Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns

KayYen Wong[†] **Sameen Maruf**[‡] **Gholamreza Haffari**[‡]
Faculty of Information Technology, Monash University, VIC, Australia
[†]kywon63@student.monash.edu
[‡]firstname.lastname@monash.edu

Abstract

The advent of context-aware NMT has resulted in promising improvements in the overall translation quality and specifically in the translation of discourse phenomena such as pronouns. Previous works have mainly focused on the use of past sentences as context with a focus on anaphora translation. In this work, we investigate the effect of future sentences as context by comparing the performance of a contextual NMT model trained with the future context to the one trained with the past context. Our experiments and evaluation, using generic and pronoun-focused automatic metrics, show that the use of future context not only achieves significant improvements over the context-agnostic Transformer, but also demonstrates comparable and in some cases improved performance over its counterpart trained on past context. We also perform an evaluation on a targeted cataphora test suite and report significant gains over the context-agnostic Transformer in terms of BLEU.

1 Introduction

Standard machine translation (MT) systems typically translate sentences in isolation, ignoring essential contextual information, where a word in a sentence may reference other ideas or expressions within a piece of text. This locality assumption hinders the accurate translation of referential pronouns, which rely on surrounding contextual information to resolve cross-sentence references. The issue is further exacerbated by differences in pronoun rules between source and target languages, often resulting in morphological disagreement in the quantity and gender of the subject being referred to (Vanmassenhove et al., 2018).

Rapid improvements in NMT have led to it replacing SMT as the dominant paradigm. With this, context-dependent NMT has gained traction, overcoming the locality assumption in SMT through the use of additional contextual information. This has led to improvements in not only the overall translation quality but also pronoun translation (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018; Miculicich et al., 2018). However, all these works have neglected the context from *future* sentences, with Voita et al. (2018) reporting it to have a negative effect on the overall translation quality.

In this work, we investigate the effect of future context in improving NMT performance. We particularly focus on pronouns and analyse corpora from different domains to discern if the future context could actually aid in their resolution. We find that for the Subtitles domain roughly 16% of the pronouns are cataphoric. This finding motivates us to investigate the performance of a context-dependent NMT model (Miculicich et al., 2018) trained on the future context in comparison to its counterpart trained on the past context. We evaluate our models in terms of overall translation quality (BLEU) and also employ three types of automatic pronoun-targeted evaluation metrics. We demonstrate strong improvements for all metrics, with the model using future context showing comparable or in some cases even better performance than the one using only past context. We also extract a targeted cataphora test set and report significant gains on it with the future context model over the baseline.

2 Related Work

Pronoun-focused SMT Early work in the translation of pronouns in SMT attempted to exploit coreference links as additional context to improve the translation of anaphoric pronouns (Le Nagard and Koehn 2010; Hardmeier and Federico 2010). These works yielded mixed results which were attributed to the limitations of the coreference resolution systems used in the process (Guillou, 2012).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5971–5978
July 5 - 10, 2020.

Domain	#Sentences	Document length
	English-German	
Subtitles	9.39M/9K/14.1K	565.8/582.2/591.0
Europarl	1.67M/3.6K/5.1K	14.1/15.0/14.1
TED Talks	0.21M/9K/2.3K	120.9/96.4/98.7
	English-Portuguese	
Subtitles	15.2M/16.1K/23.6K	580.4/620.6/605.3

Table 1: Train/dev/test statistics: number of sentences (K: thousands, M: millions), and average document length (in sentences). The *#Documents* can be obtained by dividing the *#Sentences* by the *Document Length*.

Context-Aware NMT Multiple works have successfully demonstrated the advantages of using larger context in NMT, where the context comprises few previous source sentences (Wang et al., 2017; Zhang et al., 2018), few previous source and target sentences (Miculicich et al., 2018), or both past and future source and target sentences (Maruf and Haffari, 2018; Maruf et al., 2018, 2019).

Further, context-aware NMT has demonstrated improvements in pronoun translation using past context, through concatenating source sentences (Tiedemann and Scherrer, 2017) or through an additional context encoder (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018). Miculicich et al. (2018) observed reasonable improvements in generic and pronoun-focused translation using three previous sentences as context. Voita et al. (2018) observed improvements using the previous sentence as context, but report decreased BLEU when using the following sentence. We, on the other hand, observe significant gains in BLEU when using the following sentence as context on the same data domain.

