CUED@WMT19:EWC&LMs

Felix Stahlberg[†] and **Danielle Saunders**[†] and **Adrià de Gispert**[‡] and **Bill Byrne**[‡†]
[†]Department of Engineering, University of Cambridge, UK
[‡]SDL Research, Cambridge, UK
{fs439, ds636, wjb31}@cam.ac.uk, {agispert, bbyrne}@sdl.com

Abstract

Two techniques provide the fabric of the Cambridge University Engineering Department's (**CUED**) entry to the **WMT19** evaluation campaign: elastic weight consolidation (**EWC**) and different forms of language modelling (**LMs**). We report substantial gains by fine-tuning very strong baselines on former WMT test sets using a combination of checkpoint averaging and EWC. A sentence-level Transformer LM and a document-level LM based on a modified Transformer architecture yield further gains. As in previous years, we also extract n-gram probabilities from SMT lattices which can be seen as a source-conditioned n-gram LM.

1 Introduction

Both fine-tuning and language modelling are techniques widely used for NMT. Fine-tuning is often used to adapt a model to a new domain (Luong and Manning, 2015), while ensembling neural machine translation (NMT) with neural language models (LMs) is an effective way to leverage monolingual data (Gulcehre et al., 2015, 2017; Stahlberg et al., 2018a). Our submission to the WMT19 news shared task relies on ideas from these two lines of research, but applies and combines them in novel ways. Our contributions are:

- Elastic weight consolidation (Kirkpatrick et al., 2017, EWC) is a domain adaptation technique that aims to avoid degradation in performance on the original domain. We report large gains from fine-tuning our models on former English-German WMT test sets with EWC. We find that combining fine-tuning with checkpoint averaging (Junczys-Dowmunt et al., 2016b,a) yields further significant gains. Fine-tuning is less effective for German-English.
- Inspired by the shallow fusion technique by Gulcehre et al. (2015, 2017) we ensemble our neural translation models with neural language models. While this technique is effective for single models, the gains are diminishing under NMT ensembles trained with large amounts of back-translated sentences.
- To incorporate document-level context in a light-weight fashion, we propose a modification to the Transformer (Vaswani et al., 2017) that has separate attention layers for inter- and intra-sentential context. We report large perplexity reductions compared to sentence-level LMs under the new architecture. Our document-level LM yields small BLEU gains on top of strong NMT ensembles, and we hope to benefit even more from it in document-level human evaluation.
- Even though the performance gap between NMT and traditional statistical machine translation (SMT) is growing rapidly on the task at hand, SMT can still improve very strong NMT ensembles. To combine NMT and SMT we follow Stahlberg et al. (2017a, 2018b) and build a specialized n-gram LM for each sentence that computes the risk of hypotheses relative to SMT lattices.
- While data filtering was central in last year's evaluation (Koehn et al., 2018b; Junczys-Dowmunt, 2018b), in our experiments this year we found that a very simple filtering approach based on a small number of crude heuristics can perform as well as dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018a,b).
- We confirm the effectiveness of source-side noise for scaling up back-translation as proposed by Edunov et al. (2018).

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 364–373
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

2 Document-level Language Modelling

MT systems usually translate sentences in isolation. However, there is evidence that humans also take context into account, and judge translations from humans with access to the full document higher than the output of a state-of-the-art sentence-level machine translation system (Läubli et al., 2018). Common examples of ambiguity which can be resolved with cross-sentence context are pronoun agreement or consistency in lexical choice. This year's WMT competition encouraged submissions of translation systems that are sensitive to cross-sentence context. We explored the use of document-level language models to enhance a sentence-level translation system. We argue that this is a particularly light-weight way of incorporating document-level context. First, the LM can be trained independently on monolingual target language documents, i.e. no parallel or source language documents are needed. Second, since our document-level decoder operates on the n-best lists from a sentence-level translation system, existing translation infrastructure does not have to be changed – we just add another (document-level) decoding pass. On a practical note, this means that, by skipping the second decoding pass, our system would work well even for the translation of isolated sentences when no document context is available.

