The University of Edinburgh-Uppsala University's Submission to the WMT 2020 Chat Translation Task

Nikita Moghe, Christian Hardmeier, Rachel Bawden

The University of Edinburgh-Uppsala University's Submission to the WMT 2020 Chat Translation Task

Nikita Moghe[1] **Christian Hardmeier**[2] **Rachel Bawden**[1]
[1]School of Informatics, University of Edinburgh, Scotland
[2]Department of Linguistics and Philology, Uppsala University
{nikita.moghe,rachel.bawden}@ed.ac.uk
christian.hardmeier@lingfil.uu.se

Abstract

This paper describes the joint submission of the University of Edinburgh and Uppsala University to the WMT'20 chat translation task for both language directions (English↔German). We use existing state-of-the-art machine translation models trained on news data and fine-tune them on in-domain and pseudo-in-domain web crawled data. We also experiment with (i) adaptation using speaker and domain tags and (ii) using different types and amounts of preceding context. We observe that contrarily to expectations, exploiting context degrades the results (and on analysis the data is not highly contextual). However using domain tags does improve scores according to the automatic evaluation. Our final primary systems use domain tags and are ensembles of 4 models, with noisy channel reranking of outputs. Our en-de system was ranked second in the shared task while our de-en system outperformed all the other systems.[1]

1 Introduction and challenges

The task's aim is to create machine translation (MT) systems to enable task-oriented communication between a service agent and a customer speaking different languages (English and German respectively). Like most dialogues, the texts can show strong context sensitivities, as the customer and the agent engage in a common activity and continually react to each other's utterances (Hardmeier, 2014; Bawden, 2018). However, the dialogues, which relate to ordering or reserving products and services from a limited set of providers, also follow fairly strong scripts and are anchored in a small discourse universe defined by the products on offer. Their context sensitivity is therefore counterbalanced by domain-specific conventions and expectations.

Our design choices are informed by an initial manual inspection of the training data and a baseline translation, which revealed that the main challenges relate to idiomaticity: incorrect or poor translation of English idioms, named entities and politeness markers (e.g. formal vs. informal forms of address, or poor translation of English *sir*) and an incorrect use of domain-specific terminology. Almost always, the problems were the result of an excessively literal translation of the source text, and this literalness also frequently affected the reference translations themselves too. Surprisingly, we found few instances of phenomena explicitly requiring context to be correctly translated (e.g. we did not find pronominal anaphora to be a major problem in the dialogues examined).[2] The context-dependent instances we did find were more task-specific (e.g. English *Enjoy!* should be translated differently depending on whether it is about a pizza (*Guten Appetit!*) or a film (*Viel Spaß!*)).

We therefore focus on domain adaptation and general context modelling strategies. Our submissions are based on existing state-of-the-art MT systems for news translation, which we fine-tune on in-domain and pseudo-in-domain data. We also experiment with (i) adapting the models to the different speaker roles and to the different tasks during fine-tuning and (ii) exploiting preceding context through a simple but effective method of concatenating previous sentences to the current one. Our code and models are publicly available.[3]

[1] http://www.statmt.org/wmt20/chat-task_results_DA.html

[2] We tested AllenNLP's coreference resolution tool (Gardner et al., 2018) on a few examples where pronoun resolution seemed relevant and found that it performed very poorly in these cases, confirming similar conclusions by Bawden (2016). We therefore decided not to model coreference explicitly.

[3] http://github.com/chardmeier/WMT2020-Chat

2 Data

The task data consists of parallel task-oriented dialogues between an agent (English) and a customer (German) across six domains: (i) ordering pizza, (ii) making auto repair appointments, (iii) ordering a taxi, (iv) ordering movie tickets, (v) ordering coffee and (vi) making restaurant reservations. The dialogues were initially in English, retrieved from a subset of the TaskMaster-1 dataset (Byrne et al., 2019) and then manually translated into German at Unbabel.[4] Although the speaker tags are provided for each utterance, the conversations are not explicitly marked with their task domain. The task being to translate the agent's utterances from English into German and the customer's utterances from German to English, we evaluate each translation direction separately, using only the agent's utterances for en–de translation and the customer's utterances for de–en. For training however, we use the full set of 13,845 utterances for both directions.