3 Contextual Analysis of Corpora

To motivate our use of the future context for improving the translation of cataphoric pronouns in particular and NMT in general, we first analyse the distribution of coreferences for anaphoric and cataphoric pronouns over three different corpora - OpenSubtitles2018[1] (Lison and Tiedemann, 2016), Europarl (Koehn, 2005) and TED Talks (Cettolo et al., 2012) - for English-German. For Europarl and TED Talks, we use the publicly available document-aligned version of the corpora (Maruf et al., 2019). For Subtitles, we align the English and German subtitles at the document-level using publicly available alignment links.[2] To control for the length and coherency of documents, we keep subtitles with a run-time less than 50 minutes (for English) and those with number of sentences in the hundreds. The corpus is then randomly split into training, development and test sets in the ratio 100:1:1.5. Table 1 presents the corpora statistics.

[1] http://www.opensubtitles.org/

[2] http://opus.nlpl.eu/OpenSubtitles2018.php

Pronoun	Subtitles	Europarl	TED Talks
Intrasentential	30.1	75.6	64.1
Anaphora (< 0)	54.3	19.6	28.5
Cataphora (> 0)	15.6	4.7	7.4

Table 2: Percentage of different pronoun types.

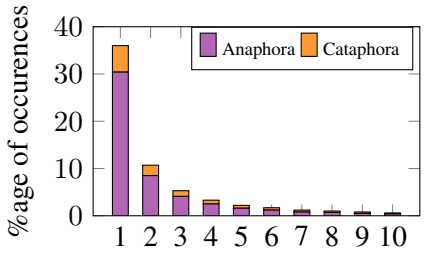


Figure 1: Plot showing proportion of intersentential English pronouns versus size of coreference resolution window for the Subtitles corpus (plots for Europarl and TED Talks are in the appendix).

Analysis of Coreferences We find the smallest window within which a referential English pronoun is resolved by an antecedent or postcedent using `NeuralCoref`.[3] Table 2 shows that the majority of pronouns in Europarl and TED Talks corpora are resolved intrasententially, while the Subtitles corpus demonstrates a greater proportion of intersentential coreferences. Further, anaphoric pronouns are much more frequent compared to cataphoric ones across all three corpora. For Subtitles, we also note that a good number of pronouns (15.6%) are cataphoric, ∼37% of which are resolved within the following sentence (Figure 1). This finding motivates us to investigate the performance of a context-aware NMT model (trained on Subtitles) for the translation of cataphoric pronouns.

4 Experiments

Datasets We experiment with the Subtitles corpus on English-German and English-Portuguese language-pairs. To obtain English-Portuguese data, we employ the same pre-processing steps as reported in §3 (corpus statistics are in Table 1). We use 80% of the training data to train our models and the rest is held-out for further evaluation as discussed later in § 4.2.[4] The data is truecased using

[3] https://github.com/huggingface/neuralcoref

[4] Due to resource contraints, we use about two-thirds of the final training set (∼8M sentence-pairs) for En-Pt.

Lang. Pair	Baseline	HAN(k = +1)	HAN(k = -1)
English→German	31.87	**32.53**	32.48
German→English	35.92	**36.64**♠	36.32
English→Portuguese	35.45	36.04	**36.21**
Portuguese→English	39.34	**39.96**♠	39.69

Table 3: BLEU for the Transformer baseline and Transformer-HAN with following sentence (k = +1) and previous sentence (k = -1). ♠: Statistically significantly better than HAN (k = -1).

the Moses toolkit (Koehn et al., 2007) and split into subword units using a joint BPE model with 30K merge operations (Sennrich et al., 2016).[5]

Description of the NMT systems As our baseline, we use the `DyNet` (Neubig et al., 2017) implementation of Transformer (Vaswani et al., 2017).[6] For the context-dependent NMT model, we choose the Transformer-HAN encoder (Miculicich et al., 2018), which has demonstrated reasonable performance for anaphoric pronoun translation on Subtitles. We extend its `DyNet` implementation (Maruf et al., 2019) to a single context sentence.[7][8] For training, Transformer-HAN is initialised with the baseline Transformer and then the parameters of the whole network are optimised in a second stage as in Miculicich et al. (2018) (details of model configuration are in Appendix A.1). For evaluation, we compute BLEU (Papineni et al., 2002) on tokenised truecased text and measure statistical significance with $p < 0.005$ (Clark et al., 2011).