Our document-level LMs are trained on the concatenations of all sentences in target language documents, separated by special sentence boundary tokens. Training a standard Transformer LM (Vaswani et al., 2017) on this data already yields significant reductions in perplexity compared to sentence-level LMs. However, the attention layers have to capture two kinds of dependencies – the long-range cross-sentence context and the short-range context within the sentence. Our modified Intra-Inter Transformer architecture (Fig. 1) splits these two responsibilities into two separate layers using masking. The "Intra-Sentential Attention" layer only allows to attend to the previous tokens in the current sentence, i.e. the intra-sentential attention mask activates the tokens between the most recent sentence boundary marker and the current symbol. The "Inter-Sentential Attention" layer is restricted to the tokens in all previous *complete* sentences, i.e. the mask enables all tokens from the document beginning to the most recent sentence boundary marker. As usual (Vaswani et al., 2017), during training the attention masks are also designed to prevent attending to future tokens. Fig. 2 shows an example of the different masks. Note that as illustrated in Fig. 1, both attention layers are part of the same layer stack which allows a tight integration of both types of context. An implication of this design is that they also use the same positional embedding – the positional encoding for the first unmasked item for intra-sentential attention may not be zero. For example, 'Lonely' has the position 10 in Fig. 2 although it is the first word in the current sentence.

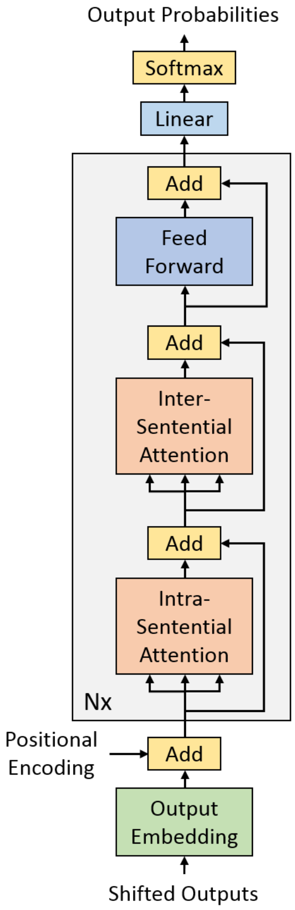

Figure 1: Our modified Intra-Inter Transformer architecture with two separate attention layers.

We use our document-level LMs to rerank n-best lists from a sentence-level translation system. Our initial document is the first-best sentence hypotheses. We greedily replace individual sentences with lower-ranked hypotheses (according to the translation score) to drive up a combination of translation and document LM scores. We start with the sentence with the minimum difference between the first- and second-best translation scores.

	Vinyl	destination	:	who	is	actually	buying	records	?	$</s>$	Lonely	,	middle-aged	men	love	'???'
Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Intra-sentential	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	-
Inter-sentential	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	-

Figure 2: Intra-sentential and inter-sentential attention masks for an English example from `news-test2017`. Document-level context helps to predict the next word ('vinyl').

We stop when the translation score difference to the first-best translation exceeds a threshold.[1]

3 Experimental Setup

Our experimental setup is essentially the same as last year (Stahlberg et al., 2018b): Our preprocessing includes Moses tokenization, punctuation normalization, truecasing, and joint subword segmentation using byte pair encoding (Sennrich et al., 2016c) with 32K merge operations. We compute cased BLEU scores with `mteval-v13a.pl` that are directly comparable with the official WMT scores.[2] Our models are trained with the TensorFlow (Abadi et al., 2016) based Tensor2Tensor (Vaswani et al., 2018) library and decoded with our SGNMT framework (Stahlberg et al., 2017b, 2018c). We delay SGD updates (Saunders et al., 2018) to use larger training batch sizes than our technical infrastructure[3] would normally allow with vanilla SGD by using the `MultistepAdam` optimizer in Tensor2Tensor. We use Transformer (Vaswani et al., 2017) models in two configurations (Tab. 1). Preliminary experiments are carried out with the 'Base' configuration while we use the 'Big' models for our final system. We use `news-test2017` as development set to tune model weights and select checkpoints and `news-test2018` as test set.

	Base	Big
T2T HParams set	`trans._base`	`trans._big`
# physical GPUs	4	4
Batch size	4,192	2,048
SGD delay factor	2	4
# training iterations	300K	1M
Beam size	4	8

Table 1: Transformer setups.

