Exploring the Importance of Source Text in Automatic Post-Editing for Context-Aware Machine Translation

Chaojun Wang[1] **Christian Hardmeier**[2,3] **Rico Sennrich**[4,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computer Science, IT University of Copenhagen
[3]Department of Linguistics and Philology, Uppsala University
[4]Department of Computational Linguistics, University of Zurich
zippo_wang@foxmail.com, chrha@itu.dk, sennrich@cl.uzh.ch

Abstract

Accurate translation requires document-level information, which is ignored by sentence-level machine translation. Recent work has demonstrated that document-level consistency can be improved with automatic post-editing (APE) using only target-language (TL) information. We study an extended APE model that additionally integrates source context. A human evaluation of fluency and adequacy in English–Russian translation reveals that the model with access to source context significantly outperforms monolingual APE in terms of adequacy, an effect largely ignored by automatic evaluation metrics. Our results show that TL-only modelling increases fluency without improving adequacy, demonstrating the need for conditioning on source text for automatic post-editing. They also highlight blind spots in automatic methods for targeted evaluation and demonstrate the need for human assessment to evaluate document-level translation quality reliably.

1 Introduction

Neural machine translation (NMT) has significantly improved the state of the art in MT (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) on the sentence level. However, accurate translation requires looking at larger units than individual sentences (Hardmeier, 2014), and context-aware NMT has recently become a popular research direction (Miculicich et al., 2018; Scherrer et al., 2019; Junczys-Dowmunt, 2019).

One approach to discourse-level processing in NMT is automatic post-editing of the output of a sentence-level system. DocRepair (Voita et al., 2019a) is a monolingual sequence-to-sequence model to correct inconsistencies in groups of adjacent sentence-level translations, showing improvements for specific discourse-level phenomena such as the generation of inflections in elliptic sentences.

The hypotheses explored in this work are (1) that the coherence of the translation can be further improved by exploiting context in the source language, and (2) that the omission of source context disproportionately affects adequacy in a way that is not measured adequately by the existing automatic evaluation procedures.

Our post-editing model is a document-level adaptation of Transference (Pal et al., 2019), a successful three-way transformer architecture from the WMT 2019 Automatic Post-Editing (APE) task (Chatterjee et al., 2019). To keep the model from over-correcting the hypothesis, we use data weighting (Junczys-Dowmunt, 2018) and a conservativeness penalty (Junczys-Dowmunt and Grundkiewicz, 2016). We evaluate on the same training and evaluation sets as Voita et al. (2019a), including a general test set validated by BLEU score and contrastive sets for several discourse phenomena.

Our experimental results confirm both hypotheses. Despite similar BLEU, human evaluation demonstrates that our Transference model significantly outperforms DocRepair in terms of adequacy, whilst both models show a comparable improvement in fluency over a baseline without APE. The automatic evaluation on discourse-specific test sets suggests that source-side information is particularly useful for predicting omitted verb phrases; however, even the targeted discourse-specific evaluation does not reflect the adequacy gain found by human evaluators. This is especially true since some of the discourse-specific test sets of Voita et al. (2019a) have a very narrow focus on problems for which source context is unlikely to help.

2 Transference

Transference (Pal et al., 2019) (Figure 1) is a multi-source transformer (Vaswani et al., 2017) architec-

ture which exploits both source *src* and the MT output *mt* to predict the reference *ref*. It is composed of (1) a source encoder (enc_{src}) to generate the *src* representation, (2) a second encoder ($enc_{src \to mt}$) which is a standard transformer decoder architecture without mask to produce the representation of *mt* incorporating *src* information, and (3) a decoder (dec_{ref}) which captures the final representation from $enc_{src \to mt}$ via cross-attention.

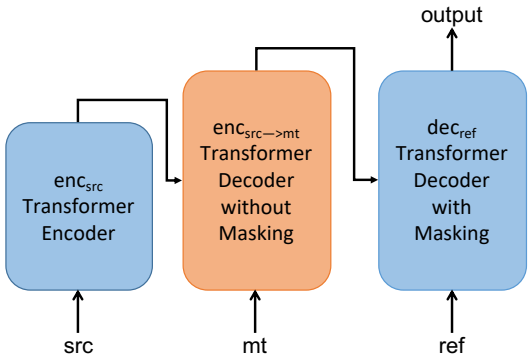

Figure 1: Transference architecture for multi-source document-level repair model.