4.1 Results

We consider two versions of the Transformer-HAN respectively trained with the following and previous source sentence as context. From Table 3, we note both context-dependent models to significantly outperform the Transformer across all language-pairs in terms of BLEU. Further, HAN (k = +1) demonstrates statistically significant improvements over the HAN (k = -1) when translating to English. These results are quite surprising as Voita et al. (2018) report decreased translation quality in terms of BLEU when using the following sentence for English→Russian Subtitles. To identify if this discrepancy is due to the language-pair or the model, we conduct experiments with English→Russian in the same data setting as Voita et al. (2018) and find that HAN (k = +1) still significantly outperforms the Transformer and is comparable to HAN (k = -1) (more details in Appendix A.2).

Model	APT	Precision	Recall	F1-score	CRC
English→German					
Baseline	60.8	47.4	54.3	50.7	-
+HAN(k = +1)	61.4	**48.3**	54.3	**51.1**	-
+HAN(k = -1)	**62.0**	48.0	**54.6**	**51.1**	-
German→English					
Baseline	77.9	56.9	50.4	53.4	50.4
+HAN(k = +1)	**78.3**	57.9	**50.6**	**54.0**	50.9
+HAN(k = -1)	**78.3**	**58.0**	50.5	**54.0**	**51.0**
English→Portuguese					
Baseline	46.4	54.8	**56.0**	55.4	-
+HAN(k = +1)	47.0	55.8	55.2	55.5	-
+HAN(k = -1)	**47.3**	**56.0**	55.4	**55.7**	-
Portuguese→English					
Baseline	64.3	54.9	51.1	53.0	50.2
+HAN(k = +1)	**64.6**	**55.7**	**51.5**	**53.5**	50.9
+HAN(k = -1)	64.3	55.6	51.2	53.4	**51.6**

Table 4: Pronoun-focused evaluation on generic test set for models trained on different types of context.

4.2 Analysis

Pronoun-Focused Automatic Evaluation For the models in Table 3, we employ three types of pronoun-focused automatic evaluation:

1. **Accuracy of Pronoun Translation (APT)** (Miculicich Werlen and Popescu-Belis, 2017)[9]. This measures the degree of overlapping pronouns between the output and reference translations obtained via word-alignments.
2. **Precision, Recall and F1 scores**. We use a variation of AutoPRF (Hardmeier and Federico, 2010) to calculate precision, recall and F1-scores. For each source pronoun, we compute the clipped count (Papineni et al., 2002) of overlap between candidate and reference translations. To eliminate word alignment errors, we compare this overlap over the set of dictionary-matched target pronouns, in contrast to the set of target words aligned to a given source pronoun as done by AutoPRF and APT.
3. **Common Reference Context (CRC)** (Jwalapuram et al., 2019). In addition to the previous

[5] Tokenisation is provided by the original corpus.

[6] https://github.com/duyvuleo/Transformer-DyNet

[7] Where in the original architecture, k sentence-context vectors were summarised into a document-context vector, we omit this step when using only one sentence in context.

[8] The code and data are available at https://github.com/sameenmaruf/acl2020-contextnmt-cataphora.

[9] https://github.com/idiap/APT

Model	BLEU	APT	F1-score
Baseline (Transformer)	31.87	61.6	49.1
+HAN($k = \emptyset$)	32.30	61.6	49.1
+HAN(k = +1,+2)	32.56♠	**62.0**	49.8
+HAN(k = -2,-1)	32.47	61.9	49.8
+HAN(k = -2,-1,+1,+2)	**32.59**♠	**62.0**	**49.9**

Table 5: Evaluation on English→German generic test set for HAN trained with $k = \{$-2,-1,+1,+2$\}$ but decoded with varying context. ♠: Statistically significantly better than HAN with no context ($k = \emptyset$).

two measures which rely on computing pronoun overlap between the target and reference translation, we employ an ELMo-based (Peters et al., 2018) evaluation framework that distinguishes between a good and a bad translation via pairwise ranking (Jwalapuram et al., 2019). We use the CRC setting of this metric which considers the same reference context (one previous and one next sentence) for both reference and system translations. However, this measure is limited to evaluation only on the English target-side.[10]

The results using the aforementioned pronoun evaluation metrics are reported in Table 4. We observe improvements for all metrics with both HAN models in comparison to the baseline. Further, we observe that the HAN (k = +1) is either comparable to or outperforms HAN (k = -1) on APT and F1 for De→En and Pt→En, suggesting that for these cases, the use of following sentence as context is at least as beneficial as using the previous sentence. For En→De, we note comparable performance for the HAN variants in terms of F1, while for En→Pt, the past context appears to be more beneficial.[11] In terms of CRC, we note HAN (k = -1) to be comparable to (De→En) or better than HAN (k = +1) (Pt→En). We attribute this to the way the metric is trained to disambiguate pronoun translations based on only the previous context and thus may have a bias for such scenarios.

