Character-Level Translation with Self-attention

Yingqiang Gao[†‡], **Nikola I. Nikolov**[‡], **Yuhuang Hu**[‡], **Richard H.R. Hahnloser**[‡]
[†]Department of Informatics, Technical University of Munich
[‡]Institute of Neuroinformatics, University of Zurich and ETH Zurich
yingqiang.gao@in.tum.de {niniko, yuhuang.hu, rich}@ini.ethz.ch

Abstract

We explore the suitability of self-attention models for character-level neural machine translation. We test the standard transformer model, as well as a novel variant in which the encoder block combines information from nearby characters using convolutions. We perform extensive experiments on WMT and UN datasets, testing both bilingual and multilingual translation to English using up to three input languages (French, Spanish, and Chinese). Our transformer variant consistently outperforms the standard transformer at the character-level and converges faster while learning more robust character-level alignments.[1]

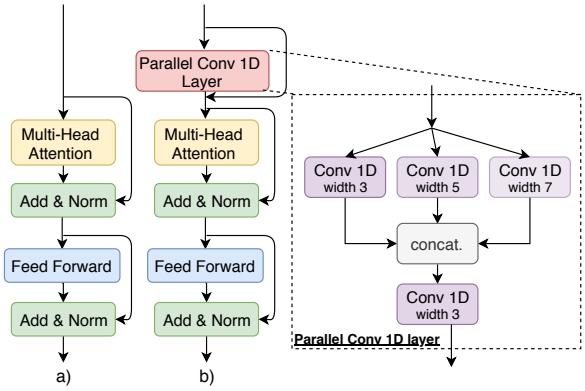


Figure 1: A comparison of the encoder blocks in the standard transformer (a) and our novel modification, the convtransformer (b), which uses 1D convolutions to facilitate character interactions.

1 Introduction

Most existing Neural Machine Translation (NMT) models operate on the word or subword-level, which tends to make these models memory inefficient because of large vocabulary sizes. Character-level models (Lee et al., 2017; Cherry et al., 2018) instead work directly on raw characters, resulting in a more compact language representation, while mitigating out-of-vocabulary (OOV) problems (Luong and Manning, 2016). Character-level models are also very suitable for multilingual translation since multiple languages can be modeled using the same character vocabulary. Multilingual training can lead to improvements in the overall performance without an increase in model complexity (Lee et al., 2017), while also circumventing the need to train separate models for each language pair.

Models based on self-attention have achieved excellent performance on a number of tasks, including machine translation (Vaswani et al., 2017) and representation learning (Devlin et al., 2019; Yang et al., 2019). Despite the success of these models, their suitability for character-level translation remains largely unexplored, with most efforts having focused on recurrent models (*e.g.*, Lee et al. (2017); Cherry et al. (2018)).

In this work, we perform an in-depth investigation of the suitability of self-attention models for character-level translation. We consider two models: the standard transformer from Vaswani et al. (2017) and a novel variant that we call the *convtransformer* (Figure 1, Section 3). The convtransformer uses convolutions to facilitate interactions among nearby character representations.

We evaluate these models on both bilingual and multilingual translation to English, using up to three input languages: French (FR), Spanish (ES), and Chinese (ZH). We compare the performance when translating from close (*e.g.*, FR and ES) and on distant (*e.g.*, FR and ZH) input languages (Section 5.1) and we analyze the learned character alignments (Section 5.2). We find that self-attention models work surprisingly well for character-level translation, achieving competitive

[1]Code available at https://github.com/CharizardAcademy/convtransformer

performance to equivalent subword-level models while requiring up to 60% fewer parameters (under the same model configuration). At the character-level, the convtransformer outperforms the standard transformer, converges faster, and produces more robust alignments.

2 Background

2.1 Character-level NMT

Fully character-level translation was first tackled in Lee et al. (2017), who proposed a recurrent encoder-decoder model. Their encoder combines convolutional layers with max-pooling and highway layers to construct intermediate representations of segments of nearby characters. Their decoder network autoregressively generates the output translation one character at a time, utilizing attention on the encoded representations.

Lee et al. (2017)'s approach showed promising results on *multilingual translation* in particular. Without any architectural modifications or changes to the character vocabularies, training on multiple source languages yielded performance improvements while also acting as a regularizer. Multilingual training of character-level models is possible not only for languages that have almost identical character vocabularies, such as French and Spanish, but even for distant languages that can be mapped to a common character-level vocabulary, for example, through latinizing Russian (Lee et al., 2017) or Chinese (Nikolov et al., 2018).

