

Applying the Transformer to Character-level Transduction

Shijie Wu⁵ Ryan Cotterell^{5, D} Mans Hulden²

⁵Johns Hopkins University ^DUniversity of Cambridge

¹ETH Zürich ²University of Colorado Boulder

shijie.wu@jhu.edu ryan.cotterell@inf.ethz.ch mans.hulden@colorado.edu

Abstract

The transformer (Vaswani et al., 2017) has been shown to outperform recurrent neural network-based sequence-to-sequence models in various word-level NLP tasks. The model offers other benefits as well: It trains faster and has fewer parameters. Yet for character-level transduction tasks, e.g. morphological inflection generation and historical text normalization, few shows success on outperforming recurrent models with the transformer. In an empirical study, we uncover that, in contrast to recurrent sequence-to-sequence models, the batch size plays a crucial role in the performance of the transformer on character-level tasks, and we show that with a large enough batch size, the transformer does indeed outperform recurrent models. We also introduce a simple technique to handle feature-guided character-level transduction that further improves performance. With these insights, we achieve state-of-the-art performance on morphological inflection and historical text normalization. We also show that the transformer outperforms a strong baseline on two other character-level transduction tasks: grapheme-to-phoneme conversion and transliteration. Code is available at <https://github.com/shijie-wu/neural-transducer>.

1 Introduction

The transformer (Vaswani et al., 2017) has become a popular architecture for sequence-to-sequence transduction in NLP. It has achieved state-of-the-art performance on a range of common word-level transduction tasks: neural machine translation (Barrault et al., 2019), question answering (Devlin et al., 2019) and abstractive summarization (Dong et al., 2019). In addition, the transformer forms the backbone of the widely-used BERT (Devlin et al., 2019). Yet for character-level transduction tasks like morphological inflection, the dominant model

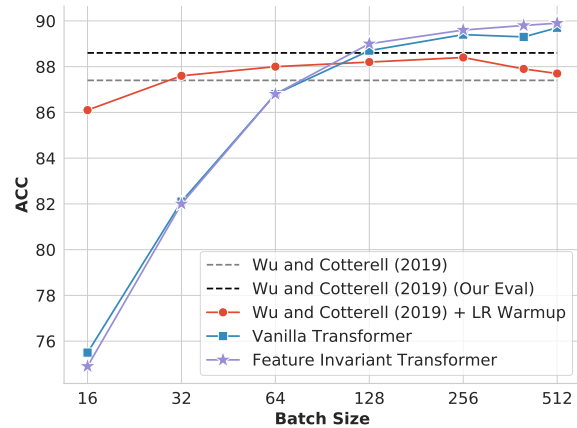


Figure 1: Development set accuracy for 5 languages on morphological inflection with different batch sizes. We evince our two primary contributions: (1) we set the **new state of the art** morphological inflection using the transformer and (2) we demonstrate the transformer’s **dependence on the batch size**.

has remained a recurrent neural network-based sequence-to-sequence model with attention (Cotterell et al., 2018). This is not for lack of effort—but rather, it is the case that the transformer has consistently underperformed in experiments on average (Tang et al., 2018b).¹ As anecdotal evidence of this, we note that in 2019, the most recent addition of the SIGMORPHON shared task on cross-lingual transfer for morphological inflection, no participating system was based on the transformer (McCarthy et al., 2019).

Character-level transduction tasks often have fewer data than their word-level counterparts: In contrast to machine translation, where millions of training samples are available, the 2018 SIGMORPHON shared task (Cotterell et al., 2018) high-resource setting only provides $\approx 10k$ training examples per language. It is also not obvious that

¹This claim is also based on the authors’ personal communication with other researchers in morphology in the corridors of conferences and through email.