3.1 ParaCrawl Corpus Filtering

Junczys-Dowmunt (2018a,b) reported large gains from filtering the ParaCrawl corpus. This year, the WMT organizers made version 3 of the ParaCrawl corpus available. We compared two different filtering approaches on the new data set. First, we implemented dual cross-entropy filtering (Junczys-Dowmunt, 2018a,b), a sophisticated data selection criterion based on neural language model and neural machine translation model scores in both translation directions. In addition, we used the "naive" filtering heuristics proposed by Stahlberg et al. (2018b):

- Language detection (Nakatani, 2010) in both source and target language.
- No words contain more than 40 characters.
- Sentences must not contain HTML tags.
- The minimum sentence length is 4 words.
- The character ratio between source and target must not exceed 1:3 or 3:1.
- Source and target sentences must be equal after stripping out non-numerical characters.
- Sentences must end with punctuation marks.

Tab. 2 indicates that our systems benefit from ParaCrawl even without filtering (rows 1 vs. 2). Our best 'Base' model uses both dual and naive filtering. However, the difference between filtering techniques diminishes under stronger 'Big' models with back-translation (rows 6 and 7).

4 Results

4.1 Back-translation

Back-translation (Sennrich et al., 2016b) is a well-established technique to use monolingual target language data for NMT. The idea is to automatically generate translations into the source language with an inverse translation model, and add these synthetic sentence pairs to the training data. A major limitation of vanilla back-translation is that the amount of synthetic data

[1] Tensor2Tensor implementation: `https://github.com/fstahlberg/ucam-scripts/blob/master/t2t/t2t_refine_with_glue_lm.py`

[2] `http://matrix.statmt.org/`

[3] The Cambridge HPC service (`http://www.hpc.cam.ac.uk/`) allows parallel training on up to four physical P100 GPUs.

	Model	ParaCrawl	Naive filtering	BLEU			
				test15	test16	test17	test18
1	Base	No		29.3	34.1	27.8	41.9
2	Base	Full		30.0	35.3	28.2	43.1
3	Base	Full	✓	30.3	35.6	28.6	43.5
4	Base	Dual x-ent filtering		30.2	35.5	28.7	43.6
5	Base	Dual x-ent filtering	✓	30.6	35.7	28.8	43.8
6	Big (with back-translation)	Full	✓	32.4	38.5	31.2	46.6
7	Big (with back-translation)	Dual x-ent filtering	✓	32.7	38.1	31.1	46.6

Table 2: Comparison of ParaCrawl filtering techniques. The rest of the training data is over-sampled to roughly match the size of the filtered ParaCrawl corpus. In the 'Dual x-ent filtering' experiments we selected the 15M best sentences according the dual cross-entropy filtering criterion of Junczys-Dowmunt (2018a).

	news-2016 (35M sentences)	news-2017 (20M sentences)	news-2018 (37M sentences)	Noise	BLEU			
					test15	test16	test17	test18
1					30.2	35.7	28.7	43.8
2		✓			30.8	36.2	29.8	44.3
3			✓		30.4	35.8	29.4	43.2
4		✓	✓		30.3	35.9	29.5	43.1
5		✓		✓	31.0	36.6	29.7	44.8
6			✓	✓	30.7	36.6	29.5	44.7
7		✓	✓	✓	30.6	36.6	29.5	44.4
8	✓	✓		✓	31.3	37.4	30.0	45.2
9	✓	✓	✓	✓	31.3	37.3	30.3	45.2

Table 3: Using different corpora for back-translation. We back-translated with a 'base' model for `news-2017` and the big single Transformer model of Stahlberg et al. (2018b) for `news-2016` and `news-2018`.

	Fine-tuning	Checkpoint averaging	BLEU (test18)	
			En-De	De-En
1	No		46.7	46.5
2	No	✓	46.6	46.4
3	Cont'd train.		47.1	46.6
4	Cont'd train.	✓	47.3	46.8
5	EWC		47.1	46.4
6	EWC	✓	47.8	46.8

Table 4: Fine-tuning our models on former WMT test sets using continued training and EWC.

has to be balanced with the amount of real parallel data (Sennrich et al., 2016b,a; Poncelas et al., 2018). Edunov et al. (2018) had overcome this limitation by adding random noise to the synthetic source sentences. Tab. 3 shows that using noise improves the BLEU score by between 0.5 and 1.5 points on the `news-test2018` test set (rows 2-4 vs. 5-7).[4] Our final model uses a very large number (92M) of (noisy) synthetic sentences (row 9), although the same performance could already be reached with fewer sentences (row 8).