More recently, (Cherry et al., 2018) performed an in-depth comparison between different character- and subword-level models. They showed that, given sufficient computational time and model capacity, character-level models can outperform subword-level models, due to their greater flexibility in processing and segmenting the input and output sequences.

2.2 The Transformer

The transformer (Vaswani et al., 2017) is an attention-driven encoder-decoder model that has achieved state-of-the-art performance on a number of sequence modeling tasks in NLP. Instead of using recurrence, the transformer uses only feedforward layers based on self-attention. The standard transformer architecture consists of six stacked encoder layers that process the input using self-attention and six decoder layers that autoregressively generate the output sequence.

The original transformer (Vaswani et al., 2017) computes a scaled dot-product attention by taking as input query Q, key K, and value V matrices:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where $\sqrt{d_k}$ is a scaling factor. For the encoder, Q, K and V are equivalent, thus, given an input sequence with length N, Attention performs N^2 comparisons, relating each word position with the rest of the words in the input sequence. In practice, Q, K, and V are projected into different representation subspaces (called heads), to perform Multi-Head Attention, with each head learning different word relations, some of which might be interpretable (Vaswani et al., 2017; Voita et al., 2019).

Intuitively, attention as an operation might not be as meaningful for encoding individual characters as it is for words, because individual character representations might provide limited semantic information for learning meaningful relations on the sentence level. However, recent work on language modeling (Al-Rfou et al., 2019) has surprisingly shown that attention can be very effective for modeling characters, raising the question of how well the transformer would work on character-level bilingual and multilingual translation, and what architectures would be suitable for this task. These are the questions this paper sets out to investigate.

3 Convolutional Transformer

To facilitate character-level interactions in the transformer, we propose a modification of the standard architecture, which we call the *convtransformer*. In this architecture, we use the same decoder as the standard transformer, but we adapt each encoder block to include an additional sub-block. The sub-block (Figure 1, b), inspired from Lee et al. (2017), is applied to the input representations M, before applying self-attention. The sub-block consists of three 1D convolutional layers, C_w, with different context window sizes w. In order to maintain the temporal resolution of convolutions, the padding is set to $\lfloor \frac{w-1}{2} \rfloor$.

We apply three separate convolutional layers, C_3, C_5 and C_7, in parallel, using context window sizes of 3, 5 and 7, respectively. The different context window sizes aim to resemble character-level interactions of different levels of granularity, such as on the subword- or word-level. To

compute the final output of the convolutional sub-block, the outputs of the three layers are concatenated and passed through an additional 1D convolutional layer with context window size 3, C_3', which fuses the representations:

$$\text{Conv}(M) = M + C_3'(\text{Concat}(C_3(M), C_5(M), C_7(M))).$$

For all convolutional layers, we set the number of filters to be equal to the embedding dimension size d_{model}, which results in an output of equal dimension as the input M. Therefore, in contrast to Lee et al. (2017), who use max-pooling to compress the input character sequence into segments of characters, here we leave the resolution unchanged, for both transformer and convtransformer models. Finally, for additional flexibility, we add a residual connection (He et al., 2016) from the input to the output of the convolutional block.

4 Experimental Set-up

Datasets. We conduct experiments on two datasets. First, we use the **WMT15 DE→EN** dataset, on which we test different model configurations and compare our results to previous work on character-level translation. We follow the preprocessing in Lee et al. (2017) and use the newstest-2014 dataset for testing. Second, we conduct our main experiments using the **United Nations Parallel Corporus** (UN) (Ziemski et al., 2016), for two reasons: (i) UN contains a large number of parallel sentences from six languages, allowing us to conduct multilingual experiments; (ii) all sentences in the corpus are from the same domain. We construct our training corpora by randomly sampling one million sentence pairs from the FR, ES, and ZH parts of the UN dataset, targeting translation to English. To construct multilingual datasets, we combine the respective bilingual datasets (*e.g.*, FR→EN, and ES→EN) and shuffle them. To ensure all languages share the same character vocabulary, we latinize the Chinese dataset using the Wubi encoding method, following (Nikolov et al., 2018). For testing, we use the original UN test sets provided for each pair.

