Revisiting Context Choices for Context-aware Machine Translation

Matīss Rikters[λτ], **Toshiaki Nakazawa**[τ]
[λ] Artificial Intelligence Research Center,
National Institute of Advanced Industrial Science and Technology
{matiss.rikters}@aist.go.jp
[τ] The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{nakazawa}@logos.t.u-tokyo.ac.jp

Abstract

One of the most popular methods for context-aware machine translation (MT) is to use separate encoders for the source sentence and context as multiple sources for one target sentence. Recent work has cast doubt on whether these models actually learn useful signals from the context or are improvements in automatic evaluation metrics just a side effect. We show that multi-source transformer models improve MT over standard transformer-base models even with empty lines provided as context, but the translation quality improves significantly (1.51 - 2.65 BLEU) when a sufficient amount of correct context is provided. We also show that even though randomly shuffling in-domain context can also improve over baselines, the correct context further improves translation quality and random out-of-domain context further degrades it.

Keywords: Machine Translation, Document-level, Evaluation

1. Introduction

There are two main approaches for incorporating context in neural machine translation (NMT) along with several others, which have not been adapted as widely (Lopes et al., 2020). The most common methods are: 1) data concatenation without changing the model architecture (Tiedemann and Scherrer, 2017), which can even be extended to full documents on both source and target (Junczys-Dowmunt, 2019); 2) training models with multiple separate encoders for main sentences and context (Zhang et al., 2018); and 3) other methods, such as cache-based (Tu et al., 2018), hierarchical attention (Miculicich et al., 2018; Maruf et al., 2019). The first approach faces the challenge of encoding longer than usual inputs and separating where context ends and content begins. In this work, we focus on the multi-encoder approach and aim to explore three main research questions: 1) is there an optimal amount of previous context sentences; 2) how is translation quality affected by training models using random in-domain context sentences as opposed to random out-of-domain context sentences; and 3) how will translation quality change when using empty lines as context.

Recently there have been several studies (Kim et al., 2019; Li et al., 2020; Jwalapuram et al., 2020) that attribute the success of context-aware NMT to regularisation and noise generated by the additional encoders rather than the actual context. While we are confident that the model alone plays a substantial role in this, we do believe that the correct data matters and even speculate that the second encoder may act as a sort of domain adaptation mechanism on the context data.

We wish to direct our research towards the Japanese-English language pair, as it is very common in the Japanese language to omit pronouns and obvious arguments to verbs which can be inferred from context.

2. Related Work

Multi-source and other multi-encoder models have been widely used in several language processing tasks, such as automatic post-editing (Junczys-Dowmunt and Grundkiewicz, 2018; Shin and Lee, 2018), speech recognition (Zhou et al., 2020), multilingual MT (Zoph and Knight, 2016) and multi-modal MT (Yao and Wan, 2020). There are also many studies on using multi-encoder models for context-aware MT (Jean et al., 2017; Zhang et al., 2018) with various degrees of success.

Most related work focuses on either only 1-2 previous/next sentences as context (Tiedemann and Scherrer, 2017; Voita et al., 2018) or training on full documents (Junczys-Dowmunt, 2019; Macé and Servan, 2019) with nothing in between. From our preliminary investigation of the English corpus (OntoNotes 5.0) and findings of the previous work Hangyo et al. (2014), more than 20% of the antecedents of anaphoras appear more than two sentences before the current sentence where the anaphoras appear. The detailed distribution of the position of the antecedents is shown in Figure 1. This indicates that the existing models which consider only 1-2 previous sentences are not sufficient. In this paper, we explore variations of using

LREC-COLING 2024, pages 14073–14079
20-25 May, 2024.

0-4 previous sentences as context.

Li et al. (2020) experimented with training models using a fixed sentence or randomly sampled words from the vocabulary as context and compared the results to using the actual previous sentence as context. They found that the model still improves over the baseline transformer even with incorrect context and in some cases even outperforms models trained with correct context.

Kim et al. (2019) explore how removing specific parts of the context impacts multi-encoder MT performance. They find that actual utilisation of document-level context is rarely interpretable, but filtering out stop-words and most frequent words from the context or keeping only named entities or specific parts of speech (POS) does not strongly impact translation quality. They also experiment with using longer context of up to 20 sentences and show that performance drops with more than 1-2 sentence context when using full sentences, but is more stable and even improves when retaining only specific POS in the context.

Stojanovski and Fraser (2020) train multi-domain models for translating from English into German using the concatenation approach with 1, 5 and 10 context sentences and separate embeddings to specify the domain of each sentence. They find that the best results are from either the 5 or 10 context sentence model, depending on the domain of the evaluation data. They also perform an ablation experiment by providing context from a different domain at evaluation time for a model trained on the correct context.

