The MLLP-UPV Supervised Machine Translation Systems for WMT19 News Translation Task

Javier Iranzo-Sánchez, Gonçal V. Garcés Díaz-Munío, Jorge Civera, Alfons Juan
Machine Learning and Language Processing (MLLP) research group
Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{jairsan,ggarces,jcivera,ajuan}@vrain.upv.es

Abstract

This paper describes the participation of the MLLP research group of the Universitat Politècnica de València in the WMT 2019 News Translation Shared Task. In this edition, we have submitted systems for the German ↔ English and German ↔ French language pairs, participating in both directions of each pair. Our submitted systems, based on the Transformer architecture, make ample use of data filtering, synthetic data and domain adaptation through fine-tuning.

1 Introduction

In this paper we describe the supervised Statistical Machine Translation (MT) systems developed by the MLLP research group of the Universitat Politècnica de València for the News Translation Shared Task of the *ACL 2019 Fourth Conference on Machine Translation* (WMT19). For this year's edition, we participated in both directions of the German ↔ English and German ↔ French language pairs, using Neural Machine Translation (NMT) models following the Transformer (Vaswani et al., 2017) architecture. Following the lessons learned from last year, we have continued working on data filtering, and we have experimented with additional synthetic data techniques and bigger neural network architectures trained with multi-GPU machines.

This paper is organized as follows. Section 2 describes the data processing steps (including data filtering and synthetic data generation) carried out prior to system training. Section 3 describes the architecture and settings used for our NMT models, and the different experiments and evaluations performed are detailed in Section 4. Our conclusions for this shared task are outlined in Section 5.

2 Data preparation

Data preprocessing, corpus filtering and data augmentation are described in the following sections.

2.1 Corpus preprocessing

The data was processed using the standard Moses pipeline (Koehn et al., 2007). Specifically, we normalized punctuation, and tokenized and truecased data. Additionally, we applied 40K BPE operations (Sennrich et al., 2016b), learned jointly over the source and target languages, and excluded from the vocabulary all subwords that did not appear at least 10 times in the training data. BPE operations are learned before adding the data extracted using corpus filtering, described in Section 2.2. Sentences longer than 100 subwords were excluded from the training data.

2.2 Corpus filtering

The addition of the ParaCrawl corpus to the WMT shared tasks has placed an increasing importance in filtering and data selection techniques in order to take advantage of this additional data. This is highlighted by the fact that a majority of participating systems in the WMT18 News Translation Task (Bojar et al., 2018) apply filtering techniques to ParaCrawl. Additionally, the experiments carried out for our 2018 submission (Iranzo-Sánchez et al., 2018) show that using a noisy corpus such as ParaCrawl without filtering can result in a worse performance compared with a baseline system that simply excludes the noisy corpus from the training data.

We have compared two different approaches to corpus filtering:

- **LM-based filtering** (Iranzo-Sánchez et al., 2018): This approach uses language models for estimating the quality of a sentence pair, under the assumption that a low-perplexity

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 218–224
Florence, Italy, August 1-2, 2019.

sentence is more likely to be an adequate sentence for training. Using in-domain data, we train one language model for each language, and then use them to score the corresponding side of the sentence pair, giving us perplexity scores (s, t). The score (perplexity) of a sentence pair is the geometric mean $\sqrt{s \cdot t}$. We select sentence pairs with the lowest score. This is the approach we used for our WMT18 submission.

- **Dual Conditional Cross-Entropy filtering** (Junczys-Dowmunt, 2018): This approach computes the sentence pair score by means of a product of a series of partial scores.

$$f(x, y) = \prod_i f_i(x, y) \quad (1)$$

We have used the same configuration sent for the WMT18-filtering task, which uses 3 partial scores: a language identification score (*lang*), a dual conditional cross-entropy score (*adq*), and a cross-entropy difference score (*dom*) with a cut-off value of 0.25. The full details of each of these partial scores is given in Junczys-Dowmunt (2018). The translation models for the *adq* score are Transformer Base models trained with the Europarl portion of WMT19. In terms of the data for the *dom* score, we randomly sampled 1M sentences from NewsCrawl 2016 as in-domain data, and 1M sentences from ParaCrawl as out-of-domain data.