	Vanilla										Feature Invariant									
Token	<s>	V	V.PTCP	PST	s	m	e	a	r	</s>	<s>	V	V.PTCP	PST	s	m	e	a	r	</s>
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Position	0	1	2	3	4	5	6	7	8	9	0	0	0	0	1	2	3	4	5	6
												+	+	+	+	+	+	+	+	
Type												F	F	F	C	C	C	C	C	

Figure 2: Handling of feature-guided character-level transduction with special position and type embeddings in the encoder. F denotes features while C denotes characters. We use morphological inflection as an example, inflecting *smear* into its past participle form, *smear*_d.

non-recurrent architectures such as the transformer should provide an advantage at many character-level tasks: For instance, Gehring et al. (2017) and Vaswani et al. (2017) suggest that transformers (and convolutional models in general) should help remember long-range dependencies better. In the case of morphology, none of these considerations seem relevant: inflecting a word (a) requires little capacity to model long-distance dependencies and is largely monotonic transduction; (b) it involves no semantic disambiguation, the tokens in question being letters; (c) it is not a task for which parallelization during training appears to help, since training time has never been an issue in morphology tasks.²

In this work, we provide state-of-the-art numbers for morphological inflection and historical text normalization, a novel result in the literature. We also show the transformer outperforms a strong recurrent baseline on two other character-level tasks: grapheme-to-phoneme (g2p) conversion and transliteration. We find that a single hyperparameter, batch size, is largely responsible for the previous poor results. Despite having fewer parameters, the transformer outperforms the recurrent sequence-to-sequence baselines on all four tasks. We conduct a short error analysis on the task of morphological inflection to round out the paper.

2 The Transformer for Characters

The Transformer. The transformer, originally described by Vaswani et al. (2017), is a self-attention-based encoder-decoder model. The encoder has N layers, consisting of a multi-head self-attention layer and a two-layer feed-forward layer with ReLU activation, both equipped with a skip connection. The decoder has a similar structure as the encoder except that, in each decoder

layer between the self-attention layer and feed-forward layer, a multi-head attention layer attends to the output of the encoder. Layer normalization (Ba et al., 2016) is applied to the output of each skip connection. Sinusoidal positional embeddings are used to incorporate positional information without the need for recurrence or convolution. Here, we describe two modifications we make to the transformer for character-level tasks.

A Smaller Transformer. As the dataset sizes in character-level transduction tasks are significantly smaller than in machine translation, we employ a smaller transformer with $N = 4$ encoder-decoder layers. We use 4 self-attention heads. The embedding size is $d_{model} = 256$ and the hidden size of the feed-forward layer is $d_{FF} = 1024$. In the preliminary experiments, we found that using layer normalization before self-attention and the feed-forward layer performed slightly better than the original model. It is also the default setting of a popular implementation of the transformer (Vaswani et al., 2018). The transformer alone has around 7.37M parameters, excluding character embeddings and the linear mapping before the softmax layer. We decode the model left to right in a greedy fashion.

Feature Invariance. Some character-level transduction is guided by features. For example, in the case of morphological reinflection, the task requires a set of morphological attributes that control what form a citation form is inflected into (see Fig. 2 for an example). The order of the features is irrelevant. In a recurrent neural network, features are input in some predefined order as special characters and pre- or postpend to the input character sequence representing the citation form. The same is true for a vanilla transformer model, as shown on the left-hand side of Fig. 2. This leads to different relative distances between a character and

²Many successful CoNLL-SIGMORPHON shared task participants report training their models on laptop CPUs.

LS	β_2	Vanilla	Feature Invariant
0	0.999	89.34	89.80
0	0.98	89.62	89.92
0.1	0.999	89.48	90.02
0.1	0.98	89.98	90.28

Table 1: Average development accuracy on morphological inflection with different LS and β_2 , which denote hyperparameter of label smoothing and Adam optimizer respectively.

a set of features.³ To avoid such an inconsistency, we propose a simple remedy: We set the positional encoding of features to 0 and only start counting the positions for characters. Additionally, we add a special token to indicate whether a symbol is a word character or a feature. The right-hand side of Fig. 2 evinces how we have the same relative distance between characters and features.