Tasks. Our experiments are designed as follows: (i) bilingual scenario, in which we train a model with a single input language; (ii) multilingual scenario, in which we input two or three languages at the same time without providing any language identifiers to the models and without increasing the number of parameters. We test combining input languages that can be considered as more similar in terms of syntax and vocabulary (*e.g.* FR and ES) as well as more distant (*e.g.*, ES and ZH).

	Model	BLEU	#par
character-level	Lee et al. (2017)	25.77	69M
	transformer-6-layer	28.8	49M
	convtransformer-6-layer	29.23	68M
	transformer-12-layer	29.81	93M
	convtransformer-12-layer	30.16	131M
bpe	transformer-6-layer	30.06	121M
	transformer-12-layer	31.60	165M

Table 1: Comparison of architecture variants on the WMT15 DE→EN dataset. **#par** is the number of model parameters.

5 Results

5.1 Automatic evaluation

Model comparison. In Table 1, we compare the BLEU performance (Papineni et al., 2002) of diverse character-level architectures trained on the WMT dataset. For reference, we include the recurrent character-level model from Lee et al. (2017), as well as transformers trained on the subword level using a vocabulary of 50k byte-pair encoding (BPE) tokens (Sennrich et al., 2016). All models were trained on four Nvidia GTX 1080X GPUs for 20 epochs.

We find character-level training to be 3 to 5 times slower than subword-level training due to much longer sequence lengths. However, the standard transformer trained at the character level already achieves very good performance, outperforming the recurrent model from Lee et al. (2017). On this dataset, our convtransformer variant performs on par with the character-level transformer. Character-level transformers also perform competitively with equivalent BPE models while requiring up to 60% fewer parameters. Furthermore, our 12-layer convtransformer model matches the performance of the 6-layer BPE transformer, which has a comparable number of parameters.

Multilingual experiments. In Table 2, we report our BLEU results on the UN dataset using the 6-layer transformer/convtransformer character-level models (Appendix A contains example model outputs). All of our models were trained for 30 epochs. Multilingual models are

	Model Input lang.	#P	transformer t-FR	transformer t-ES	transformer t-ZH	convtransformer t-FR	convtransformer t-ES	convtransformer t-ZH
bilingual	*FR*	1M	32.48	-	-	33.69	-	-
	ES	1M	-	39.90	-	-	41.41	-
	ZH	1M	-	-	38.70	-	-	**41.01**
multilingual	*FR+ES*	2M	33.51	40.83	-	**34.69**	**41.84**	-
	FR+ZH	2M	32.89	-	37.92	33.98	-	40.56
	ES+ZH	2M	-	40.43	38.23	-	41.49	40.41
	FR+ES+ZH	3M	33.69	40.71	38.01	34.38	41.73	39.87

Table 2: BLEU scores on the UN dataset, for different input training languages (first column), and evaluated on three different test sets (t-FR, t-ES and t-ZH). The target language is always English. **#P** is the number of training pairs. The best overall results for each language are in bold.

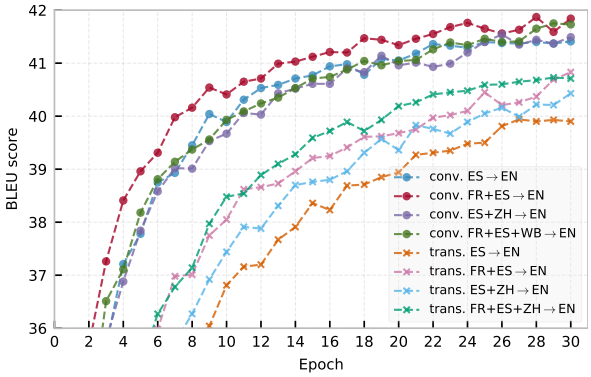


Figure 2: BLEU scores on the UN dataset as a function of epoch number, for bilingual and multilingual character-level translation from ES to EN. `conv.` is the convtransformer, while `trans.` is the original transformer.

evaluated on translation from all possible input languages to English.

Although multilingual translation can be realized using subword-level models through extracting a joint segmentation for all input languages (e.g., as in Firat et al. (2016); Johnson et al. (2017)), here we do not include any subword-level multilingual baselines, for two reasons. First, extracting a good multilingual segmentation is much more challenging for our choice of input languages, which includes distant languages such as Chinese and Spanish. Second, as discussed previously, subword-level models have a much larger number of parameters, making a balanced comparison with character-level models difficult.