We hypothesise that it is actually not the random or fixed sentences provided as context that improves multi-encoder MT output over the baseline, but rather the larger model architecture. To prove this, we extend these experiments by training models with empty lines as context and show that that alone is enough to outperform the baseline transformer model.

We also believe that using random tokens from the same training corpus as context improves the final translation by essentially performing as domain adaptation. To verify this claim, we train separate models using 1) randomly sampled sentences from the same corpus and 2) randomly sampled sentences from a completely different corpus as context. We find clear differences between the two randoms as well as a difference between the best random and best actual context model.

3. Multi-source Transformer Model

There are several different ways to implement multi-source encoder models for MT like concatenating outputs from multiple encoders (Pal et al., 2018) or averaging them. For our experiments, we follow the approach that Junczys-Dowmunt and Grundkiewicz (2018) used for automatic post-editing, where the original transformer (Vaswani et al., 2017) is supplemented by a second encoder and a second multi-head attention block is stacked above the previous multi-head attention block.

We consider two main configurations for training our models, which differ only by the data that is provided to the second encoder as context. The first is *n-context*, where *n* is 0-4 specifying the maximum number[1] of previous context sentences for each content sentence. For the second configuration, we chose to use an *n* of 3 (due to good performance in the first configuration and too few antecedents being further as shown in Figure 1) and train models with random in-domain context (*3-random-ind*) and random out-of-domain context (*3-random-ood*).

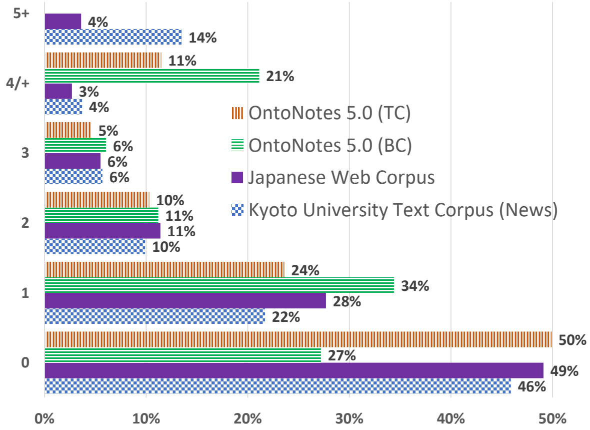

Figure 1: Distribution of the position of the antecedent from the current sentence in Japanese and English. 0 means the antecedent of an anaphora is in the same sentence, 1 means the previous sentence and so on. For Japanese corpora (Kyoto University Text Corpus and Japanese Web Corpus), the antecedents of the omitted element (zero-anaphora) are investigated. For English corpora (OntoNotes 5.0 Broadcast Conversation and Telephone Conversations), the antecedent of all coreference relations are investigated.

4. Experiments

We experiment with training EN↔JA models with several slightly differing configurations. We used the document-aligned Japanese-English conversation corpus (Rikters et al., 2020) as training data, which contains about 220k parallel sentences from about 3k documents. We used the development and evaluation data from the corpus, each containing about 2k sentences from 69 documents, for

[1]Note that even if *n* is 4, the first sentence of each document will always have 0 context, the second will have 1 and so on.

development and evaluation of our models.
For pre-processing we used only Sentencepiece (Kudo and Richardson, 2018) to create a shared vocabulary 16k tokens. We did not perform other tokenisation or truecasing for the training data. We used Mecab (Kudo, 2006) to tokenise the Japanese side of the evaluation data, which we used only for scoring. The English side remained as-is. The parameter count was about 53M for the baseline transformers and 78M for the multi-source models. We use Marian (Junczys-Dowmunt et al., 2018) to train transformer-base models as baselines and seven different configurations of multi-source transformer models using up to 4 previous sentences as context, up to 3 random sentences from the same training data as context, and up to 3 random sentences from JParacrawl (Morishita et al., 2020) (a different, unrelated corpus made up of web-crawled texts) as context. In all experiments with more than one context sentence, the context sentences were provided in a single file to the additional encoder divided by the tabulation symbol. Each model was trained using three random seeds (347155, 42, 9457) on two TITAN Xp GPUs until convergence (loss not improving for 7 checkpoints) with training time of about one day per model. We use the SacreBLEU[2] tool (Post, 2018) to evaluate automatic translations and calculate BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and ChrF (Popović, 2015) scores.
Experiment results are summarised in Table 1 and the most distinctive results of BLEU scores are visualised in Figure 2. We apply paired bootstrap resampling (Koehn, 2004) to calculate significance intervals of BLEU and NIST scores. Here we see that all results significantly outperform the baseline models, even the *0-context* and random context ones. However, if we consider models with *0-context* as our true baseline, then models with out-of-domain random context are within the margin of error while all JA→EN models with in-domain context score significantly higher than that. For EN→JA , there is a slight overlap of 0.06 BLEU in the error intervals between the *0-context* model and the highest scoring model which used 2 context sentences, but according to the NIST score, there is a significant difference. We further verify the significance of this difference in the human evaluation section.
The results also show that there are differences in automatic evaluation scores between all models trained using 1-4 or random 3 sentence actual context, but they are within the margin of error of each other. Nevertheless, we can see that in both translation directions using 3 actual context sentences is slightly better than 3 random in-domain context sentences and using random in-domain context is better than using random out-of-domain context. We also found that for the given training/development/evaluation data combination the best result for EN→JA is achieved by using 2 context sentences, but for JA→EN - by using 3.
For reference, we also trained baseline models on WMT20[3] data (∼13M parallel sentences; WMT20 rows in Table 1) and a mix of all data (WMT20+ rows). While these do outperform baselines trained only on the document-aligned data, the difference in automatic evaluation results is not too outstanding.

