Enriching Character-based Neural Machine Translation with Modern Documents for Achieving an Orthography Consistency in Historical Documents*

Miguel Domingo and Francisco Casacuberta

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

Abstract. The nature of human language and the lack of a spelling convention make historical documents hard to handle for natural language processing. Spelling normalization tackles this problem by adapting their spelling to modern standards in order to get an orthography consistency. In this work, we compare several character-based machine translation approaches, and propose a method to profit from modern documents to enrich neural machine translation models. We tested our proposal with four different data sets, and observed that the enriched models successfully improved the normalization quality of the neural models. Statistical models, however, yielded a better result.

1 Introduction

The linguistic variation in historical documents has always been a concern for scholars in humanities [3]. On the one hand, human language evolves over time. On the other hand, spelling conventions were not created until recently. Therefore, orthography changes depending on the author and time period. Sometimes, this variety is astonishing. Laing [21] pointed out that, for instance, 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*.

Historical documents are an important part of our cultural heritage. Thus, interest in effective natural language processing for these documents is on the rise [3]. However, the aforementioned linguistic problems suppose an additional challenge. Spelling normalization aims to solve these problems. Its goal is to achieve an orthography consistency by adapting the document's spelling to modern standards. Fig. 1 shows an example.

In this work, we compare several normalization approaches that rely on character-based machine translation (MT), and propose a method for enriching neural machine translation (NMT) systems by profiting from modern documents. Our main contributions are as follow:

* Author version of the paper published in *Proceedings of the International Conference on Pattern Recognition. International Workshop on Pattern Recognition for Cultural Heritage, 2019*. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-30754-7_7.

Bien responde la esperança	Bien responde la esperanza
en que engañado he viuido	en que engañado he vivido
al cuydado que he tenido	al cuidado que he tenido
de tu estudio y tu criança!	de tu estudio y tu crianza!

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 [11].

- Comparison of several character-based MT normalization approaches.
- New character-based NMT approach enriched with modern documents.

The rest of this document is structured as follows: Section 2 introduces the related work. Then, in Section 3 we present the different 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]. Porta et al. [32] made use of a weighted finite-state transducer, combined with a modern lexicon, a phonological transcriber and a set of rules. Scherrer and Erjavec [37] combined a list of historical words, a list of modern words and character-based statistical machine translation (SMT). Bollman and Søgaard [4] took a multi-task learning approach using a deep bi-LSTM applied at a character level. Ljubevsic et al. [25] applied a token/segment-level character-based SMT approach to normalize historical and user-created words. Korchagina [20] made use of rule-based MT, character-based SMT (CBSMT) and character-based NMT (CBNMT). 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. [43], however, compared many different neural architectures and reported that the NMT models are much better than SMT models in terms of CER. Finally, Hämäläinen et al. [15] 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.

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 [44,26], its interest has increased with NMT. Some approaches to CBNMT consist in using hierarchical NMT [22], 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].

Backtranslation [38] is a useful technique to increase the training data by creating synthetic text from monolingual data. It has become the norm in MT when build-

ing state-of-the-art NMT systems, especially in resource-poor scenarios [31]. Given a monolingual corpus in the target language, and an MT system trained to translate from target to source, the synthetic data is generated by translating the monolingual corpus with the MT system. After that, the synthetic data is used as the source part of the corpus, and the monolingual data as the target part. Finally, this new corpus is mixed with the available training data in order to train a new MT system.

3 Normalization Approaches

In this section, we review different approaches to tackle the orthography problem inherent in historical documents and achieve a spelling consistency, and propose a new method which profits from modern documents to enrich its system.

These approaches rely on MT, which aims at finding the most likely translation $\hat{\mathbf{y}}$ [5] for a given source sentence $\mathbf{x}$:

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

3.1 Existing Approaches

Character-based SMT CBSMT focuses to compute Eq. (1) at a character level, using models that rely on a log-linear combination of different models [28]: namely, phrase-based alignment models, reordering models and language models; among others [47,19].

Considering the document's language as the source language and its normalized version as the target language, this approach follows a CBSMT strategy. In order to have the same conditions in both SMT and NMT approaches, the character-based strategy that is usually followed is the simplest approach: to split words into characters and, then, apply conventional SMT.

Character-based NMT Like CBSMT, CBNMT focuses to compute Eq. (1) at a character level, but modeling this expression with a neural network. This neural network usually follows an encoder-decoder architecture, featuring recurrent networks [1,41], convolutional networks [13] or attention mechanisms [45]. Model parameters are jointly estimated on large parallel corpora, using stochastic gradient descent [35,36]. At decoding time, the system obtains the most likely translation using a beam search method.