4.2 Fine-tuning with EWC and Checkpoint Averaging

Fine-tuning (Luong and Manning, 2015) is a domain adaptation technique that first trains a model until it converges on a training corpus A, and then continues training on a usually much smaller corpus B which is close to the target domain. Similarly to Schamper et al. (2018); Koehn et al. (2018a), we fine-tune our models on former WMT test sets (2008-2016) to adapt them to the target domain of high-quality news translations. Due to the very small size of corpus B, much care has to be taken to avoid over-fitting. We experimented with different techniques that keep the model parameters in the fine-tuning phase close to the original ones. First, we fine-tuned our models for about 1K-2K iterations (depending on the performance on the `news-test2017` dev set) and dumped checkpoints every 500 steps. Averaging all fine-tuning checkpoints together with the last unadapted checkpoint yields minor gains over fine-tuning without averaging (rows 3 vs. 4 in Tab. 4). However, we obtain the best results by combining checkpoint averaging with another regularizer – elastic weight consolidation (Kirkpatrick et al., 2017, EWC) – that explicitly penalizes the distance of the model parameters θ to the optimized but unadapted model parameters θ^*_A. The regularized training objective according EWC is:

$$L(\theta) = L_B(\theta) + \lambda \sum_i F_i(\theta_i - \theta^*_{A,i})^2 \quad (1)$$

[4] We use Sergey Edunov's `addnoise.py` script available at `https://gist.github.com/edunov/d67d09a38e75409b8408ed86489645dd`

Model	Context	Perplexity (per subword)							
		German				English			
		test15	test16	test17	test18	test15	test16	test17	test18
Standard (Big)	Sentence-level	36.23	35.69	36.17	34.77	39.94	37.19	35.34	42.38
Standard(Big)	Document-level	26.63	27.85	25.43	28.36	43.37	34.55	31.27	39.74
Intra-Inter (Big)	Document-level	23.54	22.39	22.05	22.56	34.25	31.16	29.31	34.47

Table 5: Language model perplexities of different neural language models. 'Intra-Inter' denotes our modified Transformer architecture from Sec. 2. The standard model has 448M parameters, Intra-Inter has 549M parameters.

		English-German			German-English		
		Base	Big (with EWC)		Base	Big (with EWC)	
		Single	Single	4-Ensemble	Single	Single	4-Ensemble
1	Using back-translation?	No	Yes	Yes	No	Yes	Yes
2	NMT	43.8	47.8	48.8	40.7	47.4	48.3
3	+ Sentence-level LM	44.7	47.8	48.8	41.4	47.6	48.3
4	+ PBSMT (MBR-based)	45.1	48.0	49.1	42.1	47.6	48.5
5	+ Document-level Intra-Inter LM	45.7	47.6	49.3	42.1	47.3	48.6

Table 6: Using different kinds of language models for translation on `news-test2018`. The PBSMT baseline gets 26.7 BLEU on English-German and 27.5 BLEU on German-English.

where $L_B(\theta)$ is the normal cross-entropy training loss on task B and $F_i = \mathbb{E}\big[\nabla^2 L_A(\theta_i)\big]$ is an estimate of task A Fisher information, which represents the importance of parameter θ_i to A. On English-German, fine-tuning with EWC and checkpoint averaging yields an 1.1 BLEU improvement (rows 1 vs. 6 in Tab. 4). Gains are generally smaller on German-English.

4.3 Language modelling

We introduced our new Intra-Inter Transformer architecture for document-level language modelling in Sec. 2. Tab. 5 shows that our architecture achieves much better perplexity than both a sentence-level language model and a document-level vanilla Transformer model. Tab. 6 summarizes our translation results with various kinds of language models. Adding a Transformer sentence-level LM to NMT helps for the single Base model without back-translation, but is less effective on top of (ensembles of) Big models with back-translation (row 2 vs. 3). Extracting n-gram probabilities from traditional PBSMT lattices as described by Stahlberg et al. (2017a) and using them as source-conditioned n-gram LMs yields gains even on top of our ensembles (row 4). Our document-level Intra-Inter language models improve the ensembles and the single En-De Base model, but hurt performance slightly for the single Big models (row 5).