3 Approaches

We explore four approaches, each of which is detailed below: (i) pretraining using additional data sources, (ii) speaker adaptation, (iii) domain adaptation and (iv) incorporating previous context.

Pretraining To account for the limited in-domain data, we use pre-existing MT models trained for the WMT'19 news task (Barrault et al., 2019) and then continue training on pseudo-in-domain web crawled data from the Paracrawl project[5] (Bañón et al., 2020), before fine-tuning on the in-domain chat training data. We compare two different base systems for each language direction: UEDIN models[6] ((Bawden et al., 2019a) and FAIR models (Ng et al., 2019). The pseudo-in-domain data on which training is continued is created by filtering Paracrawl data using dual conditional noisy cross-entropy filtering (Junczys-Dowmunt, 2018). This consists in training a neural language model for each language on the task training data, and jointly scoring each parallel sentence in Paracrawl using the two models. We take the top scoring 2.5 million subset of the original 34 million en–de sentences (those that most resemble the task data).

Speaker adaption Distinguishing between the two speaker roles is important as they have different contributions to the dialogue; the customer's utterances are short, interrogative and informal, while the agent's utterances are often long, informative and more formal. We adapt our models to each speaker by using the speaker identity (provided with the task data) as a pseudo-token (Sennrich et al., 2016a): we prepend a speaker tag to each utterance on both the source and the target side.

Domain adaptation Knowing which task the dialogue belongs to (e.g. pizza, film) can be important for disambiguation, as described in Section 1. Similarly to speaker adaptation, we adapt to the different tasks (i.e. domains) by prepending a domain tag to each utterance on both the source and target side. We also consider a setup where all the utterances are tagged with speaker and domain-tags (see the example in Table 1). The dataset consists of chats across six different domains (pizza, auto, taxi, movie, coffee, and restaurant). As the domains are not indicated in the task dataset, we obtain domain tags by automatically classifying each dialogue as belonging to one of the six tasks using the English side of the data and a baseline German translation.

The dialogue classifier is trained by unsupervised k-means clustering of the training set dialogues with scikit-learn (Pedregosa et al., 2011). As features, we use the nouns in the texts (as recognised by the SpaCy PoS tagger[7]), which works substantially better than using all words. The 6 clusters are initialised to the word sets {*pizza*}, {*auto, car, repair*}, {*ride*}, {*movie*}, {*coffee*}, {*dinner, restaurant*}. Dialogues in the test set are then assigned to the cluster with the nearest centroid. To evaluate the classifier, we manually annotated 49 dialogues from the training set. Training only on the remainder of the training data, we achieved perfect accuracy on the annotated set.

To simulate an online translation scenario, we also experimented with classification using only the initial utterances of each dialogue. In this setting, it was beneficial to project the feature space to a very low dimension using Latent Semantic Analysis (LSA). The best results with a macro-averaged F-score of 0.862 (precision 0.896; recall 0.867) were obtained by using the first 4 sentences and an LSA dimensionality of 5. However, since there was no online constraint in the shared task, we ultimately decided to use the more accurate

[4] https://github.com/Unbabel/BConTrasT

[5] https://www.paracrawl.eu

[6] Although the WMT'19 submission included only de–en, we also use the similarly trained model for en–de.

[7] https://spacy.io

Adaptation	Source text	Target text
Speaker	<speaker=customer> Perfect. Okay, got it.	<speaker=customer> Perfekt. In Ordnung, verstanden.
Domain	<taxi> Perfect. Okay, got it.	<taxi> Perfekt. In Ordnung, verstanden.
Speaker+domain	<taxi> <speaker=customer> Perfect. Okay, got it.	<taxi> <speaker=customer> Perfekt. In Ordnung, verstanden.

Table 1: Examples from the dataset annotated with variants of speaker and domain tags.

full-dialogue classifier for our submission.

