliches Institut, Ruhr-Universität (2018)
4. Bollmann, M., Søgaard, A.: Improving historical spelling normalization with bi-directional lstms and multi-task learning. In: Proceedings of the International Conference on the Computational Linguistics. pp. 131–139 (2016)
5. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2), 263–311 (1993)
6. Chatterjee, R., Farajian, M.A., Negri, M., Turchi, M., Srivastava, A., Pal, S.: Multi-source neural automatic post-editing: Fbks participation in the wmt 2017 ape shared task. In: Proceedings of the Second Conference on Machine Translation. pp. 630–638 (2017)
7. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 1693–1703 (2016)
8. Costa-Jussà, M.R., Aldón, D., Fonollosa, J.A.: Chinese–spanish neural machine translation enhanced with character and word bitmap fonts. Machine Translation **31**, 35–47 (2017)
9. Costa-Jussà, M.R., Fonollosa, J.A.: Character-based neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 357–361 (2016)
10. Domingo, M., Casacuberta, F.: Spelling normalization of historical documents by using a machine translation approach. In: Proceedings of the Annual Conference of the European Association for Machine Translation. pp. 129–137 (2018)

11. Domingo, M., Casacuberta, F.: Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents. In: Proceedings of the International Workshop on Pattern Recognition for Cultural Heritage. pp. 59–69 (2019)
12. Domingo, M., García-Martínez, M., Peris, Á., Helle, A., Estela, A., Bié, L., Casacuberta, F., Herranz, M.: A user study of the incremental learning in NMT. In: Proceedings of the European Association for Machine Translation. pp. 319–328 (2020)
13. F. Jehle, F.: Works of Miguel de Cervantes in Old- and Modern-spelling. Indiana University Purdue University Fort Wayne (2001)
14. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning (2017), *arXiv:1705.03122*
15. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. Neural computation **12**(10), 2451–2471 (2000)
16. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Normalizing early english letters to present-day english spelling. In: Proceedings of the Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 87–96 (2018)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: Proceedings of the Association for Computational Linguistics: System Demonstration. pp. 67–72 (2017)
19. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 177–180 (2007)
20. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. pp. 48–54 (2003)
21. Korchagina, N.: Normalizing medieval german texts: from rules to deep learning. In: Proceedings of the Nordic Conference on Computational Linguistics Workshop on Processing Historical Language. pp. 12–17 (2017)
22. Laing, M.: The linguistic analysis of medieval vernacular texts: Two projects at edinburgh'. In: Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, edited by M. Rissanen, M. Kytd, and S. Wright. St Catharines College Cambridge. vol. 25427, pp. 121–141 (1993)
23. Ling, W., Trancoso, I., Dyer, C., Black, A.W.: Character-based neural machine translation. arXiv preprint arXiv:1511.04586 (2015)
24. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: Proceedings of the International Conference on Language Resources Association. pp. 923–929 (2016)
25. Ljubešić, N., Zupan, K., Fišer, D., Erjavec, T.: Dataset of normalised slovene text KonvNormSl 1.0. Slovenian language resource repository CLARIN.SI (2016), http://hdl.handle.net/11356/1068
26. Ljubešic, N., Zupan, K., Fišer, D., Erjavec, T.: Normalising slovene data: historical texts vs. user-generated content. In: Proceedings of the Conference on Natural Language Processing. pp. 146–155 (2016)
27. Nakov, P., Tiedemann, J.: Combining word-level and character-level models for machine translation between closely-related languages. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 301–305 (2012)

28. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 160–167 (2003)
29. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 295–302 (2002)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
31. Porta, J., Sancho, J.L., Gómez, J.: Edit transducers for spelling variation in old spanish. In: Proceedings of the workshop on computational historical linguistics. pp. 70–79 (2013)
32. Post, M.: A call for clarity in reporting bleu scores. In: Proceedings of the Third Conference on Machine Translation. pp. 186–191 (2018)
33. Riezler, S., Maxwell, J.T.: On some pitfalls in automatic evaluation and significance testing for mt. In: Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 57–64 (2005)
34. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics pp. 400–407 (1951)
35. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
36. Scherrer, Y., Erjavec, T.: Modernizing historical slovene words with character-based smt. In: Proceedings of the Workshop on Balto-Slavic Natural Language Processing. pp. 58–62 (2013)
37. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 1715–1725 (2016)
38. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the Association for Machine Translation in the Americas. pp. 223–231 (2006)
39. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing. pp. 257–286 (2002)
40. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 27, pp. 3104–3112 (2014)
41. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
42. Tang, G., Cap, F., Pettersson, E., Nivre, J.: An evaluation of neural machine translation models on historical spelling normalization. In: Proceedings of the International Conference on Computational Linguistics. pp. 1320–1331 (2018)
43. Tiedemann, J.: Character-based PSMT for closely related languages. In: Proceedings of the Annual Conference of the European Association for Machine Translation. pp. 12–19 (2009)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
45. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X.,

Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation (2016), *arXiv:1609.08144*

46. Zens, R., Och, F.J., Ney, H.: Phrase-based statistical machine translation. In: Proceedings of the Annual German Conference on Advances in Artificial Intelligence. vol. 2479, pp. 18–32 (2002)