5 Related Work

Regularized fine-tuning Our approach to fine-tuning is a combination of EWC (Kirkpatrick et al., 2017) and checkpoint averaging (Junczys-Dowmunt et al., 2016b,a). In our context, both methods aim to avoid *catastrophic forgetting*[5] (Goodfellow et al., 2013; French, 1999) and over-fitting by keeping the adapted model close to the original, and can thus be seen as *regularized* fine-tuning techniques. Khayrallah et al. (2018); Dakwale and Monz (2017) regularized the output distributions during fine-tuning using techniques inspired by knowledge distillation (Bucilu et al., 2006; Hinton et al., 2014; Kim and Rush, 2016). Barone et al. (2017) applied standard L2 regularization and a variant of dropout to domain adaptation. EWC as generalization of L2 regularization has been used for NMT domain adaptation by Thompson et al. (2019); Saunders et al. (2019). In particular, Saunders et al. (2019) showed that EWC is not only more effective than L2 in reducing catastrophic forgetting but even yields gains on the general domain when used for fine-tuning on a related domain.

Document-level MT Various techniques have been proposed to provide the translation system with inter-sentential context, for example by initializing encoder or decoder states (Wang et al., 2017a), using multi-source encoders (Bawden et al., 2018; Jean et al., 2017), as additional decoder input (Wang et al., 2017a), with memory-augmented neural networks (Tu et al., 2018; Maruf and Haffari, 2018; Kuang et al., 2017), hierar-

[5] Catastrophic forgetting occurs when the performance on the specific domain is improved after fine-tuning, but the performance of the model on the general domain has decreased drastically.

chical attention (Miculicich et al., 2018; Maruf et al., 2019), deliberation networks (Xiong et al., 2018), or by simply concatenating multiple source and/or target sentences (Tiedemann and Scherrer, 2017; Bawden et al., 2018). Context-aware extensions to Transformer encoders have been proposed by Voita et al. (2018); Zhang et al. (2018). Techniques also differ in whether they use source context only (Jean et al., 2017; Wang et al., 2017a; Voita et al., 2018; Zhang et al., 2018), target context only (Tu et al., 2018; Kuang et al., 2017), or both (Bawden et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Tiedemann and Scherrer, 2017; Maruf et al., 2019). Several studies on document-level NMT indicate that automatic and human sentence-level evaluation metrics often do not correlate well with improvements in discourse level phenomena (Bawden et al., 2018; Läubli et al., 2018; Müller et al., 2018). Our document-level LM approach is similar to the work of Xiong et al. (2018) in that cross-sentence context is only used in a second pass to improve translations from a sentence-level MT system. Our method is light-weight as, similarly to Tiedemann and Scherrer (2017), we do not modify the architecture of the core NMT system.

English-German	
Team	**BLEU**
MSRA	44.9
Microsoft	43.9
NEU	43.5
UCAM	**43.0**
Facebook FAIR	42.7
JHU	42.5
eTranslation	41.9
8 more...	

German-English	
Team	**BLEU**
MSRA	42.8
Facebook FAIR	40.8
NEU	40.5
UCAM	**39.7**
RWTH	39.6
MLLP-UPV	39.3
DFKI	38.8
4 more...	

Table 7: English-German and German-English primary submissions to the WMT19 shared task.

Year	Best in competition	This work	Δ
2017	28.3	32.8	**+4.5**
2018	48.3	49.3	**+1.0**
2019	44.9	43.0	**-1.9**

Table 8: Comparison of our English-German system with the winning submissions over the past two years.

NMT-SMT hybrid systems Popular examples of combining a fully trained SMT system with independently trained NMT are rescoring and reranking methods (Neubig et al., 2015; Stahlberg et al., 2016b; Khayrallah et al., 2017; Grundkiewicz and Junczys-Dowmunt, 2018; Avramidis et al., 2016; Marie and Fujita, 2018; Zhang et al., 2017), although these models may be too constraining if the neural system is much stronger than the SMT system. Loose combination schemes include the edit-distance-based system of Stahlberg et al. (2016a) or the minimum Bayes-risk approach of Stahlberg et al. (2017a) we adopted in this work. NMT and SMT can also be combined in a cascade, with SMT providing the input to a post-processing NMT system (Niehues et al., 2016; Zhou et al., 2017) or vice versa (Du and Way, 2017). Wang et al. (2017b, 2018) interpolated NMT posteriors with word recommendations from SMT and jointly trained NMT together with a gating function which assigns the weight between SMT and NMT scores dynamically. The AMU-UEDIN submission to WMT16 let SMT take the lead and used NMT as a feature in phrase-based MT (Junczys-Dowmunt et al., 2016b). In contrast, Long et al. (2016) translated most of the sentence with an NMT system, and just used SMT to translate technical terms in a post-processing step. Dahlmann et al. (2017) proposed a hybrid search algorithm in which the neural decoder expands hypotheses with phrases from an SMT system.