If document-level APE is trained on a small subset of the parallel data, or only synthetic data, and therefore presumably weaker as a general model of translation than the sentence-level main model, we need to control how aggressively APE can modify *mt* to prevent over-correction. We adopt two strategies from the APE literature to achieve this. A *conservativeness penalty* (Junczys-Dowmunt and Grundkiewicz, 2016), denoted c, penalises the score of each prediction that is not in *src* or *mt*. Formally, let $V_c = V_{src} \cup V_{mt}$ be the subset of the full vocabulary V that occurs in an input segment. Given a $|V|$-sized vector of candidates h_t at time step t, the score of each candidate v is defined as:

$$h_t(v) = \begin{cases} h_t(v) - c & \text{if } v \in V \backslash V_c \\ h_t(v) & \text{otherwise.} \end{cases} \quad (1)$$

Second, similar to Lopes et al. (2019), we apply a *data weighting strategy* during training. We assign each training sample a weight that is defined as $\text{BLEU}_{\text{smooth}}(mt, ref)$ (Lin and Och, 2004) to upweight samples that require little post-editing.

3 Data and Preprocessing

We use all of the English-to-Russian data released by Voita et al. (2019a)[1], including: (1) 6M context-agnostic and 1.5M context-aware (4 consecutive sentences in each sample) data from the OpenSubtitles2018 corpus (Lison et al., 2018); (2) Russian monolingual data in 30M groups of 4 consecutive sentences gathered by Voita et al. (2019a). We reuse the synthetic training data for APE generated by Voita et al. (2019a), treating Russian monolingual data as *ref*, a sentence-level English back-translation as *src*, and the Russian roundtrip translation as *mt*. The evaluation data consists of general test sets extracted from the training data and four contrastive test sets to evaluate specific contextual phenomena.

[1] https://github.com/lena-voita/good-translation-wrong-in-context

Model	Deixis	Lex.c.	Ell.infl.	Ell.VP	BLEU
Results reported by Voita et al. (2019a):					
Baseline	50.0	45.9	53.0	28.4	32.41
DocRepair	91.8	80.6	86.4	75.2	34.60
Our experiments:					
DocRepair	88.6	70.5	83.8	69.0	32.69
DocRepair (+P)	87.6	67.6	82.2	71.8	32.38
Transference	86.8	62.9	81.6	73.0	30.56
Transference (+P)	87.8	65.4	84.8	82.8	32.53

Experiments marked +P use the ParData corpus.

Table 1: BLEU score on general test set and accuracy on contrastive test sets (deixis, lexical consistency, ellipsis (inflection), and VP ellipsis).

The four contrastive test sets have a narrow focus on specific discourse-level phenomena. The "Deixis" set targets consistent use of formal and informal second-person pronouns (T-V distinction) in Russian (however without regard to the social acceptability of the selected form). "Lexical cohesion" targets the consistent transliteration of proper names into Cyrillic script. These two sets are independent of source context by design, as the model is only evaluated on the generation of consistent repetitions of a form it has committed to, regardless of its adequacy in the context. The "Ellipsis VP" set targets elliptic verb phrases, where Russian requires the production of a lexical verb form not found in English. The "Ellipsis inflection" set tests the generation of noun inflections in sentences where the governing verb has been elided.

The training data is tokenised and truecased with Moses (Koehn et al., 2007), and encoded using byte-pair encoding (Sennrich et al., 2016b) with source and target vocabularies of 32000 tokens. Like Voita et al. (2019a), we report lowercased, tokenised BLEU (Papineni et al., 2002) with *multi-bleu.perl* from the Moses toolkit.

4 Model

The sentence-level baselines (EN→RU) and model used for RU→EN back-translation are Transformer base models (Vaswani et al., 2017).