Like with the CBSMT approach, this normalization approach considers the original language as the source and its normalized version as the target, and focuses on a character-based strategy. The difference is that this approach follows an NMT strategy in stead of a SMT one.

3.2 Character-based NMT Enriched with Modern Documents

Our normalization proposal is an extension of the CBNMT approach (see Section 3.1). The scarce availability of parallel training data is a frequent problem when working with

historical documents [4]. This problem is specially troublesome for NMT approaches, which need an abundant quantity of parallel training data. To tackle this problem, we propose to use modern documents to enrich the NMT systems.

Following a backtranslation strategy [38], we propose to enrich the NMT models using modern documents to create synthetic data. With this aim, 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.
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 MT Systems

We trained our SMT systems with `Moses` [18], following the standard procedure: we estimated a 5-gram language model—smoothing it with the improved KneserNey method—using `SRILM` [40], and optimized the weights of the log-lineal model with MERT [27].

NMT systems were built using `OpenNMT-py` [17]. We used long short-term memory units [14], with all model dimensions set to 512. We trained the system using Adam [16] with a fixed learning rate of 0.0002 [46] and a batch size of 60. We applied label smoothing of 0.1 [42]. At the inference time, we used a beam search with a beam size of 6.

Finally, 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. Nonetheless, as a second baseline, we implemented a statistical dictionary. Using `mgiza` [12], we computed *IBM's model 1* [29] to obtain word alignments from source and target of the training set. Then, for each source word, we selected as its translation the target word which had the highest alignment probability with that source word. Finally, at translation time, we translated each source word with the translation that appeared in the dictionary. If a given word did not appear in the dictionary, then we left it untranslated.

4.2 Corpora

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

Entremeses y Comedias [11]: 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 [11]: The 17^{th} century Spanish two-volumes novel by Miguel de Cervantes.

Bohorič [24]: A collection of 18^{th} century Slovene texts written in the old Bohorič alphabet.

Gaj [24]: A collection of 19^{th} century Slovene texts written in the Gaj alphabet.

As reflected in Table 1, the size of the corpora is small. Thus, the use of backtranslation to increase the training data. As *modern documents*, we selected half a million sentences from OpenSubtitles [23], a collection of movie subtitles in different languages. We selected the same Spanish sentences for *Entremeses y Comedias* and *Quijote*, and the same Slovene sentences for *Bohoric* and *Gaj*.

		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.

4.3 Metrics

We made use of the following well-known metrics in order to assess our proposal:

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

Translation Error Rate (TER) [39]: 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 consistent BLEU scores, we used `sacreBLEU` [33]. Additionally, we applied approximate randomization tests [34]—with $10,000$ repetitions and using a p-value of 0.05—to determine whether two systems presented statistically significant differences.

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, as a second baseline, we made use of a statistical dictionary for normalizing the spelling. With one exception in which CER yielded worse results, the statistically dictionary presented significant gains for all data sets according to all the metrics (up to 5 points according to CER, 28 points according to TER and 38 points according to BLEU).

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
SD	7.8	18.9	66.8	3.9	5.5	89.3	16.2	20.7	56.1	7.6	8.8	79.8
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	2.4	8.0	84.8	4.2	7.6	85.1	37.0	45.1	40.1	39.0	42.5	45.4
Enriched CBNMT	1.9	7.2	85.9	3.3	4.5	91.9	28.7	37.3	49.0	36.4	40.7	47.3

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. SD is the statistical dictionary. All results are significantly different between all systems. Best results are denoted in **bold**.

The CBSMT approach yielded the most significant improvements for all data sets and according to all the metrics, with gains of up to 19 points according to CER, 40 points according to TER and 62 points according to BLEU.

With two exception, the CBNMT approach yielded better results than both baselines, but worse than the CBSMT approach. Those exceptions were with *Bohorič* and *Gaj*, for which it yielded worse results than both baselines according to all the metrics. This behavior was already noticed by Domingo and Casacuberta [10]. Most likely, it is related with the small size of the corpora, and the nature of the Slovene language—specially in the case of *Bohorič*, whose documents were written while the Slovene language was having a big restructuring.

Following a backtranslation approach to enrich the neural systems using modern documents significantly improved the results, yielding gains of up to 8 points according to CER and TER, and 9 points according to BLEU. However, in the case of *Bohorič* and *Gaj*, these results are still worse than both baselines according to all the metrics. Nonetheless, these results are encouraging, since they show that we can profit from modern documents to improve neural systems. We shall further investigate this approach in a future work.