6 Conclusion

Our WMT19 submission focused on regularized fine-tuning and language modelling. With our novel Intra-Inter Transformer architecture for document-level LMs we achieved significant reductions in perplexity and minor improvements in BLEU over very strong baselines. A combination of checkpoint averaging and EWC proved to be an effective way to regularize fine-tuning. Our systems are competitive on both English-German and German-English (Tab. 7), especially considering the immense speed with which our field has been advancing in recent years (Tab. 8).

Acknowledgments

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) grant EP/L027623/1 and has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service[6] funded by EPSRC Tier-2 capital grant EP/P020259/1.

[6] http://www.hpc.cam.ac.uk

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, pages 265–283. USENIX Association.

Eleftherios Avramidis, Vivien Macketanz, Aljoscha Burchardt, Jindrich Helcl, and Hans Uszkoreit. 2016. Deeper machine translation and evaluation for German. In *Proceedings of the 2nd Deep Machine Translation Workshop*, pages 29–38. ÚFAL MFF UK.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM.

Leonard Dahlmann, Evgeny Matusov, Pavel Petrushkov, and Shahram Khadivi. 2017. Neural machine translation leveraging phrase-based models in a hybrid search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1411–1420, Copenhagen, Denmark. Association for Computational Linguistics.

Praveen Dakwale and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, page 117.

Jinhua Du and Andy Way. 2017. Neural pre-translation for hybrid machine translation. *In Proceedings of MT Summit XVI*, 1:27–40.

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

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137 – 148.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Marcin Junczys-Dowmunt. 2018a. 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. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *International Workshop on Spoken Language Translation IWSLT*.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016b. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany. Association for Computational Linguistics.

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn. 2017. Neural lattice search for domain adaptation in machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 20–25, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44, Melbourne, Australia. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn, Kevin Duh, and Brian Thompson. 2018a. The JHU machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 438–444, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018b. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Cache-based document-level neural machine translation. *arXiv preprint arXiv:1711.11221*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. 2016. Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 47–57. The COLING 2016 Organizing Committee.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Benjamin Marie and Atsushi Fujita. 2018. A smorgasbord of features to combine phrase -based and neural machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, Boston, US.

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. 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.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Belgium, Brussels. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for Java.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 35–41, Kyoto, Japan. Workshop on Asian Translation.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836. The COLING 2016 Organizing Committee.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018. Multi-representation ensembles and delayed SGD updates improve syntax-based NMT. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–325. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The RWTH Aachen university supervised machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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

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

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018a. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211. Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018b. The University of Cambridge's machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 504–512. Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017a. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368. Association for Computational Linguistics.

Felix Stahlberg, Eva Hasler, and Bill Byrne. 2016a. The edit distance transducer in action: The University of Cambridge English-German system at WMT16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 377–384. Association for Computational Linguistics.

Felix Stahlberg, Eva Hasler, Danielle Saunders, and Bill Byrne. 2017b. SGNMT – A flexible NMT decoding platform for quick prototyping of new models and search strategies. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016b. Syntactically guided neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 299–305. Association for Computational Linguistics.

Felix Stahlberg, Danielle Saunders, Gonzalo Iglesias, and Bill Byrne. 2018c. Why not be versatile? Applications of the SGNMT decoder for machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 208–216. Association for Machine Translation in the Americas.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

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

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. 2017a. 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.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2017b. Neural machine translation advised by statistical machine translation. In *AAAI*, pages 3330–3336.

Xing Wang, Zhaopeng Tu, and Min Zhang. 2018. Incorporating statistical machine translation word knowledge into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2255–2266.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2018. Modeling coherence for discourse neural machine translation. *arXiv preprint arXiv:1811.05683*.

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.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Improving neural machine translation through phrase-based forced decoding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 152–162, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384. Association for Computational Linguistics.