Setting	BLEU	NIST	ChrF2
		JA→EN	
baseline	12.35 ± 0.77	3.48 ± 0.11	40.88
0-context	15.31 ± 0.85	4.02 ± 0.14	44.74
1-context	17.29 ± 0.87	4.39 ± 0.14	47.12
2-context	17.34 ± 0.88	4.43 ± 0.14	47.33
3-context	17.96 ± 1.02	4.51 ± 0.14	47.73
4-context	17.14 ± 0.92	4.36 ± 0.14	46.83
3-rand-ind	17.65 ± 0.91	4.52 ± 0.13	47.67
3-rand-ood	16.56 ± 0.92	4.28 ± 0.14	46.45
WMT20	16.29	4.33	45.54
WMT20+	18.44	4.81	48.12
		EN→JA	
baseline	11.86 ± 0.71	3.87 ± 0.10	29.68
0-context	14.00 ± 0.76	4.17 ± 0.10	32.21
1-context	14.93 ± 0.78	4.36 ± 0.10	33.44
2-context	15.51 ± 0.81	4.45 ± 0.10	34.30
3-context	15.26 ± 0.82	4.42 ± 0.11	33.92
4-context	15.19 ± 0.79	4.36 ± 0.11	33.68
3-rand-ind	15.18 ± 0.78	4.36 ± 0.10	33.86
3-rand-ood	14.43 ± 0.80	4.26 ± 0.10	32.85
WMT20	12.99	3.98	31.09
WMT20+	15.33	4.40	33.97

Table 1: Automatic evaluation results.

5. Human Evaluation

We perform human evaluation to compare the 0-context baseline and our highest scoring models (*3-context* for JA→EN and *2-context* for EN→JA). Following the pairwise evaluation method from the WAT workshop (Nakazawa et al., 2019), we randomly sample 400 sentences from each translation direction and employ 5 evaluators to perform a blind comparative evaluation task by specifying if the translation is better or worse than the baseline (-1 or 1) or are they equal (0). Note that the evaluators had access to the context sentences so they could consider the context for the evaluation, however, they had no access to the system names. The final decision for a sentence is determined as a win if the sum of evaluations S $\geq$ 2, a loss if S

[2]Version string: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.21

[3]http://www.statmt.org/wmt20/translation-task.html

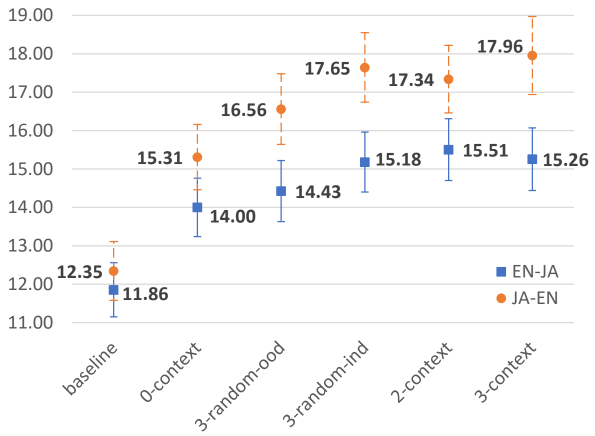

Figure 2: Best EN↔JA results compared to the baseline and 0 context, random out-of-domain (ood) context, and random in-domain (ind) context.

