LIUM's Contributions to the WMT2019 News Translation Task: Data and Systems for German↔French Language Pairs

Fethi Bougares
LIUM, Le Mans Université
fethi.bougares@univ-lemans.fr

Jane Wottawa
LIUM, Le Mans Université
jane.wottawa@univ-lemans.fr

Anne Baillot
3L.AM, Le Mans Université
anne.baillot@univ-lemans.fr

Loïc Barrault
LIUM, Le Mans Université
loic.barrault@univ-lemans.fr

Abstract

This paper describes the neural machine translation (NMT) systems of the LIUM Laboratory developed for the French ↔ German news translation task of the Fourth Conference on Machine Translation (WMT 2019). The chosen language pair is included for the first time in the WMT news translation task. We describe how the training and the evaluation data was created. We also present our participation in the French ↔ German translation directions using self-attentional Transformer networks with small and big architectures.

1 Introduction

Since the start of the WMT translation shared tasks in 2006, English has been involved in the majority of translation directions. Few exceptions have been seen in 2012 and 2013 where Czech was also proposed as source and target for several language pairs. This overwhelming disparity is due to the fact that English is available in large quantity, in both monolingual and bilingual corpora.

We think that this may be problematic for research purposes since considering English (either as source or target language) may hide many linguistic problems. For example, considering gender agreement, which does not exist in English, translating from English is harder because of the lack of source side information, and translating towards English is simpler since the agreement should be ignored. Generally speaking, English is a rather morphologically impoverished language, for instance having few gender agreement cases or conjugated verb forms. This contrasts with French and German where number and gender agreements are very frequent. That is why we introduced two new translation directions involving two European languages, namely French and German.

2 DE↔FR language pair

Training data

The training data for this language pair was created by cross-matching the training data from the previous WMT shared tasks for the EN-FR and EN-DE language pairs. The details of the corpora are provided in Table 1 in which we provide the original sizes of EN-FR and EN-DE corpora and the extracted parallel corpora in DE-FR. Overall, we were able to create a German-French parallel corpus with **153.2M** and **171.1M** words respectively.

Development and test data

The data collected for the FR↔DE language pair has been created from several online news websites. The development and test sets have been created from news articles in both French and German. The development set is the fruit of a collaboration with the Faculty of Literature and Humanities of the University of Le Mans during several Digital Humanities (DH) lab sessions. The purpose of these quality sessions is twofold: on the first hand, students would learn and comprehend the inherent concepts of using a computer assisted translation (CAT) tool in the context of DH classes (Baillot et al., 2019). On the other hand, the translated data is intended to be used for Machine Translation research purposes. This process led to a 1512 sentences[1] development corpus distributed during the WMT2019 shared task. While creating the development data we intentionally mixed (to some degree) the translation directions, therefore 462 sentences were translated from French to German and the reverse for the remaining 1050 sentences. The same process has

[1]The translations have been revised by professors from the Faculty of Literature and Humanities in order to reach the desired quality

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 129–133
Florence, Italy, August 1-2, 2019.

	FR-EN	DE-EN	FR-DE
europarl-v7	2M (52.5M/50.3M)	1.9M (44.6M/47.9M)	1.7M (46M / 41M)
Common Crawl	3.2M (76.6M/70.7M)	2.4M (47M/51.3M)	622k (14M/12.2M)
ParaCrawl	40.4M (663M/640M)	31.8M (467M/502M)	7.2M (110.6M/99.6M)
dev08-14	– 	– 	18k (417.1k/369.5k)

Table 1: Training corpora statistics (number of sentences) for FR↔DE News translation shared task. The second line of each cell corresponds to the number of tokens in French followed by the number of tokens in German.

been followed for the test set creation: 335 of the 1701 test sentences have been produced from French documents and the 1366 remaining pairs from German documents. We note that 756 out of the German 1366 German sentences in the test set have been translated into French by professional translators[2]. The dev and test sets are freely distributed and available for download[3].

	#lines	#token FR	#token DE
dev2019	1512	33833	28733
test2019	1701	38138	31560

Table 2: FR-DE dev and test set statistics.

3 LIUM Submissions

All our systems are constrained as we only used the supplied parallel data (described in table 1) with additional back-translations created from a subset of the monolingual news data made available by the shared task organizers.