Context-level MT Finally, we explore using linguistic context (varying numbers of previous utterances) to improve translation, with the aim that previous context can provide vital information for disambiguation or adaptation. We use the approach of concatenating varying numbers of previous sentences to the current sentence, separated by a sentence boundary token `<break>` (Tiedemann and Scherrer, 2017; Bawden et al., 2018). This simple strategy was shown to be one of the most effective in a recent comparison of document-level MT approaches (Lopes et al., 2020). To distinguish between different speakers, we also add the speaker tag to the beginning of every utterance. The models are trained to translate both the context and the utterance into the target language (i.e. n-to-n strategy). The candidate utterance is then extracted from the generated output in a preprocessing step. Since the dialogues are bilingual (the agent and customer are speaking in different languages), the original versions of the previous sentences can be either in English or in German. While we always translate both the context and the current sentence into the target language on the target side, we consider two approaches to incorporate context in the source sentence: (i) `ORIG`: each previous sentence is in the original language of its speaker (if the context and current sentences are not produced by the same speaker, our input will be a mix of English and German) and (ii) `SAME`: the source context is provided in the same language as the current sentence (language consistency in the source input). At test time, this requires translating utterances sentence by sentence (as opposed to batch decoding); when the previous utterances are not from the same speaker, they must first be translated by the MT model in the opposite language direction for them to be used as context for the current sentence.

4 Experimental setup

We compare two neural MT base system types, both WMT'19 news translation task submissions: UEDIN (University of Edinburgh; Bawden et al. 2019a and FAIR (Facebook; Ng et al. 2019). All models are transformer-big models (Vaswani et al., 2017): 6 encoder and 6 decoder layers, model dimension of 1024, 16 heads except that UEDIN has a feedforward dimension of 4096 for both the encoder and decoder, and FAIR models increase this dimension to 8192 in the encoder. UEDIN models are implemented in Marian (Junczys-Dowmunt et al., 2018) and FAIR models in Fairseq (Ott et al., 2019). Both model types are trained on parallel and backtranslated monolingual data from the WMT'19 news translation shared task (Barrault et al., 2019). For our final submission (using the base FAIR model), we also use noisy channel reranking (Yee et al., 2019), which requires MT models in both directions and a (target) language model. We describe the data processing techniques in Appendix A and list the hyper-parameters in Appendix B.

5 Experimental Results and Analysis

We report automatic evaluation results in Section 5.1 and provide a qualitative manual comparison in Section 5.2.

5.1 Automatic evaluation results

We report BLEU scores (Papineni et al., 2002), calculated with SACREBLEU[8] (Post, 2018) on the dev set (beam size of 4).

Pretraining The results in Table 2 show that in-domain fine-tuning of the pretrained models always gives large gains. The pre-trained FAIR models are better than the pre-trained UEDIN models (Barrault et al., 2019). Fine-tuning on filtered paracrawl and then on the in-domain data gives a slight gain for the UEDIN models (particularly for de–en) but slightly degrades the FAIR models. We choose to take as a base the models fine-tuned on filtered paracrawl to fine-tune all subsequent models (with tags and context). Though these models perform similar to the FT_1 models, as these were trained on more data, they are likely to be more robust on unseen data. Note that all pretrained models

[8]Default parameters and case-sensitive evaluation.

outperform the baseline models trained just on the chat training data (shown in the first row).

Model	en–de UEDIN	en–de FAIR	de–en UEDIN	de–en FAIR
Chat baseline	33.2	35.8	37.4	30.9
Pretrained	42.5	41.0	44.9	48.5
+ in-domain (FT_1)	58.6	61.4	61.0	62.3
+ paracrawl (FT_2)	44.8	45.4	46.5	45.2
+ in-domain	58.8	60.8	60.9	62.2

Table 2: BLEU scores on the dev set for both pretrained models, and of each model fine-tuned on (i) in-domain data and (ii) filtered paracrawl then in-domain data.

Effect of adding tags As shown in Table 3, we observe that in general the performance of both systems improves with the addition of tags. The use of speaker tags improves the BLEU scores for UEDIN models while dialogue tags improve the BLEU scores for FAIR models. We did not observe an improvement in BLEU scores in models using both the tags over models that used a single tag.