We carried out a series of comparisons between the two techniques, and found out that the cross-entropy model provides better performance than the LM-based filtering model. This is consistent with the fact that the cross-entropy filtering was the winning submission to the WMT18 Shared Task on Parallel Corpus Filtering (Koehn et al., 2018). As a result, we have elected to use the cross-entropy filtering method for filtering the different versions of the ParaCrawl corpus present in all language pairs.

2.3 Synthetic source sentences

The use of synthetic data produced by means of the backtranslation technique (Sennrich et al., 2016a) is an effective way of benefiting from additional monolingual data. Further improvements are possible if the data is from the same domain as the test data. For this reason, we have produced synthetic data for all the language pairs we have participated in.

We used the following configuration:

- German → English: We have used 20M sentences from our WMT18 submission (Iranzo-Sánchez et al., 2018), and an additional 24M sentences generated using a system with the same configuration as WMT18, but trained with 3 GPUs instead of 1. The monolingual sentences were randomly sampled from News Crawl 2017.

- English → German: We have generated 18M sentences using our German → English system submitted to WMT18, with monolingual sentences randomly sampled from News Crawl 2017.

- German → French: We have generated 10M synthetic sentences, using the reverse direction baseline system described in Section 3. The monolingual sentences were sampled from News Crawl 2015-2018.

- French → German: We have generated 18M synthetic sentences, using the reverse direction baseline system described in Section 3. The monolingual sentences were sampled from News Crawl 2017.

Prior to selecting sentences, we filtered out from the German News Crawl 2017 all sentences that were written in a language different from German, using the `langid` tool (Lui and Baldwin, 2012). When combining bilingual and synthetic data, the original bilingual data was upsampled in order to achieve a 1:1 ratio.

3 System description

This section describes the configuration and decisions adopted for training our NMT systems. We will first begin by describing the details that are common to all systems, and we will then move on to specific details for each of the considered translation directions.

Our models follow the Transformer architecture (Vaswani et al., 2017), and are configured based on the Transformer Base and Transformer Big settings.

The Transformer Base models are trained with a batch size of 3000 tokens per GPU, whereas

the Transformer Big models use a batch size of 2300 tokens per GPU. We store a checkpoint every 10 000 updates, and inference is carried out by averaging the last 8 checkpoints.

We used the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$,$\beta_2 = 0.98$. The learning rate was updated following an inverse square-root schedule, with an initial learning rate of 0.0005, and 4000 warm-up updates. All models use 0.1 label smoothing (Szegedy et al., 2016) and 0.1 dropout (Srivastava et al., 2014), with the exception of the German ↔ French models, that use 0.3 dropout due to having less training data.

The systems from our WMT18 submission and this year's baseline systems were built using the Sockeye toolkit (Hieber et al., 2017). The rest of the systems were built using the fairseq toolkit (Ott et al., 2019), in order to train using Half Precision and gradient accumulation like in Ott et al. (2018).

3.1 Finetuning

Finetuning (training on a new set of data after system convergence) has been widely used as a method for domain-adaptation in NMT systems (Luong and Manning, 2015; Sennrich et al., 2016a). Due to the different data sources provided in the competition, and possible domain mismatch between training and test data, we have decided to carry out finetuning in order to improve model performance. The goal of adapting our models to the domain of the test data is achieved by using test sets from previous years as in-domain data for finetuning.

To carry out finetuning, we set the learning rate to the value that was being used when training finished, and we reduced the checkpoint interval in order to store a checkpoint every 20 updates. Finetuning continues as long as the performance does not decrease in the appropriate dev set. For the German ↔ English systems, we follow the setup of Schamper et al. (2018), and use test sets from previous years (newstest08-16) as training data for the finetuning step. Since this is the first time the German ↔ French language pair is included in WMT, we do not have available test sets from previous editions, so we resort to using the dev1 set as training data, and stop finetuning when performance drops in dev2 (see Section 4).

4 Experimental evaluation

This section describes the experiments and evaluation carried out for each of the language directions, with special emphasis placed in the German ↔ English systems.