	JA→EN	EN→JA
Wins	131	107
Losses	48	61
Ties	221	232
Score	20.75 ± 3.67	11.50 ± 3.50
Agreement	67.65	64.90
Kappa	0.47	0.51

Table 2: Human evaluation results comparing wins, losses and ties for the *2-context* EN→JA model and *3-context* JA→EN model against the *0-context* models.

$\leq$ -2, or a tie otherwise. We calculate a pairwise score in a range of -100 to 100 as follows:

$$Pairwise = 100 \times \frac{W - L}{W + L + T},$$

where a negative value favours the *0-context* baseline and a positive value - the *2/3-context* model.
Table 2 shows that in both directions models with actual context significantly outperform models with empty lines as context, even for EN→JA where the difference in BLEU scores was not significant.
We also calculated the Free-Marginal Kappa (Randolph, 2005) values for the evaluations to measure inter-annotator agreement between evaluators. The results (EN→JA overall agreement - 64.90%, Free-marginal kappa - 0.47; JA→EN overall agreement - 67.65%, Free-marginal kappa - 0.51) show intermediate to good agreement.

6. Conclusion

In this paper, we explored how the data that is provided as context in the second source encoder of multi-source transformer models impacts the final translation quality. Firstly, we found that using only one previous sentence as context is not the optimal choice - two or three seem to be better, but this obviously depends on the data used, languages in question and translation direction.
Another interesting finding is that the multi-source transformer model significantly outperformed the baseline transformer without any additional data at all. Our intuition is that this is due to the larger model architecture which sees the second empty source as noise and therefore learns clearer distinctions in the actual training data.
Lastly, we have shown that not all random data provided as previous context to multi-source transformer models has equal effect. Using in-domain random context led to 0.75 to 1.09 more BLEU than using out-of-domain random context, and both versions of random context were still slightly worse (0.08 - 0.31 BLEU) than the same corresponding models that used the correct context. This encourages us to perhaps focus more on considering similar-domain comparable data for context-aware modelling in future work as opposed to directly parallel data, which is often more expensive and more difficult to acquire.

7. Future Work

For future work, we plan to perform similar experiments on different less explored language pairs, which is challenging due to the requirement of a decent amount of document-aligned data, preferably with document boundaries. We would also be interested in probing the trained models and exploring what was learned by training on empty context lines. Another interesting avenue to explore would be to verify if other context sentences from the same paragraph or even the same document are more beneficial than other random sentences from elsewhere in the same corpus as opposed to other random sentences from other unrelated corpora.

Acknowledgements

This work was supported by “Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation”, the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN.

Limitations

In this work, we only considered training our models on data that is confirmed to be document-level and has reliable alignments. Also, since hyperparameter tuning on training large models is computationally very costly, we opt for choosing mostly default parameters in our experiments. While it would be interesting to compare the dual-encoder model with a single-encoder model of similar parameter count, it is difficult to choose which part of the model to upscale, since the parameter count difference is 1.5x. Ablation experiments of varying layer counts, hidden, feed-forward sizes, etc. would be required.

Ethics Statement

Our work fully complies with the ACL Code of Ethics[4]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. All human data annotators were fairly compensated in accordance with market rates.

8. Bibliographical References

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2014. Building and analyzing a diverse document leads corpus annotated with semantic relations. *Journal of Natural Language Processing*, 21(2):213–247.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context?

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. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev,

[4] https://www.aclweb.org/portal/content/acl-code-ethics

André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2020. Can your context-aware mt system pass the dip benchmark tests? : Evaluation benchmarks for discourse phenomena in machine translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. arXiv.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Valentin Macé and Christophe Servan. 2019. Using Whole Document Context in Neural Machine

Translation. In *16th International Workshop on Spoken Language Translation 2019*. Zenodo.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. 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 3092–3102, Minneapolis, Minnesota. 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.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

Santanu Pal, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2018. A transformer-based multi-source automatic post-editing system. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 827–835, Belgium, Brussels. 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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the Joensuu Learning and Instruction Symposium*, volume 2005.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2020. Document-aligned japanese-english conversation parallel corpus. In *Proceedings of the Fifth Conference on Machine Translation*, pages 637–643, Online. Association for Computational Linguistics.

Jaehun Shin and Jong-Hyeok Lee. 2018. Multi-encoder transformer network for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 840–845, Belgium, Brussels. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2020. Addressing zero-resource domains using document-level context in neural machine translation.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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*, volume 30, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Xinyuan Zhou, Emre Yılmaz, Yanhua Long, Yijie Li, and Haizhou Li. 2020. Multi-Encoder-Decoder Transformer for Code-Switching Speech Recognition. In *Proc. Interspeech 2020*, pages 1042–1046.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

9. Language Resource References