Model	en–de UEDIN	en–de FAIR	de–en UEDIN	de–en FAIR
FT_2 + no tag	58.8	60.8	60.9	62.2
FT_2 + speaker	59.4	61.3	60.1	62.1
FT_2 + domain	59.6	**61.5**	60.8	**62.7**
FT_2 + speaker + domain	59.6	61.1	61.4	61.6

Table 3: Dev set BLEU scores for fine-tuning with tags.

Context-level MT As shown in Table 4, the contextual models perform similarly to the baseline for FAIR models while the performance degrades slightly with the UEDIN models. Increasing the number of contextual sentences degrades BLEU scores, most likely due to the necessity to translate longer sentences. It is also likely that the MT systems do not benefit from the addition of previous sentences because the particular chat dataset used contains utterances that do not need context to be correctly translated, contrary to expectations but in line with findings by Mosig et al. (2020). Using context in the same language (SAME) was more beneficial than the original context (ORIG). It is evident that SAME would perform better than ORIG as the pre-trained models were never exposed to such mix-language utterances. Despite fine-tuning a monolingual encoder on mix-language utterances, ORIG systems perform well.

Final submission Table 5 shows the results of our primary submission on both the dev and test

Model	en–de UEDIN	en–de FAIR	de–en UEDIN	de–en FAIR
FT_2 + in-domain	58.8	60.8	60.9	62.2
In-domain data uses previous context (ORIG language)				
FT_2 + 1 prev	58.2	60.3	58.9	**61.8**
FT_2 + 2 prev	56.1	60.2	58.7	61.5
FT_2 + 3 prev	53.3	59.5	56.7	61.7
In-domain data uses previous context (SAME language)				
FT_2 + 1 prev	58.1	61.0	59.2	**62.2**
FT_2 + 2 prev	57.5	60.1	59.1	61.5
FT_2 + 3 prev	55.4	60.5	57.3	62.1

Table 4: Dev set BLEU scores for contextual MT models. The numbers before "prev" are the number of previous utterances used as context.

Model(FAIR)	en–de dev	en–de test	de–en dev	de–en test
FT_2 + domain-tags	61.5	**60.3**	62.7	60.6
+ noisy-channel re-ranking	62.0	60.1	62.9	61.8
+ ensemble [primary]	62.1	60.2	63.1	**62.4**
FT_1 [contrastive]	61.4	60.2	62.3	61.8
FT_2 + 1-same [contrastive]	61.0	59.8	62.2	61.5

Table 5: The method-wise ablation of our final submission: a 4-model ensemble of FAIR based FT_2 models fine-tuned with in-domain training data tagged with domain tags. The outputs are obtained through noisy-channel reranking.

sets: a 4-model ensemble, each model trained by first fine-tuning the pre-existing FAIR model on filtered paracrawl data, then on in-domain training data tagged with dialogue tags and then reranked using noisy channel reranking (n=20) (Yee et al., 2019). We note that noisy channel reranking is more effective for en–de than for de–en. Ensembling provides limited gains. We report our contrastive submissions for comparison. Our models were chosen on their respective performances on the dev set. We observe that the trends for dev set and test set are similar except for FT_2 + domain-tags model without the noisy channel re ranking.

5.2 Qualitative Evaluation

As the gains in BLEU scores with different configurations are limited, it is difficult to identify if the models exhibit qualitative improvement. We created an evaluation set by selecting around 40 peculiar utterances in each translation direction from the development set and conducted an informal human evaluation by assigning scores of -1, 0 or 1 to poor, acceptable or particularly good translations. The average score was used to guide model selection. As per the qualitative evaluation, there

were few and similar errors across different models to draw any significant conclusions. Notably, the number of errors was higher for the en–de direction due to the production of literal translations. Our primary submission achieved a score of 85.357 on human evaluation using direct assessment.