For the German ↔ English systems, we have used newstest2017 as dev set, and newstest2018 as test set. Additionally, we report results on this year's test set, newstest 2019. For the German ↔ French systems, we splitted in half the supplied euelections dev set into two sets, dev 1 and dev 2, and used the former as dev set and the latter as test set. We also report the results obtained in the official test set newstest2019. We report BLEU scores (Papineni et al., 2002) computed using SacreBLEU (Post, 2018).

4.1 German → English

Table 1 shows the results obtained by our systems trained for the German → English direction. As baselines, we take our WMT18 system, trained with 1 GPU (this is the configuration that was used for our WMT18 submission), and the same setup trained with 3 GPUs. The increase in effective batch size from 3000 to 9000 tokens results in an improvement of 1.7 BLEU in newstest2018 and 2.0 BLEU in newstest2019 without any other change in hyperparameters.

We began our WMT19 experiments by building a system following the Transformer Big architecture, trained in a 4-GPU machine and using the 20M backtranslations produced for WMT18. This results in an increase of 0.3 BLEU in newstest2018 and 0.6 BLEU in newstest2019. We then applied gradient accumulation by setting the Update Frequency (UF) to 2. Under this setting, the model's weights are updated every two steps (this simulates a batch size equivalent to training on 8 GPUs). This model obtains a significant improvement in the dev (+0.7 BLEU), and test sets (+1.4 BLEU), however the performance decreases by 0.7 BLEU when evaluating on newstest2019. We have found no explanation for this phenomenon. Finetuning on the news in-domain data results improves all previous results, resulting in 47.8 BLEU in newstest2018 and 39.4 BLEU in newstest2019.

For our final submission, we trained a system with noisy backtranslations, following the work of Edunov et al. (2018). We used the previous 20M backtranslations and appended an additional 24M generated with the system in row 2 of Table 1. We

System	GPUs	BLEU newstest2018	BLEU newstest2019
WMT18 (Transformer Base)	1	44.2	35.6
WMT18 (Transformer Base)	3	45.9	37.6
Transformer Big, 20M backtrans	4	46.2	38.3
+ UF=2	4	47.6	37.7
+ finetuned	4	47.8	39.4
+ 24M backtrans, noise (non-converged)	4	47.5	39.9
+ finetuned	4	48.0	39.3
+ 24M backtrans, noise (converged)	4	48.0	40.2
+ finetuned	4	47.9	40.1

Table 1: Evaluation results of German → English systems

added noise to the source side of the synthetic sentence pairs using the technique described by Lample et al. (2018). Following the setup of Edunov et al. (2018), bilingual data was not upsampled, resulting in a ratio of around 1:3 original to synthetic sentences. The system had not converged at the time of the shared task deadline, so we report results both from our submission, which was generated when the system was still training, as well as the results from the converged system, obtained after the competition ended.

The system trained with noisy backtranslation obtains 47.5 BLEU in newstest2018 and 39.9 BLEU in newstest2019. An additional finetuning step improves the results in newstest2018 by 0.5 BLEU. Due to having obtained the best results in the test set, this was the system we submitted to the competition. However, when evaluating the finetuned version with this year's test set, we find a decrease of 0.6 BLEU. Allowing the system to train for additional epochs leaves us with a final result of 48.0 BLEU and 40.2 BLEU in newstest2018 and newstest2019, and 47.9 and 40.1 BLEU, respectively, after finetuning.

We observe that, in the case of the noisy system, finetuning seems to obtain mixed results, in contrast with other trained systems and language directions (see Sections 4.2, 4.3 and 4.4), where finetuning achieves a performance increase in all cases. We theorize this could be due to the fact that the system was first trained with a ratio that included 3 times as many noisy sentences as clean data, but the finetuning was carried out only with clean data, without any added noise.

4.2 English → German

Table 2 shows the results obtained by our systems trained for the English → German direction. We began with a baseline system trained using our WMT18 configuration and data, plus an additional 18M backtranslations. This system obtains 45.2 BLEU in newstest2018 and 39.3 BLEU in newstest2019. For our WMT19 submission, we trained a Transformer Big model, using the WMT19 data (including 10M filtered sentences from ParaCrawl), as well as the already mentioned 18M backtranslations. This system was trained with 2 GPUs and an Update Frequency of 2, giving us an effective batch size equivalent to 4 GPUs. This system obtains an improvement of 0.4 BLEU in newstest2018 and 0.1 BLEU in newstest2019 over the baseline. Increasing the number of GPUs from 2 to 4 shows no significant differences in either newstest2018 or newstest2019. Our final submission was generated after applying a finetuning step to the previous configuration. This finetuning resulted in an increase of 2.4 BLEU in newstest2018 and 2.3 BLEU in newstest2019 when compared with the non-finetuned model.