Ablation Study We would like to investigate whether a context-aware NMT model trained on a wider context could perform well even if we do not have access to the same amount of context at decoding. We thus perform an ablation study for English→German using the HAN model trained with two previous and next sentences as context and decoded with variant degrees of context.

From Table 5, we note that reducing the amount of context at decoding time does not have adverse effect on the model's performance. However, when no context is used, there is a statistically significant drop in BLEU, while APT and F1-scores are equivalent to that of the baseline. This suggests that the model does rely on the context to achieve the improvement in pronoun translation. Further, we find that the future context is just as beneficial as the past context in improving general translation performance.

	English→German			
Model	**Cataphora**	**DET**	**PROPN**	**NOUN**
Baseline	32.33	32.14	33.02	32.93
+HAN(k = +1)	32.93♠	32.68♠	33.98♠	33.76♠
	German→English			
Model	**Cataphora**	**DET**	**PROPN**	**NOUN**
Baseline	36.91	36.35	38.81	38.84
+HAN(k = +1)	37.68♠	37.19♠	39.51	39.45
	English→Portuguese			
Model	**Cataphora**	**DET**	**PROPN**	**NOUN**
Baseline	36.29	35.91	37.91	37.60
+HAN(k = +1)	37.08♠	36.70♠	38.49	38.19
	Portuguese→English			
Model	**Cataphora**	**DET**	**PROPN**	**NOUN**
Baseline	40.74	40.12	42.77	42.63
+HAN(k = +1)	41.63♠	41.06♠	43.60♠	43.42♠

Table 6: BLEU on the cataphora test set and its subsets for the Transformer and Transformer-HAN (k = +1). ♠: Statistically significantly better than the baseline.

Cataphora-Focused Test Suite To gauge if the improvements in Table 3 for the HAN (k = +1) model are coming from the correct translation of cataphoric pronouns, we perform an evaluation on a cataphoric pronoun test suite constructed from the held-out set mentioned earlier in § 3. To this end, we apply `NeuralCoref` over the English side to extract sentence-pairs which have a cataphoric pronoun in one sentence and the postcedent in the next sentence. This is further segmented into subsets based on the part-of-speech of the postcedent, that is, determiner (DET), proper noun (PROPN) or all nouns (NOUN) (more details in the appendix).[12]

From Table 6, we observe HAN (k = +1) to outperform the baseline for all language-pairs when evaluated on the cataphora test suite. In general, we observe greater improvements in BLEU when trans-

[10] We use the same English pronoun list for all pronoun-focused metrics (provided by Jwalapuram et al. (2019) at https://github.com/ntunlp/eval-anaphora). All pronoun sets used in our evaluation are provided in Appendix A.4.

[11] It should be noted that for Portuguese, adjectives and even verb forms can be marked by the gender of the noun and these are hard to account for in automatic pronoun-focused evaluations.

[12] We note that there may be some overlap between the three pronoun subsets as a test sentence may contain more than one type of pronoun.

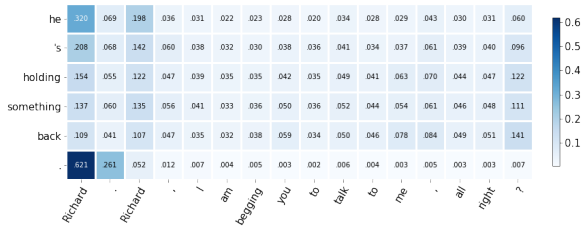


Figure 2: Example attention map between source (*y*-axis) and context (*x*-axis). The source pronoun *he* correctly attends to the postcedents *Richard* in the context.

lating to English, which we attribute to the simplification of cross-lingual pronoun rules when translating from German or Portuguese to English.[13] We also observe fairly similar gains in BLEU across the different pronoun subsets, which we hypothesise to be due to potential overlap in test sentences between different subsets. Nevertheless, we note optimum translation quality over the noun subsets (PROPN and NOUN), while seeing the greatest percentage improvement on the DET subset. For the latter, we surmise that the model is able to more easily link pronouns in a sentence to subjects prefixed with possessive determiners, for example, "his son" or "their child".

