The MLLP-UPV German-English Machine Translation System for WMT18

Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, Alfons Juan
Machine Learning and Language Processing research group
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{jiranzo,pbaquero,ggarces,amartinez1,jcivera,ajuan}@dsic.upv.es

Abstract

This paper describes the statistical machine translation system built by the MLLP research group of Universitat Politècnica de València for the German→English news translation shared task of the EMNLP 2018 Third Conference on Machine Translation (WMT18). We used an ensemble of Transformer architecture–based neural machine translation systems. To train our system under "constrained" conditions, we filtered the provided parallel data with a scoring technique using character-based language models, and we added parallel data based on synthetic source sentences generated from the provided monolingual corpora.

1 Introduction

In this paper we describe the statistical machine translation (SMT) system built by the MLLP research group of Universitat Politècnica de València for the German→English news translation shared task of the EMNLP 2018 Third Conference on Machine Translation (WMT18).

Neural Machine Translation (NMT) has made great advances over the last few years, and in particular it has come to outperform Phrase-Based Machine Translation (PBMT) and PBMT-NMT combinations in the most recent WMT shared news translation tasks (Bojar et al., 2016, 2017). Taking this into account, we decided to build an NMT system taking as a basis the Transformer architecture, which has been shown to provide state-of-the-art SMT results while requiring relatively short times to train (Vaswani et al., 2017).

Apart from the SMT system itself, we also describe our work on parallel-corpus preprocessing and filtering, an aspect which has gained importance in WMT18 with the addition of the much larger and noisier parallel corpus ParaCrawl. Regarding data augmentation, we report as well how we extended the provided parallel dataset with data based on synthetic source sentences generated from the provided target-language monolingual corpora (in compliance with this shared task's "constrained" conditions).

This paper is organized as follows: in Section 2, we outline the data preparation techniques that were used (corpus preprocessing, corpus filtering, and data augmentation with synthetic source sentences); Section 3 shows the architecture and parameters of our NMT system and our system combination; in Section 4, we report our experiments and results (including on data preparation and on final system evaluation); and we draw our final conclusions in Section 5.

2 Data preparation

In this section, we describe the techniques that we used to prepare the provided WMT18 German↔English data (parallel and monolingual) to improve our SMT system results: corpus preprocessing (Section 2.1), corpus filtering (Section 2.2) and parallel data augmentation with synthetic source sentences (Section 2.3).

Corpus preprocessing and filtering has acquired a new relevance in WMT18, due to the addition of the new ParaCrawl parallel corpus, which sextuplicates the amount of German↔English parallel data that was available in WMT17 and previous editions: there are approx. 36 million sentence pairs in ParaCrawl, versus approx. 6 million in the rest of the parallel corpora (for a total sum of approx. 42 million sentence pairs in the full WMT18 training data). This is illustrated in Table 1, which summarizes the number of sentences of each corpus in the provided parallel dataset.

While the large size of the ParaCrawl parallel corpus makes it a valuable resource for system training in WMT18, it is much noisier than

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 418–424

https://doi.org/10.18653/v1/W18-64041

Table 1: Size by corpus of the WMT18 parallel dataset

Corpus	Sentences (M)
News Commentary v13	0.3
Rapid (press releases)	1.3
Common Crawl	1.9
Europarl v7	2.4
ParaCrawl	36.4
WMT18 total	42.3

the rest of the WMT corpora. By noise here we mean misaligned sentences, wrong languages, meaningless sentences...; that is, sentence pairs which hinder system training for the purpose of German→English translation. In our experiments, we have observed that preprocessing and filtering the ParaCrawl corpus is necessary in order to make it useful as training data with the goal of increasing translation quality. In fact, using the ParaCrawl corpus "as is", we not only did not find any improvement in translation quality, but we even observed a degradation in all metrics of quality (as we will detail in Section 4.2).

Regarding data augmentation, the usage of relevant in-domain monolingual data has been shown to be important in order to improve NMT system results (Sennrich et al., 2016a). The provided WMT18 dataset contains large amounts of monolingual data which we can take advantage of to increase system accuracy. This fact led us to use these monolingual resources to generate additional synthetic data from target-language sentences.

2.1 Corpus preprocessing

Our preprocessing was done as suggested by the WMT18 organization (WMT18 organizers, 2018) using the provided scripts, with punctuation normalization, tokenization, cleaning and truecasing using standard Moses scripts.