4.3 German → French

Table 3 shows the results obtained by our systems trained for the German → French direction. Our baseline system is a Transformer Base model trained with all the WMT19 data excluding ParaCrawl. This system obtains 31.3 BLEU in dev2 and 32.1 BLEU in newstest2019. We then moved on to training a Transformer Big model, adding 1M sentences filtered from ParaCrawl, and 10M backtranslations generated with the French → German baseline system. This system was trained with 2 GPUs and an Update Frequency

System	GPUs	BLEU newstest2018	BLEU newstest2019
WMT18 (Transformer Base), 18M backtrans	3	45.2	39.3
Transformer Big, 18M backtrans, UF=2	2	45.6	39.4
+ GPU=4	4	45.7	39.4
+ finetuned	4	48.1	41.7

Table 2: Evaluation results of English → German systems

System	GPUs	BLEU dev2	BLEU nt2019
WMT19 - {ParaCrawl}	1	31.1	32.1
Transformer Big, UF=2	2	33.3	34.4
+ finetuning	2	33.5	34.5

Table 3: Evaluation results of German → French systems

System	GPUs	BLEU dev2	BLEU nt2019
WMT19 - {ParaCrawl}	1	22.8	25.7
Transformer Big, UF=2	2	24.9	26.9
+ finetuning	2	25.4	27.5

Table 4: Evaluation results of French → German systems

of 2. This results in an increase of 2.2 BLEU in dev2 and 2.3 BLEU in newstest2019. An additional finetuning step, carried out using the dev1 data, results in an increase of 0.2 BLEU in dev2 and 0.1 BLEU in newstest2019, and constituted our submission to the competition.

4.4 French → German

Table 4 shows the results obtained by our systems trained for the French → German direction. The approach and configurations for this language directions mirror those of the German → French direction (Section 4.3). We began with a baseline Transformer Base model, that obtains 22.8 BLEU in dev2 and 25.7 BLEU in newstest2019. The Transformer Big model obtains an improvement of 2.1 BLEU in dev2 and 1.2 BLEU in newstest2019, and the finetuning step results in an additional increase of 0.5 BLEU in dev2 and 0.6 BLEU in newstest2019.

5 Conclusions

The experiments carried out this year have allowed us to explore one of the missing pieces of our WMT18 submission, which is the interaction between the Transformer architecture and different batch sizes. The results show that the performance of models following the Transformer architecture is highly dependent on the batch size used to train the model, requiring multiple GPUs or gradient accumulation in order to fully take advantage of this architecture. This result is consistent with other works such as Popel and Bojar (2018).

As future work, we would like to look further into using massive amounts of synthetic data jointly with noise, as our experiments this year have not provided conclusive results. Overall, the finetuning steps loos like an effective way of obtaining translation improvements, at the expense of only a small amount of computation. This domain adaptation step can be carried out as long as we have some amount of in-domain data available. More work needs to be carried out to explore the interaction between finetuning and adding noise to the data. Another avenue for improvement is to look into the optimal amount of filtered data to extract from ParaCrawl, as well as the upsampling ratio to mix bilingual and synthetic data. These aspects were not explored in our WMT19 submission due to time constraints.

Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 761758 (X5gon); the Government of Spain's research project Multisub, ref. RTI2018-094879-B-I00 (MCIU/AEI/FEDER, EU); and the Universitat Politècnica de València's PAID-01-17 R&D support programme.

References

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Pro-*

ceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pages 272–303.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adria A. Martinez-Villaronga, Jorge Civera, and Alfons Juan. 2018. The MLLP-UPV german-english machine translation system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 418–424.

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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 888–895.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, 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, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 726–739.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*.

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. *arXiv preprint arXiv:1904.01038*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9.

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, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

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. 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 496–503.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural

networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.