For document-level APE, DocRepair is a Transformer base model that operates on groups of adjacent sentences, mapping from *mt* to *ref*. We use the Nematus toolkit (Sennrich et al., 2017) for DocRepair and our implementation of the Transference architecture, using the same configuration as Pal et al. (2019).[2] Detailed hyperparameters are listed in Appendix A. We train our document-level models on the 30M pairs of synthetic data. For some models, we also include the subset of the parallel data (1.5M pairs) for which context sentences are available, referred to as *ParData*. The *mt* part of *ParData* is generated by randomly sampling 20 translations with our EN→RU baseline system.

In preliminary experiments, adding noise to the training data improved model generalisation. We generated noise with two strategies. Following Voita et al. (2019a), *mt* in both synthetic data and *ParData* is randomly selected from 20 translations, and noise is added by making random token substitutions with probability of 10%. Following Edunov et al. (2018), noise is added to the *src* in synthetic data by three operations: (1) replacing a token; (2) deleting a token; (3) swapping adjacent token pairs, with a probability of 10%.

5 Automatic evaluation

Table 1 shows the results in terms of accuracy on the contrastive test sets and BLEU on the general test set. For DocRepair, we were unable to replicate the exact results of Voita et al. (2019a). Our conclusions are based on our own implementation.

On the general test set, trained on only synthetic training data, Transference achieves about 2 BLEU points less than DocRepair. We suspect that this derives from the mismatch of the training and test data for Transference. Specifically, during training, the "source" seen by Transference is the result of noisy back-translation from Russian, whereas at test time, the source is an original English sentence. When *ParData* is included, Transference and DocRepair achieve comparable BLEU.

In accuracy on the test sets for T/V pronouns ("deixis") and transliteration consistency ("lexical cohesion"), Transference does not improve over DocRepair, which is unsurprising considering how those test sets are constructed. However, adding source knowledge does improve results on both ellipsis test sets, for VP ellipsis even without adding the *ParData* data. The improvement is generally greater for VP ellipsis than for noun inflection.

6 Human evaluation

To gain a better picture of the merits of the different systems, we conducted a manual evaluation. We randomly selected 720 sentences from the general test set and 100 sentences from the discourse test set and had them evaluated separately for adequacy and fluency by two native speakers of Russian. To avoid priming between the fluency and adequacy conditions, the test set was split between the annotators, and no sentence was annotated for adequacy and fluency by the same annotator. To determine the inter-annotator agreement, there are 100 overlapping sentences for two annotators. Table 5 shows inter-annotator agreement results while Table 4 shows the intra-annotator agreement. According to Landis and Koch (1977), all groups of human evaluation results are fair ($\kappa > 0.2$).

The sentences were presented to the annotators in random order along with 3 sentences of preceding context. The sentence to be evaluated was highlighted, and the Russian translations of the three systems (Baseline, DocRepair (+*ParData*) and Transference (+*ParData*)) were displayed next to each other, ordered randomly. In the adequacy condition only, the English source text was also shown. The annotators received instructions according to Table 2 and were told to assign the same rank if two translations were of equal quality. Once the annotation was complete, the rankings were converted into pairwise comparisons. Duplicate assessments from the inter- and intra-annotator sets were counted once if their annotations agreed, and discarded if they disagreed.

Table 3 shows the outcome of pairwise comparisons between the systems, including the number of times the output of one system was preferred over that of the other by the annotator. The results were tested for significance with a sign test. We find the same pattern of results for both test sets. In the *Fluency* evaluation, both monolingual DocRepair and bilingual Transference significantly improve over the Baseline. The comparison between DocRepair and Transference is not significant in this condi-

[2]Code available at `https://github.com/zippotju/Context-Aware-Bilingual-Repair-for-Neural-Machine-Translation`

Adequacy: Please rank the three translations according to *how adequately the translation of the last sentence reflects the meaning of the source, given the context.*

Fluency: Please rank the three translations according to *how fluent the last sentence is, in terms of grammaticality, naturalness and consistency, taking into account the context of the previous sentences.*