The convtransformer consistently outperforms the character-level transformer on this dataset, with a gap of up to 2.3 BLEU on bilingual translation (ZH→EN) and up to 2.6 BLEU on multilingual translation (FR+ZH→EN). Training multilingual models on similar input languages (FR + ES→EN) leads to improved performance for both languages, which is consistent with (Lee et al., 2017). Training on distant languages is surprisingly still effective in some cases. For example, the models trained on FR+ZH→EN outperform the models trained just on FR→EN; however they perform worse than the bilingual models trained on ZH→EN. Thus, distant-language training seems to be helpful mainly when the input language is closer to the target translation language (which is English here).

The convtransformer is about 30% slower to train than the transformer (see Figure 2). Nevertheless, the convtransformer reaches comparable performance in less than half of the number of epochs, leading to an overall training speedup compared to the transformer.

5.2 Analysis of Learned Alignments

To gain a better understanding of the multilingual models, we analyze their learned character alignments as inferred from the model attention probabilities. For each input language (*e.g.*, FR), we compare the alignments learned by each of our multilingual models (*e.g.*, FR + ES → EN model) to the alignments learned by the corresponding bilingual model (*e.g.*, FR → EN). Our intuition is that the bilingual models have the greatest flexibility to learn high-quality alignments because they are not distracted by other input languages. Multilingual models, by contrast, might learn lower quality alignments because either (i) the architecture is not robust enough for multilingual training; or (ii) the languages are too dissimilar to allow for effective joint training, prompting the model to learn alternative alignment strategies to accommodate for all languages.

We quantify the alignments using canonical correlation analysis (CCA) (Morcos et al., 2018). First, we sample 500 random sentences from each of our UN testing datasets (FR, ES, or ZH) and then produce alignment matrices by extracting the encoder-decoder attention from the last layer of each model. We use CCA to project each alignment matrix to a common vector space and infer the correlation. We analyze our transformer and convtransformer models separately. Our results are in Figure 3, while Appendix B contains example alignment visualizations.

For similar source and target languages (*e.g.*, the FR+ES→EN model), we observe a strong pos-

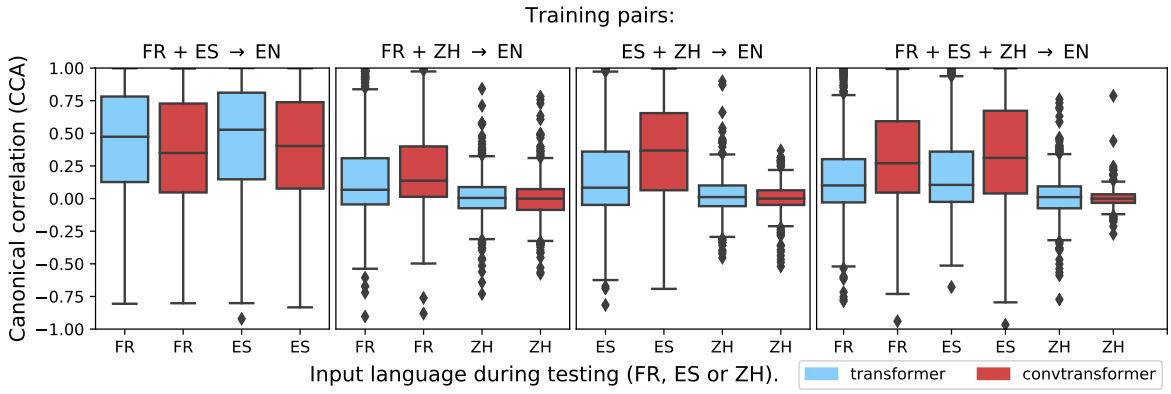


Figure 3: Canonical correlation between multilingual and bilingual translation models trained on the UN dataset.

itive correlation to the bilingual models, indicating that alignments can be simultaneously learned. When introducing a distant source language (ZH) in the training, we observe a drop in correlation, for FR and ES, and an even larger drop for ZH. This result is in line with our BLEU results from Section 5.1, suggesting that multilingual training on distant input languages is more challenging than multilingual training on similar input languages. The convtransformer is more robust to the introduction of a distant language than the transformer ($p < 0.005$ for FR and ES inputs, according to a one-way ANOVA test). Our results also suggest that more sophisticated attention architectures might need to be developed when training multilingual models on several distant input languages.