Additionally, we removed from the training corpus any sentence that contained strange characters, defined as those lying outside the Latin UTF interval (u0000-u20AC) plus the euro sign (€). This allowed us to reduce the vocabulary size by eliminating unnecesary characters belonging to languages other than German or English that are not required for the translation of online news.

2.2 Corpus filtering

In regard to data filtering, we aimed to filter out noisy sentence pairs from the parallel corpora. To this end, we trained two separate 9-gram character-based language models (one for German, one for English) on the newstest2014 development set, based on which we computed the perplexity for each sentence in the full WMT18 dataset (including ParaCrawl), in a manner similar to the techniques described by Yasuda et al. (2008), Foster et al. (2010) and Axelrod et al. (2011). The software used was the SRI Language Modelling Toolkit (Stolcke et al., 2011).

The idea was that the lower the perplexity for a given sentence with respect to a reference news test corpus, the lower the odds of this sentence being noise (for the purpose of training a German→English SMT system). At the same time, this method could be considered to provide some degree of domain adaptation, since we score the sentences with respect to an in-domain reference corpus.

To produce the final score for each sentence pair, we combined the perplexity scores (s, t) with the geometric mean ($\sqrt{s \cdot t}$). The geometric mean of two character-based perplexities can be interpreted as the character-based perplexity of the concatenation, assuming both sentences have the same number of characters. This is usually not the case exactly, but it is a good enough approximation. As the square root is a monotonic function, it does not alter the order of the scores.

We then ranked all the sentence pairs in the full WMT18 dataset according to their combined perplexity score, and selected subsets of different sizes, taking in each case the n lowest-scored (less noisy) sentence pairs.

2.3 Synthetic source sentences

We augmented the WMT18 German↔English parallel training dataset (while keeping it under "constrained" conditions) with synthetic source sentences generated from the provided target-language monolingual corpora. To this end, we followed the approach outlined by Sennrich et al. (2016a).

In particular, we trained an English→German NMT system based on our best system configuration for German→English. Then, we used this system to generate our synthetic source sentences (German) from a subset of the WMT18 target-language monolingual corpora (English), which provided us with a significant amount of new sentence pairs to use as in-domain synthetic training data.

3 System description

We decided to build an NMT system based on the Transformer architecture (Vaswani et al., 2017). We opted for a pure NMT system due to the great advances this technology has made in the field of SMT over the last few years, which has led it to outperform systematically the more traditional PBMT systems and PBMT-NMT combinations, as introduced in Section 1. In particular, the Transformer architecture, based on self-attention mechanisms, can provide state-of-the-art SMT results while keeping training times relatively short. Regarding the software used, we used the Sockeye NMT framework (Hieber et al., 2017).

We based our systems on the less complex Transformer "base" configuration, which has significantly fewer parameters than the "big" configuration (65M parameters in the former case, 213M in the latter), and is thus much quicker to train (in exchange for a relatively small decrease in translation quality, in the case of the experiments described by Vaswani et al. (2017)). This was important in order to complete our experiments and the final training of our primary system in time for participation in this shared task. Thus, our models use 6 self-attentive layers both on the encoder and the decoder, with a model dimension of 512 units and a feed-forward dimension of 2048 units.

During training, we applied 0.1 dropout and 0.1 label smoothing, the Adam optimization scheme (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate annealing: we set an initial learning rate of 0.0001, and scaled this by a factor of 0.7 whenever the validation perplexity did not improve in 8 consecutive checkpoints (each checkpoint being equivalent to 2000 parameter updates). The system was trained with a word-based batch size of 3000, and a maximum sentence length of 75 tokens (subword units).

For our internal experiments, all systems were trained after applying 20K BPE operations (Sennrich et al., 2016b); but when building our final submissions, we increased this amount to 40K BPE operations (this will be detailed for each system in Section 4.4).

The final system consists of an ensemble of 4 independent training runs of our best model, based on a linear combination of the individual probabilities.

4 Experimental evaluation

In this section, we outline our experimental setup (Section 4.1); we report our experiments and results on corpus filtering (Section 4.2); we detail our setup for parallel data augmentation with synthetic source sentences (Section 4.3); and we discuss our final German→English NMT system evaluation and results (Section 4.4).

4.1 Experimental setup

For our experiments, we used newstest2015 as the development set and newstest2017 as the test set. We also report the results obtained with this year's newstest2018.

We evaluated our systems using the BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) measures, using mteval from the Moses SMT toolkit (Koehn et al., 2007) and tercom (Snover et al., 2008), respectively. All reported scores are according to the instructions on system output formatting provided by the WMT18 organization.