We also perform an auxiliary evaluation for Transformer-HAN (k = -1) trained with the previous sentence as context on the cataphora test suite and find that the BLEU improvements still hold. Thus, we conclude that Transformer-HAN (a context-aware NMT model) is able to make better use of coreference information to improve translation of pronouns (detailed results in Appendix A.3).

Qualitative Analysis We analyse the distribution of attention to the context sentence for a few test cases.[14] Figure 2 shows an example in which a source pronoun *he* attends to its corresponding postcedent in context. This is consistent with our hypothesis that the HAN (k = +1) is capable of exploiting contextual information for the resolution of cataphoric pronouns.

5 Conclusions

In this paper, we have investigated the use of future context for NMT and particularly for pronoun translation. While previous works have focused on the use of past context, we demonstrate through rigorous experiments that using future context does not deteriorate translation performance over a baseline. Further, it shows comparable and in some cases better performance as compared to using the previous sentence in terms of both generic and pronoun-focused evaluation. In future work, we plan to investigate translation of other discourse phenomena that may benefit from the use of future context.

[13] It should be noted that the cataphora test set is extracted based on the existence of cataphoric-pairs in the English-side, which may have biased the evaluation when English was in the target.

[14] Attention is average of the per-head attention weights.

Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments and feedback and to George Foster for fruitful discussions. This work is supported by a Google Faculty Research Award to G.H. It is further supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (www.massive.org.au).

References

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, Louisiana, USA. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Short Papers)*, pages 176–181. Association for Computational Linguistics.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine

translation benefit from larger context? *CoRR*, abs/1704.05135.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. 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 2957–2966, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86. AAMT.

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 ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 923–929, Portorož, Slovenia. European Language Resources Association.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics.

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, USA. 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.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

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.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

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 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.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. 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.

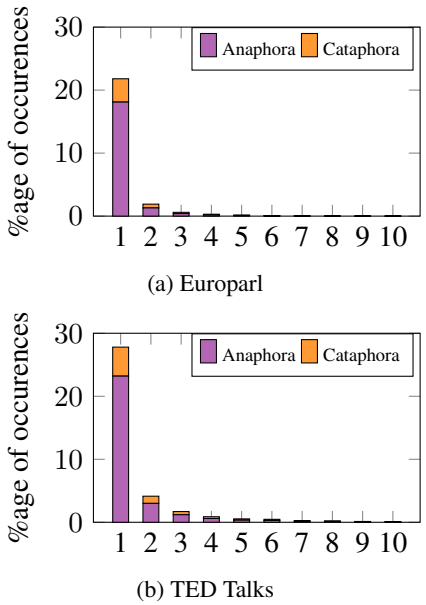
A Experiments

A.1 Model Configuration

We use similar configuration as the Transformer-base model (Vaswani et al., 2017) except that we reduce the number of layers in the encoder and decoder stack to 4 following Maruf et al. (2019). For training, we use the default Adam optimiser (Kingma and Ba, 2015) with an initial learning rate of 0.0001 and employ early stopping.

A.2 English→Russian Experiments

We wanted to compare the two variants of Transformer-HAN with k = +1 and k = -1 in the same experimental setting as done by Voita et al. (2018). The data they made available only contains the previous context sentence. Thus, we extract training, development and test sets following the procedure in this work but of roughly the same size as Voita et al. (2018) for a fair comparison of the two context settings. While they extract their test set as a random sample of sentences, we extract



(a) Europarl

(b) TED Talks

Figure 3: Plots showing proportion of intersentential English pronouns versus size of coreference resolution window for Europarl and TED Talks corpora.

from a random sample of documents, resulting in a test set which has document-level continuity between sentences. The pre-processing pipeline is the same as the one used for our English-German and English-Portuguese experiments except that we perform lowercasing (instead of truecasing) and learn separate BPE codes for source and target languages following Voita et al. (2018). We also evaluate the models trained with our training set on the test set provided by Voita et al. (2018) after removing the sentences overlapping with our train and dev sets (corpora statistics are in Table 7).