6 Conclusion

We performed a detailed investigation of the utility of self-attention models for character-level translation. We test the standard transformer architecture, as well as introduce a novel variant which augments the transformer encoder with convolutions, to facilitate information propagation across nearby characters. Our experiments show that self-attention performs very well on character-level translation, with character-level architectures performing competitively when compared to equivalent subword-level architectures while requiring fewer parameters. Training on multiple input languages is also effective and leads to improvements across all languages when the source and target languages are similar. When the languages are different, we observe a drop in performance, in particular for the distant language.

In future work, we will extend our analysis to include additional source and target languages from different language families, such as more Asian languages. We will also work towards improving the training efficiency of character-level models, which is one of their main bottlenecks, as well as towards improving their effectiveness in multilingual training.

Acknowledgements

We acknowledge support from the Swiss National Science Foundation (grant 31003A_156976) and the National Centre of Competence in Research (NCCR) Robotics. We also thank the anonymous reviewers for their useful comments.

References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level Language Modeling with Deeper Self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine

Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on Representational Similarity in Neural Networks with Canonical Correlation. In *Advances in Neural Information Processing Systems*, pages 5727–5736.

Nikola Nikolov, Yuhuang Hu, Mi Xue Tan, and Richard H.R. Hahnloser. 2018. Character-level Chinese-English Translation through ASCII Encoding. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 10–16, Belgium, Brussels. 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*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

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, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764. Curran Associates, Inc.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534.

A Example model outputs

Tables 3, 4, and 5 contain example translations produced by our different bilingual and multilingual models trained on the UN datasets.

B Visualization of Attention

In Figures 4,5, 6 and 7 we plot example alignments produced by our different bilingual and multilingual models trained on the UN datasets, always testing on translation from FR to EN. The alignments are produced by extracting the encoder-decoder attention of the last decoder layer of our transformer/convtransformer models.

We observe the following patterns: (i) for bilingual translation (Figure 4), the convtransformer has a sharper weight distribution on the matching characters and words than the transformer; (ii) for multilingual translation of close languages (FR+ES→EN, Figure 5), both transformer and convtransformer are able to preserve the word alignments, but the alignments produced by the convtransformer appear to be slightly less noisy; (iii) for multilingual translation of distant languages (FR+ZH→EN, Figure 6), the character alignments of the transformer become visually much noisier and concentrate on a few individual characters, with many word alignments dissolving. The convtransformer character alignments remain more spread out, and word align-

ment appears to be better preserved. This is another indication that the convtransformer is more robust for multilingual translation of distant languages. (iv) for multilingual translation with three inputs, where two of the three languages are close (FR+ES+ZH→EN, Figure 7), we observe a similar pattern, with word alignments being better preserved by the convtransformer.

source	Pour que ce cadre institutionnel soit efficace, il devra remédier aux lacunes en matière de réglementation et de mise en œuvre qui caractérisent à ce jour la gouvernance dans le domaine du développement durable.
reference	For this institutional framework to be effective, it will need to fill the regulatory and implementation deficit that has thus far characterized governance in the area of sustainable development.
FR→EN	
transformer	To ensure that this institutional framework is effective, it will need to address regulatory and implementation gaps that characterize governance in sustainable development.
convtransformer	In order to ensure that this institutional framework is effective, it will have to address regulatory and implementation gaps that characterize governance in the area of sustainable development.
FR+ES→EN	
transformer	To ensure that this institutional framework is effective, it will need to address gaps in regulatory and implementation that characterize governance in the area of sustainable development.
convtransformer	In order to ensure that this institutional framework is effective, it will be necessary to address regulatory and implementation gaps that characterize governance in sustainable development so far.
FR+WB→EN	
transformer	To ensure that this institutional framework is effective, gaps in regulatory and implementation that have characterized governance in sustainable development to date.
convtransformer	For this institutional framework to be effective, it will need to address gaps in regulatory and implementation that characterize governance in the area of sustainable development.
FR+ES+WB→EN	
transformer	To ensure that this institutional framework is effective, it will need to address regulatory and implementation gaps that are characterized by governance in the area of sustainable development.
convtransformer	If this institutional framework is to be effective, it will need to address gaps in regulatory and implementation that are characterized by governance in the area of sustainable development.