4.2 Results on corpus filtering

We show here the results obtained with the corpus filtering techniques explained in Section 2.2.

Table 2 summarizes the results in translation quality obtained with different subsets of the WMT18 parallel dataset. We can see that using the full WMT18 parallel dataset (42M sentence pairs), including the ParaCrawl corpus "as is", leads to a significant degradation in all metrics of quality compared to using the WMT18 dataset excluding ParaCrawl (6M sentence pairs; our baseline system for system evaluations in Section 4.4). Furthermore, we see that if we restrict ourselves to an excessively small training dataset (5M sentence pairs) using our filtering approach, there is also a degradation in quality with respect to using the unfiltered WMT18 dataset excluding ParaCrawl (6M).

We can also see (focusing on the test set results, newstest2017) that our filtering approach is effective at selecting useful training data from ParaCrawl, in the fact that the filtered datasets with sizes over the baseline's 6M sentence pairs provide significant improvements in quality (even if we limit ourselves to the small increase in size of the 7.5M subset). At the other extreme, in our experiments, going over 15M filtered sentence pairs meant setting the threshold for noise too low, as quality metrics degraded again.

Table 2: Results of 9-gram character-based language model data filtering, by number of selected sentences

Subset (no. of parallel sentences)	newstest2015 BLEU	newstest2015 TER	newstest2017 BLEU	newstest2017 TER	newstest2018 BLEU	newstest2018 TER
Full WMT18 parallel dataset (42M)	20.6	71.1	21.3	70.2	26.2	64.2
Baseline: WMT18 minus ParaCrawl (6M)	31.1	55.4	32.0	54.8	39.1	46.3
Filtered corpus (5M)	30.3	56.3	31.4	55.5	38.7	46.5
Filtered corpus (7.5M)	32.8	54.0	33.7	56.5	41.5	44.5
Filtered corpus (10M)	33.0	53.7	34.5	52.9	42.2	43.7
Filtered corpus (15M)	33.4	53.2	34.3	52.7	42.2	43.6

As Table 2 shows, we obtained our best test set results with the 10M and 15M subsets. As results were very similar in both cases, we considered that any possible improvements in quality obtained from using the larger 15M subset were too small to justify using it instead of the 33% smaller 10M subset (which has a significantly faster convergence time for system training). Thus, the 10M subset is the filtered training corpus we took as a basis for the subsequent work described in Sections 4.3 and 4.4.

As a downside, this data filtering method based on independent language models for each side of a noisy parallel corpus has the caveat of not being able to detect sentence pairs where the source and the target are valid sentences, but not actually a translation of each other. To avoid this problem, it could be useful to combine into the filtering method the score provided by a simple, quick translation model (which should provide better scores for the sentence pairs which are correctly aligned translations). While we carried out some preliminary experiments on filtering with this approach, we did not obtain conclusive improvements in time for this shared task, so we left this for future work.

We also left for future work further corpus filtering experiments with other data selection approaches, such as using the cross-entropy difference (rather than just perplexity or cross entropy) to score each sentence pair (Moore and Lewis, 2010), or the dynamic data selection method described by van der Wees et al. (2017).

4.3 Synthetic source sentence setup

Here we detail how we augmented the WMT18 German↔English parallel training dataset, based on the technique introduced in Section 2.3.

We created an English→German NMT system using our best parameters for German→English (as described in Section 3), and trained it with the 10M-sentence filtered WMT18 parallel dataset that had shown the best performance for German→English (as described in Section 4.2). For reference, the resulting English→German NMT system obtained 27.4 BLEU on newstest2017. While improving this "inverse" system with further experiments could result in better synthetic training data (Sennrich et al., 2016a), we settled on this configuration (which obtains reasonable results with respect to the best WMT17 systems) in order to be in time for participation in this shared task.

Then, using this system, we translated into German a random sample of 20M English sentences from News Crawl 2017 (the most recent in-domain corpus among the provided WMT18 monolingual corpora). This provided us with 20M sentence pairs of German→English in-domain synthetic training data.

This augmented corpus was used for the final systems the results of which are discussed in the following Section 4.4.

4.4 Final system evaluation and results

We will now describe the most significant results obtained with the German→English NMT models we trained for WMT18 (based on the architecture and parameters outlined in Section 3). These results are shown in Table 3.

Our baseline model was trained excluding the ParaCrawl corpus from the training data, since using the full WMT18 corpus (with ParaCrawl) actually led to worse results (as we saw in Section 4.2). As mentioned in Section 3, this system was trained with 20K BPE operations (as is the case with the next system we will describe).