Table 2: Instructions to human annotators

		Preference			
System A	*System B*	*A*	*B*	*Ties*	
Fluency					
General corpus:					
Baseline	**DocRepair**	30 <	62	612	($p < 0.005$)
Baseline	**Transference**	51 <	89	547	($p < 0.005$)
DocRepair	Transference	70	78	542	(n. s.)
Discourse corpus:					
Baseline	**DocRepair**	12 <	28	138	($p < 0.05$)
Baseline	**Transference**	15 <	34	120	($p < 0.01$)
DocRepair	Transference	23	25	121	(n. s.)
Adequacy					
General corpus:					
Baseline	DocRepair	24	31	655	(n. s.)
Baseline	**Transference**	34 <	67	592	($p < 0.005$)
DocRepair	**Transference**	39 <	66	592	($p < 0.05$)
Discourse corpus:					
Baseline	DocRepair	16	20	140	(n. s.)
Baseline	**Transference**	9 <	46	117	($p < 0.001$)
DocRepair	**Transference**	11 <	43	117	($p < 0.001$)

n. s. = not significant
Significance threshold: $p < 0.05$

Table 3: Human evaluation results. Winning systems in pairwise comparisons marked in bold.

tion. In the *Adequacy* evaluation, the comparison between DocRepair and the Baseline is not significant, but Transference significantly outperforms both DocRepair and the Baseline, demonstrating that knowledge of the source is essential for APE to improve the accuracy of the translations.

One of the evaluators provided qualitative comments on 32 pairs of DocRepair and Transference outputs sampled from those sentences for which the two systems were ranked differently in the human evaluation. The comments show that both systems tend to produce imperfect output for the same sentences, but the winning system often manages to fix errors partially. Both systems make a wide range of errors in terms of morphology and lexical choice, but the source information permits Transference to correct certain recurring problems more reliably, such as agreement errors, mistranslations of proper names (e.g., Lena as Sarah), or the incorrect use or omission of subjunctive mood in conditional sentences.

Per annotator:		
Annotator 1		91.1%
Annotator 2		83.9%
Per dataset:		
Fluency	General	90.0%
Fluency	Discourse	86.7%
Adequacy	General	90.0%
Adequacy	Discourse	78.3%

Table 4: Intra-annotator agreement of human evaluation

		κ	Pct.
Fluency	General	0.234	5
Fluency	Discourse	0.352	55
Adequacy	General	0.301	27
Adequacy	Discourse	0.471	93

Table 5: Inter-annotator agreement in terms of Cohen's κ (Cohen, 1960). The last column shows the percentile of our κ value in the context of a series of similar evaluations carried out at WMT 2012–2016 (Bojar et al., 2016, Table 4).

7 Related Work

Our work draws on two strands of research: automatic post-editing and context-aware MT.

Automatic post-editing has a long history in MT (Knight and Chander, 1994), with regular shared tasks (Bojar et al., 2015, 2016, 2017). Neural multi-source APE systems as first proposed by Pal et al. (2016) and Junczys-Dowmunt and Grundkiewicz (2016), some of them including source language information (Junczys-Dowmunt and Grundkiewicz, 2017; Chatterjee et al., 2017; Libovický and Helcl, 2017), have come to dominate APE. We take inspiration from the top-performing systems at the WMT19 shared task for architectures and training/decoding tricks (Chatterjee et al., 2019), and make heavy use of synthetic training data (Sennrich et al., 2016a; Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019).

Neural context-aware MT can be achieved by integrating context into the main translation model (Jean et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018, inter alia). Two-stage models with a sentence-level first pass and document-level second pass have been explored for scenarios with asymmetric training data. Voita et al. (2019b) introduces a two-pass model where, unlike in APE, the second-pass is tightly integrated with the first-pass model, reusing its hidden representations. Apart

from Voita et al. (2019a), the model closest to ours is by Junczys-Dowmunt (2019), who explored document-level APE, but only manually evaluated its efficacy as part of a large model ensemble.

8 Conclusion

Our human evaluation shows that monolingual APE oriented towards consistency beyond the sentence level improves fluency, but not adequacy, while multi-source APE with source context improves both adequacy and fluency. This shortcoming of monolingual APE in terms of adequacy was not easily visible with a consistency-focused automatic evaluation, highlighting the need for human evaluation to avoid such blind spots and reinforcing earlier findings about the inadequacy of automatic evaluation methods for discourse-level MT (Guillou and Hardmeier, 2018).