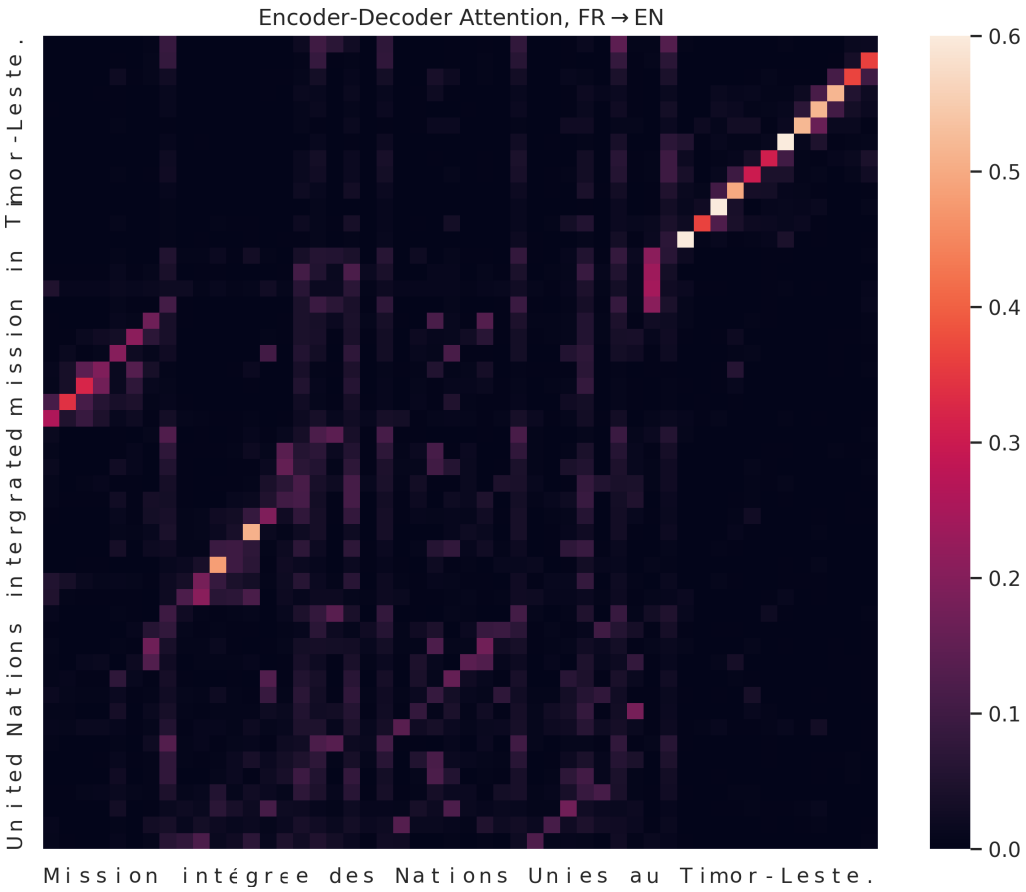
Table 3: Example character-level translation outputs on the UN dataset, FR→EN.

source	Estamos convencidos de que el futuro de la humanidad en condiciones de seguridad, la coexistencia pacífica, la tolerancia y la reconciliación entre las naciones se verán reforzados por el reconocimiento de los hechos del pasado.
reference	We strongly believe that the secure future of humanity, peaceful coexistence, tolerance and reconciliation between nations will be reinforced by the acknowledgement of the past.
ES→EN	
transformer	We are convinced that the future of humanity in conditions of security, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by recognition of the facts of the past.
convtransformer	We are convinced that the future of humanity under conditions of safe, peaceful coexistence, tolerance and reconciliation among nations will be reinforced by the recognition of the facts of the past.
FR+ES→EN	
transformer	We are convinced that the future of mankind under security, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by the recognition of the facts of the past.
convtransformer	We are convinced that the future of humanity in safety, peaceful coexistence, tolerance and reconciliation among nations will be reinforced by the recognition of the facts of the past.
ES+WB→EN	
transformer	We are convinced that the future of humanity in safety, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by the recognition of the facts of the past.
convtransformer	We are convinced that the future of humanity in safety, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by the recognition of the facts of the past.
FR+ES+WB→EN	
transformer	We are convinced that the future of mankind in safety, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by the recognition of the facts of the past.
convtransformer	We are convinced that the future of mankind in security, peaceful coexistence, tolerance and reconciliation among nations will be strengthened by the recognition of the facts of the past.