Our first step to improve these baseline results was filtering the full WMT18 corpus (including ParaCrawl), as explained in Section 4.2. In Ta-

Table 3: Results of German→English MT system evaluations

System	newstest2015		newstest2017		newstest2018	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline (WMT18 minus ParaCrawl, 6M pairs)	31.1	55.4	32.0	54.8	39.1	46.3
Filtered corpus (including ParaCrawl, 10M pairs)	33.0	53.7	34.5	52.9	42.2	43.7
+ Synthetic data (2*10M+20M pairs), 40K BPE	34.3	52.0	35.9	51.2	44.7	41.1
Ensemble (x4)	34.6	51.9	36.2	51.0	45.1	40.8

ble 3 we show the result obtained with a system trained on our best filtered corpus. As we saw in Section 4.2, the 10M filtered corpus provides an improvement of 2.5 BLEU and 1.7 TER in the test set over the baseline model. This shows how our data-filtering approach has allowed us to extract useful sentences from the noisy ParaCrawl corpus and improve our system performance.

For our final systems, we added 20M synthetic sentence pairs as described in Section 4.3, and we oversampled the previous 10M filtered bilingual training set by duplicating it, which gave us a final training set with a total of 40M sentence pairs[1]. We also increased the number of BPE operations from 20K to 40K. A single system trained with this configuration obtained 35.9 BLEU and 51.2 TER in the test set. This represents a significant improvement of 1.4 BLEU and 1.7 TER over the previous model, explained by a combination of the additional sentence pairs and the increase in vocabulary size.

As reference of the training times required, training a system with this final configuration took approx. 120 hours on a single-GPU system (Nvidia GeForce GTX 1080 Ti)[2].

Finally, our primary submission for WMT18 consists of an ensemble of 4 independent training runs with this final configuration, resulting in 36.2 BLEU and 51.0 TER in our test set, and 45.1 BLEU and 40.8 TER in newstest2018.

[1]Oversampling the 10M original training set was a measure intended to keep in check the weight of the comparatively large 20M synthetic training data. We left for future work experimenting with different ratios of synthetic versus original data, such as 1:1 (Sennrich et al., 2016a; Fadaee et al., 2017), as additional comparison terms to determine the best performing configuration.

[2]While our systems were trained on single-GPU machines, multi-GPU system training with proportionally larger batch sizes (larger than the 3000 words per batch we used, as noted in Section 3) could deliver better translation quality results (Vaswani et al., 2017). We left this for future work.

5 Conclusions

The MLLP group of the Universitat Politècnica de València has participated in the German→English WMT18 news translation shared task with an ensemble of neural machine translation models based on the Transformer architecture. Our models were trained using a filtered subset of the provided parallel training dataset, plus augmented parallel data based on synthetic source sentences generated from the provided monolingual corpora. Our primary submission was an ensemble of four independent training runs of our best model parameters.

Our results point to the usefulness of the Transformer NMT architecture to obtain highly competitive SMT results with a relatively low computational cost (which can contribute to "democratizing" access to state-of-the-art research in NMT to a higher number of research groups, even those with more modest computational equipment). We have also shown the importance of adequate corpus filtering to make the most of larger, noisier parallel corpora, employing a simple approach to filtering using character-based language models that has resulted in significant improvements in translation quality.

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 Spanish government's TIN2015-68326-R (MINECO/FEDER) research project MORE, university collaboration grant programme 2017-2018, and faculty training scholarship FPU13/06241; the Generalitat Valenciana's predoctoral research scholarship ACIF/2017/055; as well as the Universitat Politècnica de València's PAID-01-17 R&D support programme.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, Pennsylvania, USA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Marzie Fadaee, Aarianna Bisazza, and Christof Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation. *ArXiv e-prints (arXiv:1705.00440)*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, Massachussets, USA.

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

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, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Stroudsburg, Pennsylvania, USA.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, Pennsylvania, USA.

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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachussets, USA.

Matthew Snover, Shuguang Wang, and Spyros Matsoukas. 2008. Translation Error Rate. http://www.cs.umd.edu/~snover/tercom/. [Online; accessed 6-July-2018].

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark.

WMT18 organizers. 2018. WMT18 Shared Task: Machine Translation of News. http://www.statmt.org/wmt18/translation-task.html. [Online; accessed 24-July-2018].

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 655–660, Hyderabad, India.