3.1 Model Description

For our submissions we used the Transformer (Vaswani et al., 2017) sequence-to-sequence model as implemented in fairseq (Ott et al., 2019). Transformer is the state of the art NMT model which rely on a multi-headed attention applied as self-attention to source and target sentences. Our models are based on both small and big Transformer configurations. All experiments with the big transformer are models with 6 blocks in the encoder and decoder networks following the configuration described in (Ott et al., 2018). With respect to the small transformer model, we also used a 6 blocks encoder and decoder network with an embedding layer of size 512, a feed-forward layer with an inner dimension of 1024, and a multi-headed attention with 4 attention heads.

We use a vocabulary of 35K units based on a joint source and target byte pair encoding (Sennrich et al., 2016). We set the batch size to 2048 tokens and maximum sentence length to 150 BPE units, in order to fit the big Transformer configuration to our GPUs (NVIDIA GeForce GTX 1080 Ti with 11 GB RAM).

3.2 Data Preparation

Our preparation pipeline consists of a pre-processing step performed using scripts from Moses (Koehn et al., 2007). We replace the unicode punctuation, normalize the punctuation and remove the non-printing characters before the tokenization. After the tokenization step, we perform a cleaning stage where all source and target sentences with an overlapping rate higher than 65% are deleted. Statistics of the training corpora after the cleaning process are presented in table 2. These values should be contrasted with those of table 1 to assess the effect of the cleaning process. As it can be seen from tables 1 and 2, the effect of the cleaning step is more pronounced for the noisy parallel corpora (*i.e.* ParaCrawl and Common Crawl). For the europarl-v7 corpus, more than a thousand lines are removed after cleaning which mainly corresponds to English sentences in both languages: FR and DE as well as sentences with long lists of numbers.

In addition to the available parallel data, we have used monolingual News Crawl articles as additional synthetic bilingual data. We used only news 2018 from which we selected a subpart based on cross-entropy data selection method

[2]This was carried out by LinguaCustodia

[3]dev and test sets can be downloaded from https://github.com/lium-lst/euelections

	#lines	#token FR	#token DE
europarl-v7	1.7M	45.9M	40.9
Common Crawl	585k	13M	11M
ParaCrawl	6.7M	107M	95M
dev08-14	18k	417.1k	369.5k

Table 3: Training corpora statistics for FR↔DE systems after the cleaning process.

(Moore and Lewis, 2010). Data selection was performed with the *europarl* corpus as in-domain data and using the XenC Toolkit (Rousseau, 2013). By doing this, we were able to extract 3.4M German sentences out of the 38.6M sentences of the monolingual German 2018 News Crawl corpus. Similarly, 3.3M sentences were extracted out of the 8.2M monolingual French 2018 News Crawl.

4 Experiments and Results

In this section, we first present the results for German to French translation direction followed by the French to German direction. We use BLEU as evaluation metric (Papineni et al., 2002) and all reported scores are calculated using case-sensitive detokenized BLEU with multi-bleu.pl. All results use beam search with a beam width of 12 and length penalty of 1.

4.1 German to French

In this section we present the results for German to French direction. We have tried three different configurations differentiated by the training data used to create the NMT system. For each of these configurations, we trained a small and a big transformer model.

Given the prior knowledge about the noisy quality of the ParaCrawl corpus, we first tried to train some NMT systems with all available parallel data from table 3 except ParaCrawl. Table 4 contains the results for this setting. We report the results with the best checkpoint and an ensemble-decoding with 2 and 5 checkpoints for small and big Transformer versions. As expected, the big transformer outperforms the small version and we obtain an improvement of 1.69 BLEU point for the ensemble-decoding of 5 checkpoints.

Table 5 shows the BLEU scores when the ParaCrawl corpus is used. We obtain almost the same results for small transformer version while there is a small improvement of 0.46 BLEU point for the big model compared to the results reported in table 4 (without ParaCrawl).

de → fr	dev (BLEU)
1. Small Transformer (x1)	25.39
+Ensemble (x2)	25.81
+Ensemble (x5)	**25.92**
2. Big Transformer (x1)	26.27
+Ensemble (x2)	27.04
+Ensemble (x5)*	**27.61**

Table 4: BLEU results for DE→FR NMT systems using all training data but ParaCrawl corpus.

de → fr	dev (BLEU)
1. Small Transformer (x1)	25.18
+Ensemble (x2)	25.59
+Ensemble (x5)	**25.93**
2. Big Transformer (x1)	26.83
+Ensemble (x2)	27.80
+Ensemble (x5)	**28.07**

Table 5: BLEU results for DE →FR NMT systems with all training data including ParaCrawl.