5.1 Analysis

Fig. 2 shows an example of normalizing the spelling of a sentence from *Bohorič*.

Original: dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.
Normalized: dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.

SD: dobro manengo, da meni drugi, kakor je kristus cerkvene, in kakor ima katholshka cerkev navado drugi.
CBSMT: dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
CBNMT: dobro manengo, da otže krztiti, kokor je krstiti.
Enriched CBNMT: dobro manengo, da otže krstiti, kakor je kriztus goripostavil, in koker ima katoliška cerkev nava

Fig. 2. Example of modernizing a sentence from *Bohorič* with all the different approaches. Unnormalized characters that should have been normalized are denoted in red. Characters which were successfully normalized are denoted in teal.

From all the corpora, *Bohorič* has the biggest differences in its orthography due to the Slovene language having a big restructuring in the period in which their documents were written. Therefore, 23 changes are needed in order to update its spelling to match modern Slovene standards.

The statistical dictionary was able to correct 11 of these errors. However, since it is a word-based approach, it introduced more mistakes than it was able to correct: while only a few characters of some words needed a change in their spelling, the statistical dictionary suggested new words.

For this example, the CBSMT approach was able to achieve a perfect normalization.

The CBNMT approach was able to correct 2 characters, and successfully determined that the combination *sh* should be normalized as a single character. However, it made a wrong correction. Furthermore, half of the sentence is gone. This is a known miss-behavior of neural systems in MT.

Finally, the enriched CBNMT approach was able to handle the neural miss-behavior. Although the last few characters are still missing. Moreover, most of the unnormalized characters have been successfully corrected. A behavior worth noting, however, is how the system was able to successfully normalized the first appearance of the word *koker*, but not its second appearance.

6 Conclusions and Future Work

In this work, we proposed a normalization method based on backtranslation, to enrich CBNMT systems using modern documents. We tested our proposal in different data sets, observing significant gains for all metrics.

Additionally, we compared several normalization approaches, reaching the conclusion than CBSMT systems are more suitable for this task. We believe that this is specially true 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 randomly selected 500 thousand lines from modern documents, in order to balance the quantity between synthetic and real data, and use the same data for corpora who belonged to the same language. We should further investigate about how to balance synthetic and real data. Additionally, instead of randomly selecting the data, we would like to use 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 Valencia (2014–2020) under project *Sistemas de frabricacin inteligentes para la indstria 4.0* (grant agreement IDIFEDER/2018/025); and from Ministerio de Economía y Competitividad (MINECO) under project *MISMIS-FAKEnHATE* (grant agreement PGC2018-096212-B-C31). 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. F. Jehle, F.: Works of Miguel de Cervantes in Old- and Modern-spelling. Indiana University Purdue University Fort Wayne (2001)
12. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. In: Proceedings of the Association for Computational Linguistics Software Engineering, Testing, and Quality Assurance Workshop. pp. 49–57 (2008)
13. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning (2017), *arXiv:1705.03122*
14. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. Neural computation **12**(10), 2451–2471 (2000)
15. 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)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. Ling, W., Trancoso, I., Dyer, C., Black, A.W.: Character-based neural machine translation. arXiv preprint arXiv:1511.04586 (2015)
23. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: Proceedings of the International Conference on Language Resources Association (2016)
24. 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
25. 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)

26. 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)
27. 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)
28. 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)
29. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational linguistics **29**(1), 19–51 (2003)
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. Poncelas, A., Shterionov, D., Way, A., Maillette de Buy Wenniger, G., Passban, P.: Investigation backtranslation in neural machine translation. In: Proceedings of the Annual Conference of the European Association for Machine Translation. pp. 249–258 (2018)
32. 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)
33. Post, M.: A call for clarity in reporting bleu scores. In: Proceedings of the Third Conference on Machine Translation. pp. 186–191 (2018)
34. 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)
35. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics pp. 400–407 (1951)
36. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533 (1986)
37. Scherrer, Y., Erjavec, T.: Modernizing historical slovene words with character-based smt. In: Proceedings of the Biennial International Workshop on Balto-Slavic Natural Language Processing. pp. 58–62 (2013)
38. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709 (2015)
39. 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)
40. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing. pp. 257–286 (2002)
41. 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)
42. 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)
43. 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)
44. 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)
45. 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)

46. 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*
47. 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)