Clearly, a two-stage process with sentence-level translation and multi-sentence APE is a viable approach in asymmetric data settings with little document-level parallel data. However, we still required some actual document-level parallel data, and were unable to match the success of monolingual repair when using only synthetic data. Exploring the data requirements of document-level APE, and devising ways to reduce them, are worth further study.

Acknowledgments

Chaojun Wang was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MTStretch). Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930. This project has received funding from the European Union's Horizon 2020 research and innovation programme (ELITR, grant agreement no 825460), and the Royal Society (NAF\R1\180122).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638, Copenhagen, Denmark. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. Ape at scale and its implications on mt evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*, volume 15 of *Studia Linguistica Upsaliensia*. Acta Universitatis Upsaliensis, Uppsala.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.

Marcin Junczys-Dowmunt. 2018. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 425–430, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 120–129, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI*, volume 94, pages 779–784.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany. Association for Computational Linguistics.

Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2019. USAAR-DFKI – the transference architecture for English–German automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 124–131, Florence, Italy. Association for Computational Linguistics.

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

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks.

In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DISCOMT'17, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameter Search and Validation Performance

The following hyperparameters were manually tuned:

- The percentage of *ParData* mixed with the synthetic training data. of Transference.
- The conservativeness penalty.
- The decision whether to add the conservativeness penalty to the probability estimates or to the logits of the model.

The tuning bounds are shown in Table 7 in curly braces for each tuned hyperparameter. After 18 hyperparameter search trials, the best-performing models were selected considering both BLEU score on the general validation set and the accuracy on the contrastive validation sets. The validation results are shown in Table 6, and the hyperparameter configurations in Table 7.

Model	**Deixis**	**Lex.c.**	**CE.loss**	**BLEU**
DocRepair	89.0	68.0	58.2	32.01
DocRepair (+ParData)	88.8	68.8	56.3	31.63
Transference	86.0	62.2	61.0	30.37
Transference (+ParData)	85.4	64.8	50.7	31.99

Table 6: Validation performance of tested systems (CE represents Cross Entropy).

A.2 Training Time and Model Size

The two sentence-level baselines and the DocRepair model have approximately 72 million parameters each. The baseline systems are trained for around 72 hours each on a GeForce GTX 1080 Ti GPU. DocRepair and DocRepair (+*ParData*) are trained for approximately 216 hours on four TITAN X (Pascal) GPUs and 192 hours on a GeForce RTX 2080 Ti GPU, respectively.

The Transference model has around 119 million parameters. Transference and Transference (+*ParData*) were trained for around 192 and 288 hours, respectively, on three GeForce GTX 1080 Ti GPUs.

	DocRepair	Transference	Tuning bounds
Common hyperparameters			
Embedding layer size	512		
Hidden state size	512		
Tied encoder/decoder embeddings	yes	no	
Tie decoder embeddings	yes		
Loss function	per-token cross-entropy		
Label smoothing	0.1		
Optimizer	Adam		
Learning schedule	Transformer		
Warmup steps	8000		
Gradient clipping threshold	1.0		
Maximum sequence length	500		
Token batch size	15000		
Length normalization alpha	0.6		
Encoder depth	6		
Decoder depth	6		
Feed forward num hidden	2048		
Number of attention heads	8		
Embedding dropout	0.1		
Residual dropout	0.1		
ReLU dropout	0.1		
Attention weights dropout	0.1		
Beam size	4		
Percentage of ParData in training	0.3		$\{0.2, 0.3, 0.4\}$
Transference-specific hyperparameters			
Tied second encoder/decoder embeddings	yes		
Second encoder depth	6		
Conservativeness penalty	(0.2, probability)		$\{0.1, 0.2, 0.3\} \times \{\text{probability}, \text{logit}\}$

Table 7: Hyperparameter configurations for best-performing DocRepair and Transference models, and hyperparameter tuning bounds.