3 Empirical Findings

Tasks. We consider four character-level transduction tasks: morphological inflection, grapheme-to-phoneme conversion, transliteration, and historical text normalization. For morphological inflection, we use the 2017 SIGMORPHON shared task data (Cotterell et al., 2017) with 52 languages. The performance is evaluated by accuracy (ACC) and edit distance (Dist). For the g2p task, we use the unstressed CMUDict (Weide, 1998) and NETalk (Sejnowski and Rosenberg, 1987) resources. We use the splits from Wu et al. (2018). We evaluate under word error rate (WER) and phoneme error rate (PER). For transliteration, we use the NEWS 2015 shared task data (Zhang et al., 2015).⁴ For historical text normalization, we follow Bollmann (2019) and use datasets for Spanish (Sánchez-Martínez et al., 2013), Icelandic and Swedish (Pettersson et al., 2013), Slovene (Scherrer and Erjavec, 2013, 2016; Ljubešić et al., 2016), Hungarian and German (Pettersson, 2016).⁵ We evaluate using accuracy (ACC) and character error rate of incorrect prediction (CER_i).

Optimization. We use Adam (Kingma and Ba, 2014) with a learning rate of 0.001 and an inverse square root learning rate scheduler (Vaswani et al., 2017) with 4k steps during the warm-up. We train

³While the features could be encoded with a binary vector followed by MLP, it introduces a representation bottleneck for encoding features.

⁴We do not have access to the test set.

⁵We do not include English due to licensing issues.

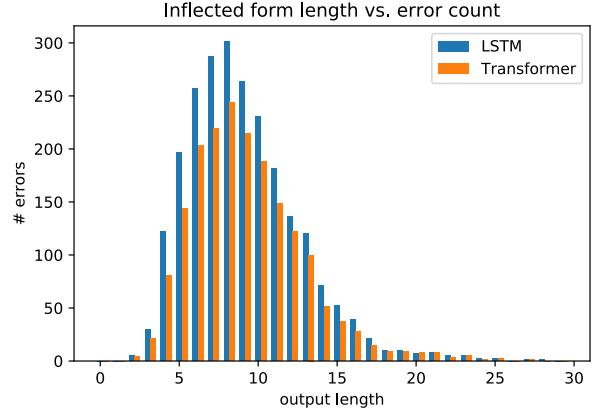


Figure 3: Distribution of incorrectly inflected forms in the test set of the inflection task over all 52 languages grouped by desired output word length.

the model for 20k gradient updates and save and evaluate the model every 400 gradient updates. We select the best model out of 50 checkpoints based on development set accuracy. The number of gradient updates and checkpoints are roughly the same as Wu and Cotterell (2019), the single model state of the art on the 2017 SIGMORPHON dataset. We use their model as a baseline model. For all experiments, we use a single predefined random seed.

3.1 A Controlled Hyperparameter Study

To demonstrate the importance of hyperparameter tuning for the transformer on character-level tasks, we perform a small controlled hyperparameter study. This is important since researchers had previously failed to achieve high-performing results with the transformer on character-level tasks. Here, we look at morphological inflection on the five languages in the 2017 SIGMORPHON dataset where submitted systems performed the worst: Latin, Faroese, French, Hungarian, and Norwegian (Nynorsk). We set the dropout to 0.3, β_2 of Adam to 0.999 (the default value), and do not use label smoothing. We do not tune any other hyperparameter except the following three hyperparameters.

The Importance of Batch Size. While recurrent models like Wu and Cotterell use a batch size of 20, halving the learning rate when stuck and employing early stopping, we find that a less aggressive learning rate scheduler, allowing the model to train longer, outperforms these hyperparameters. Fig. 1 shows that the *single most important hyperparameter* when training is the batch size. The transformer performance increases steadily as the batch size is increased, similarly to what Popel and Bojar (2018)

	ACC	Dist
Silfverberg et al. (2017)*	92.97	0.170
Wu et al. (2018)	93.60	0.128
Wu and Cotterell (2019)	94.40	0.113
Wu and Cotterell (2019) (Our eval)	94.81	0.123
Makarov et al. (2017)*	95.12	0.100
Bergmanis et al. (2017)*	95.32	0.100
Transformer (Dropout = 0.3)	95.59	0.088
Transformer (Dropout = 0.1)	95.56	0.090