Table 6 contains our results for WMT2019 training data with back-translation[4]. As expected, adding back-translations improves the results for both configurations: an increase of about 1% BLEU point is observed for small and big transformer models compared to the same systems without back-translation (see systems labeled "+Ensemble (x5)" in Table 4).

de → fr	dev (BLEU)
1. Small Transformer (x1)	26.64
+Ensemble (x2)	26.95
+Ensemble (x5)	**26.99**
2. Big Transformer (x1)	27.65
+Ensemble (x2)	28.40
+Ensemble (x5)	**28.63**

Table 6: BLEU results for DE →FR NMT systems with back-translation training data and without ParaCrawl parallel data.

[4]The FR→DE back-translations have been created using the small transformer (x1) system from table 7

Asterisk (*) in Table 4 marks our submitted model for German to French official evaluation. This model obtains a BLEU score of **33.4**. Our best system with back-translation was also submitted after the evaluation deadline and obtain a BLEU score of **34.6**.

4.2 French to German

We performed the same set of experiments as German to French. Table 7 shows the BLEU scores when NMT systems are trained without the ParaCrawl corpus. Unlike the German to French direction, only a small improvement is observed by using the big transformer architecture compared to the small one (21.18 with big model and 21.08 for small model).

fr → de	dev (BLEU)
1. Small Transformer (x1)	20.28
+Ensemble (x2)	20.73
+Ensemble (x5)	**21.09**
2. Big Transformer (x1)	20.42
+Ensemble (x2)	21.03
+Ensemble (x5)	**21.18**

Table 7: Results in terms of BLEU for FR →DE NMT systems using all the available training data except the ParaCrawl corpus.

As for the DE→Fr direction, we also trained systems by adding ParaCrawl data and results are presented in Table 9. As was formerly the case with DE→Fr, no improvement is observed by adding the Paracrawl corpus to the small transformer model. The model works less well than without Paracrawl and a drop of 0.4% BLEU points is observed when we compare the "+Ensemble (x5)" of small transformer models from tables 7 and 8. For the big transformer model there is an improvement of 0.76 BLEU point when the Paracrawl corpus is included in the training data.

Table 9 presents the results when the training set is extended with back-translated data[5]. Results shows a consistent improvement with back-translated data. We note an improvement of 0.4 BLEU points in comparison with the best small and big transformer models without back-translation. Asterisk (*) in Table 9 marks our submitted model for French to German official evaluation.

[5]The DE→FR back-translations have been created using the small transformer (x1) system from Table 4

fr → de	dev (BLEU)
1. Small Transformer (x1)	20.15
+Ensemble (x2)	20.29
+Ensemble (x5)	**20.65**
2. Big Transformer (x1)	21.37
+Ensemble (x2)	21.80
+Ensemble (x5)	**21.94**

Table 8: Results in terms of BLEU for FR →DE NMT systems using all the available training data including ParaCrawl corpus.

fr → de	dev (BLEU)
1. Small Transformer (x1)	21.15
+Ensemble (x2)	21.45
+Ensemble (x5)	**21.50**
2. Big Transformer (x1)	21.82
+Ensemble (x2)*	22.03
+Ensemble (x5)	**22.34**

Table 9: Results in terms of BLEU for the FR→DE NMT systems with back-translation training data but without ParaCrawl parallel data.

5 Conclusion

In this paper, we presented the LIUM participation to the WMT2019 news translation shared task. This year we have added for the first time the French-German language pair to the WMT news translation task. The parallel training data were created by cross-matching the EN-FR and EN-DE training data from previous WMT shared tasks. The LIUM has participated in the German ↔ French translation task with an ensemble of neural machine translation models based on the Transformer architecture. Our models were trained using a cleaned subset of the provided training dataset, and synthetic parallel data generated from the provided monolingual corpora.

Acknowledgments

We thank Franck Burlot and LinguaCustodia for translating part of the DE→FR test set. This work was supported by the French National Research Agency (ANR) through the CHIST-ERA M2CR project[6], under the contract number ANR-15-CHR2-0006-01.

[6]http://m2cr.univ-lemans.fr

References

Anne Baillot, Loïc Barrault, and Fethi Bougares. 2019. Cat tools in dh training. In *Proceedings of the 2019 Digital Humanities Conference*, Utrecht, The Netherlands. Poster.

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*, ACL '07, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*, pages 1–9. Association for Computational Linguistics.

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*, ACL '02, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *Prague Bull. Math. Linguistics*, 100:73–82.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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. Curran Associates, Inc.