Results Table 8 indicates that the model trained with the next sentence as context not only statistically significantly outperforms the Transformer baseline (+0.9 BLEU) but also demonstrates comparable performance to the HAN model trained

Origin	#Sentences	Document length
Voita et al. (2018)	2M/10K/10K	-
Ours	2M/11K/10K	606.3/620.6/631.6
Our, Voita et al. (2018)⋆	2M/11K/7.3K	606.3/620.6/-

Table 7: Train/dev/test statistics for English-Russian: number of sentences (K: thousands, M: millions), and average document length (in sentences). The first row mentions statistics of data used by Voita et al. (2018), the second row mentions statistics of data we extracted, and the third row mentions the data statistics for our train/dev sets and Voita et al. (2018)'s test after removing overlap (referred as Voita et al. (2018)⋆).

Data Setting	Baseline	HAN(k = +1)	HAN(k = -1)
Ours	23.35	**24.25**	24.18
Our, Voita et al. (2018)*	27.15	-	**28.23**

Table 8: BLEU on tokenised lowercased text for the Transformer baseline and Transformer-HAN with following sentence (k = +1) and previous sentence (k = -1) for English→Russian. All reported results for the HAN variants are statistically significantly better than the baseline.

with the previous sentence. This finding is consistent with our main results. We also evaluate the model trained with our training data on Voita et al. (2018)* test set and report almost four points jump in the absolute BLEU score for both the baseline and the context-dependent model.[15] In addition, we note that for their test set, the HAN (k = -1) has greater percentage improvement over the baseline (4%) than what they report for their context-aware model (2.3%).

A.3 Cataphora-Focused Test Suite

We segment the cataphora test set into the following subsets based on the part-of-speech of the postcedent being referred to:

- **DET** Postcedents prefixed with possessive determiners, e.g., *his son* or *their child*.
- **PROPN** Postcedents which are proper nouns, i.e., named entities.
- **NOUN** Postcedents which are nouns, including proper nouns and common nouns, such as *boy* or *child*.

A.3.1 Results for HAN (k = -1)

We evaluate Transformer-HAN (k = -1) enriched with anaphoric context on the cataphora test set (Table 9) to determine if this context-aware model is making use of coreference information to improve the overall translation quality (in BLEU). We find that HAN (k = +1) performs better than HAN (k = -1) when English is in the target-side, which we hypothesise to be because of the extraction of the cataphora test suite from the English-side. However, when English is in the source-side, both models perform comparably showing that the Tranformer-HAN (a context-aware NMT model) is able to make better use of coreference information to improve translation of pronouns.

[15]The BLEU score for the baseline on Voita et al. (2018)* is less than the one reported in their original work because of the reduced size of the test set and the different training set.

Lang. Pair	Baseline	HAN(k = -1)
English→German	32.33	32.94
German→English	36.91	37.23
English→Portuguese	36.29	37.24
Portuguese→English	40.74	41.25

Table 9: BLEU on the cataphora test set for the Transformer and Transformer-HAN (k = -1). All results for HAN (k = -1) are statistically significantly better than the baseline.

A.4 Pronoun Sets

Language	Pronouns
English	i, me, my, mine, myself, we, us, our, ours, ourselves, you, your, yours, yourself, yourselves, he, his, him, himself, she, her, hers, herself, it, its, itself, they, them, their, themselves, that, this, these, those, what, whatever, which, whichever, who, whoever, whom, whose
German	ich, du, er, sie, es, wir, mich, dich, sich, ihn, uns, euch, mir, dir, ihm, ihr, ihre, ihrer, ihnen, meiner, mein, meine, deiner, dein, seiner, sein, seine, unser, unsere, euer, euere, denen, dessen, deren, meinen, meinem, deinen, deinem, deines, unserer, unseren, unseres, unserem, ihrem, ihres, seinen, seinem, seines
Portuguese	eu, nós, tu, você, vocês, ele, ela, eles, elas, me, te, nos, vos, o, lo, no, a, la, na, lhe, se, os, los, as, las, nas, lhes, mim, ti, si, meu, teu, seu, nosso,vosso, minha, tua, sua, nossa, vossa, meus teus, seus, nossos, vossos, minhas, tuas, suas nossas, vossas, dele, dela, deles, delas, quem que, qual, quais, cujo, cujos, cuja, cujas, onde

Table 10: Pronoun sets used in our pronoun-focused automatic evaluation.