Table 2: Average test performance on morphological inflection of Transformer against models from the literature. * denotes model ensembling.

observe for machine translation. The transformer only outperforms the recurrent baseline when the batch size is above 128. Note that the model of Wu and Cotterell has 8.66M parameters, 17% more than the transformer model. To get an apples-to-apples comparison, we apply the same learning rate scheduler to Wu and Cotterell; this does not yield similar improvements and underperforms with respect to the traditional learning rate scheduler. Our feature invariant transformer also outperforms the vanilla transformer model. We set the batch size to 400 for our main experiments. Note the batch size of 400 is especially large (4% of training data) consider the training size is only 10k.

Other Hyperparameters. Vaswani et al. (2017) apply label smoothing (Szegedy et al., 2016) of 0.1 to the transformer model and show that it hurts perplexity, but improves BLEU scores for machine translation. Instead of the default 0.999 β_2 for Adam, Vaswani et al. (2017) use 0.98 and we find that both choices benefit character-level transduction tasks as well (see Tab. 1).

3.2 New State-of-the-Art Results

We train our feature invariant transformer on the four character-level tasks, exhibiting state-of-the-art results on morphological inflection and historical text normalization.

Morphological Inflection. As shown in Tab. 2, the feature invariant transformer produces state-of-the-art results on the 2017 SIGMORPHON shared tasks, improving upon ensemble-based systems by 0.27 points. We observe that as the dataset decreases in size, a model with a larger dropout value performs slightly better. A brief tally of phenomena that are difficult to learn for many machine learning models, categorized along typical linguistic dimensions (such as word-internal sound changes,

	ACC	CER _i	ACC ^s	CER _i ^s
Ljubešić et al. (2016)	91.78	0.392	90.37	0.360
Ljubešić et al. (2016) (LM)	91.56	0.399	89.93	0.368
Bollmann (2018)	91.27	0.381	89.73	0.350
Tang et al. (2018a)	91.67	0.389	90.32	0.358
Flachs et al. (2019)	-	-	90.06	-
Transformer (Dropout = 0.3)	91.30	0.340	89.99	0.330
Transformer (Dropout = 0.1)	91.85	0.352	90.61	0.334

Table 3: Average test performance on historical text normalization of Transformer against models from the literature. ^s denote subset of dataset as Flachs et al. (2019) only experiment with subset of languages.

	WER	PER	ACC	MFS
Wu et al. (2018)	28.20	0.068	41.10	0.894
Wu and Cotterell (2019)	28.20	0.069	41.20	0.895
Transformer (Dropout = 0.3)	28.08	0.070	43.39	0.897
Transformer (Dropout = 0.1)	27.63	0.069	41.35	0.891

Table 4: Average test performance on Grapheme-to-Phoneme and dev performance on Transliteration of Transformer against models from the literature.

vowel harmony, circumfixation, ablaut, and umlaut phenomena) fail to reveal any consistent pattern of advantage to the transformer model. In fact, errors seem to be randomly distributed with an overall advantage of the transformer model. Curiously, errors grouped along the dimension of word length reveal that as word forms grow longer, the transformer advantage shrinks (Fig. 3).

Historical Text Normalization. Tab. 3 shows that the transformer model with dropout of 0.1, like with morphological inflection, improves upon the previous state of the art, although the model with a dropout of 0.3 yields a slightly better CER_i.

G2P and Transliteration. Tab. 4 shows that the transformer outperforms previously published strong recurrent models on two tasks despite having fewer parameters. A dropout rate of 0.3 yields significantly better performance on the transliteration task while a dropout rate of 0.1 is stronger on the g2p task. This shows that transformers can and do outperform recurrent transducers on common character-level tasks when properly tuned.