6 Discussion and Future Work

We observe that fine-tuning the WMT'19 news-adapted models on in-domain chat data is a strong baseline. The addition of tags, though helpful, has limited gains on BLEU, and the addition of context (intuitively an important component for any dialogue related task) actually degrades results. We speculate that this is due to the nature of the original dataset, which has limited linguistic diversity and utterances that are mostly context-independent (Mosig et al., 2020). The overall translation of this dataset was of excellent quality, allowing easy understanding of the dialogues. However, the translated chats exhibit translationese and in some cases lacked naturalness, also the case of the references themselves. An interesting avenue for data collection would be a spontaneous generation of chats in two different languages which can roughly follow the same discourse as in (Bawden et al., 2019b).

Acknowledgements

We thank Ulrich Germann for providing us with the pretrained UEDIN models and FAIR for making their models publicly available. Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh (Moghe). The authors gratefully acknowledge Huawei for their support (Moghe). This work was also supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 825299 (GoURMET), 825303 and the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MT-Stretch).

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy.

Rachel Bawden. 2016. Cross-lingual pronoun prediction with linguistically informed features. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 564–570, Berlin, Germany.

Rachel Bawden. 2018. *Going beyond the sentence: Contextual Machine Translation of Dialogue*. Ph.D. thesis, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019a. The university of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy.

Rachel Bawden, Sophie Rosset, Thomas Lavergne, and Éric Bilinski. 2019b. DiaBLa: A Corpus of Bilingual Spontaneous Written Dialogues for Machine Translation. *CoRR*, abs/1905.13354.

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.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018.

AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels.

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

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR'15, San Diego, California, USA.

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.

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.

António V. 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*, Online, formerly Lisbon, Portugal.

Johannes E. M. Mosig, Vladimir Vlasov, and Alan Nichol. 2020. Where is the context? – a critique of recent dialogue datasets. *CoRR*, abs/2004.10473.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling 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 5696–5701, Hong Kong, China.

A Pre- and post-processing

We reuse the data processing of each pre-trained system (reusing subword segmentation models). For UEDIN models, the data is preprocessed using a SentencePiece (Kudo and Richardson, 2018) model with a joint vocabulary of 32k subwords. By default, we use a maximum sentence length of 100 subwords and scale this when adding previous context (e.g. 200 subwords for 1 previous sentence, 300 for 2, etc.). For FAIR models, the Moses toolkit (Koehn et al., 2007) is used for tokenisation and FastBPE [9] for subword segmentation (Sennrich et al., 2016b). A maximum length of 1024 is used for all models.

For FAIR models, we observed some inconsistencies while detokenising the generated outputs in terms of punctuation. We post-processed the output using regular expressions to ensure there was no additional space with the punctuation marks. We also standardised the production of $ in the German output such that all the prices now follow XX,XX $ convention.

B Hyper-parameters

The pretrained models are fine-tuned (first on filtered Paracrawl data, then on the task-specific training data). Adam optimiser (Kingma and Ba, 2015) is used to fine-tune all models, with a batch size of 32 (except for FAIR fine-tuning on filtered Paracrawl data where a batch size of 64 was used). For UEDIN, we use a learning rate of 0.0009, a learning rate warmup of 16000. We validate every 250k subwords decoded. The best model is chosen based on the best BLEU score and least cross-entropy loss on the side of the dev set specific to the language direction for UDEIN and FAIR respectively. For FAIR, we use a learning rate of the last epoch of the pre-trained model (9.85e-5 for en–de, 9.89e-5 for de–en) and validate per epoch.

The training parameters for each model are summarised in Table 6.

Detail\Model	UEDIN	FAIR
Preprocessing	SentencePiece[10]	Moses tokeniser[11]+ FastBPE[12]
Optimiser	Adam	Adam
Learning rate	9e-4 (warmup of 16000)	9.85e-5 (En-De), 9.89e-5 (De-En)
Batch size	32	32 (64 for paracrawl data)
Checkpoint	250k words decoded	1 training epoch
Best model	Best BLEU on dev	Smallest cross-entropy loss on dev

Table 6: Pre-processing and hyper-parameters.

[9] https://github.com/glample/fastBPE

[12] https://github.com/glample/fastBPE

[12] (Kudo and Richardson, 2018), using a joint 32k model.

[12] (Koehn et al., 2007)