4 Related Work

Character-level transduction is largely dominated by attention-based LSTM sequence-to-sequence (Luong et al., 2015) models (Cotterell et al., 2018). Character-level transduction tasks usually involve input-output pairs that share large substrings and

alignments between these are often monotonic. Models that address the task tend to focus on exploiting such structural bias. Instead of learning the alignments, [Aharoni and Goldberg \(2017\)](#) use external monotonic alignments from the SIGMORPHON 2016 shared task baseline [Cotterell et al. \(2016\)](#). [Makarov et al. \(2017\)](#) use this approach to win the CoNLL-SIGMORPHON 2017 shared task on morphological inflection ([Cotterell et al., 2017](#)). [Wu et al. \(2018\)](#) shows that explicitly modeling alignment (hard attention) between source and target characters outperforms soft attention. [Wu and Cotterell \(2019\)](#) further show that enforcing monotonicity in a hard attention model improves performance.

5 Conclusion

Using a large batch size and feature invariant input allows the transformer to achieve strong performance on character-level tasks. However, it is unclear what linguistic errors the transformer makes compared to recurrent models on these tasks. Future work should analyze the errors in detail as [Gorman et al. \(2019\)](#) does for recurrent models. While [Wu and Cotterell](#) shows that the monotonicity bias benefits character-level tasks, it is not evident how to enforce monotonicity on multi-headed self-attention. Future work should consider how to best incorporate monotonicity into the model., either by enforcing strictly ([Wu and Cotterell, 2019](#)) or by pretraining the model to copy ([Anastasopoulos and Neubig, 2019](#)).

References

- Roei 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.
- Antonios Anastasopoulos and Graham Neubig. 2019. [Pushing the limits of low-resource morphological inflection](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. [Training data augmentation for low-resource morphological inflection](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, Vancouver. Association for Computational Linguistics.
- Marcel Bollmann. 2018. *Normalization of historical texts with neural network models*. Ph.D. thesis, Bochum, Ruhr-Universität Bochum.
- Marcel Bollmann. 2019. [A large-scale comparison of historical text normalization systems](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. [The CoNLL-SIGMORPHON 2018 shared task: Universal morphological reinflection](#). In *Proceedings of the CoNLL-SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. [CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. [The SIGMORPHON 2016 shared Task—Morphological reinflection](#). In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.
- Simon Flachs, Marcel Bollmann, and Anders Søgaard. 2019. [Historical text normalization with delayed rewards](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1614–1619, Florence, Italy. Association for Computational Linguistics.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR.
- Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. [Weird inflects but OK: Making sense of morphological generation errors](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- 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 (KONVENS 2016)*, pages 146–155.
- 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 (KONVENS 2016)*, pages 146–155.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. [Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 49–57, Vancouver. Association for Computational Linguistics.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. [The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.
- Eva Pettersson. 2016. *Spelling normalisation and linguistic analysis of historical text for information extraction*. Ph.D. thesis, Acta Universitatis Upsalien-sis.
- Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODAL-IDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, 087, pages 54–69. Linköping University Electronic Press.
- Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.
- 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*.
- Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based smt. In *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*.
- Yves Scherrer and Tomaž Erjavec. 2016. Modernising historical Slovene words. *Natural Language Engineering*, 22(6):881–905.
- Terrence J. Sejnowski and Charles R. Rosenberg. 1987. Parallel networks that learn to pronounce English text. *Complex Systems*, 1.
- Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. [Data augmentation for morphological reinflection](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

- Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018a. [An evaluation of neural machine translation models on historical spelling normalization](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018b. [Why self-attention? a targeted evaluation of neural machine translation architectures](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- R.L. Weide. 1998. [The Carnegie Mellon pronouncing dictionary](#).
- Shijie Wu and Ryan Cotterell. 2019. [Exact hard monotonic attention for character-level transduction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.
- Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. [Hard non-monotonic attention for character-level transduction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4425–4438, Brussels, Belgium. Association for Computational Linguistics.
- Min Zhang, Haizhou Li, Rafael E. Banchs, and A Kumar. 2015. [Whitepaper of NEWS 2015 shared task on machine transliteration](#). In *Proceedings of the Fifth Named Entity Workshop*, pages 1–9, Beijing, China. Association for Computational Linguistics.