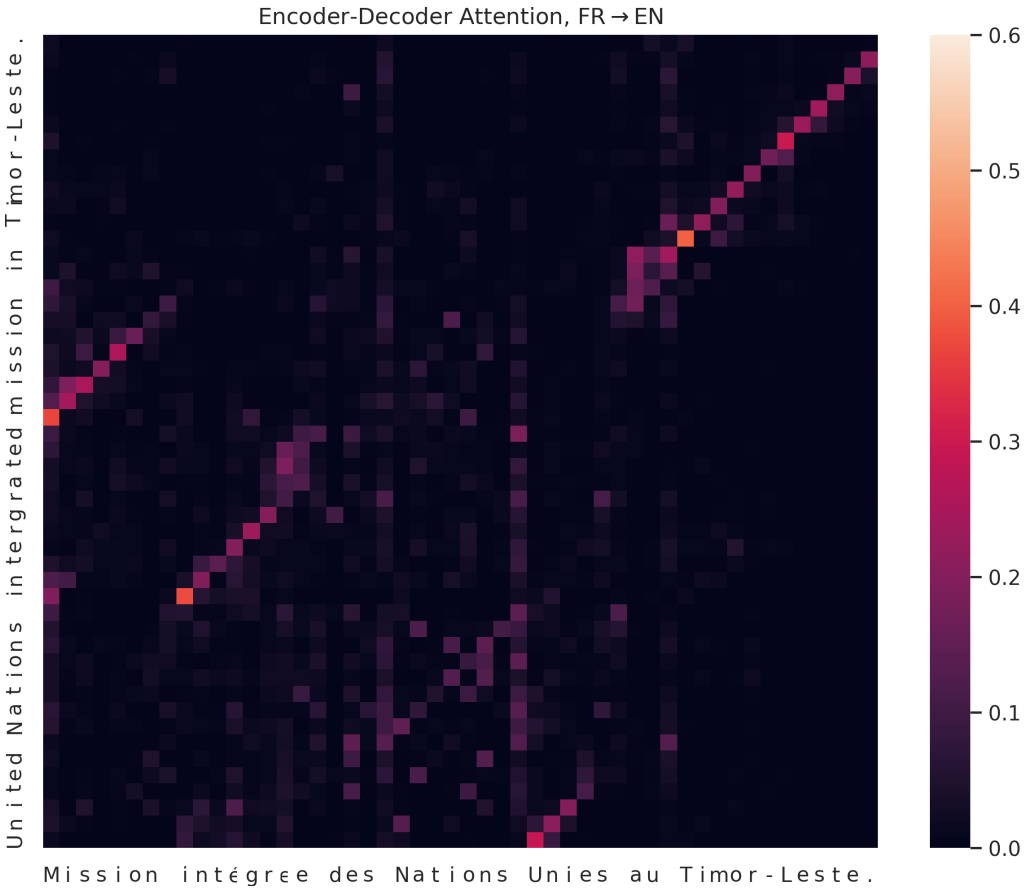
Table 4: Example character-level translation outputs on the UN dataset, ES→EN.

source ZH	利用专家管理农场对于最大限度提高生产率和灌溉水使用效率也是重要的。
source ZH	tjh|et fny|pe tp|gj pei|fnrt cf|gf jb|dd bv|ya rj|ym tg|u|yx t iak|ivc|ii wgkq0|et uqt|yx bn j tgj|s r .
reference EN	The use of expert farm management is also important to maximize land productivity and efficiency in the use of irrigation water.
ZH→EN	
transformer	The use of expert management farms is also important for maximizing productivity and irrigation use.
convtransformer	The use of experts to manage farms is also important for maximizing efficiency in productivity and irrigation water use.
FR+ZH→EN	
transformer	The use of expert management farms is also important for maximizing productivity and efficiency in irrigation water use.
convtransformer	The use of expert management farms is also important for maximizing productivity and irrigation water efficiency.
ES+ZH→EN	
transformer	The use of expert farm management is also important for maximizing productivity and irrigation water use efficiency.
convtransformer	The use of expert management farms to maximize efficiency in productivity and irrigation water use is also important.
FR+ES+ZH→EN	
transformer	The use of expert management farms is also important for maximizing productivity and irrigation water use.
convtransformer	It is also important that expert management farms be used to maximize efficiency in productivity and irrigation use.

Table 5: Example character-level translation outputs on the UN dataset, ZH→EN.

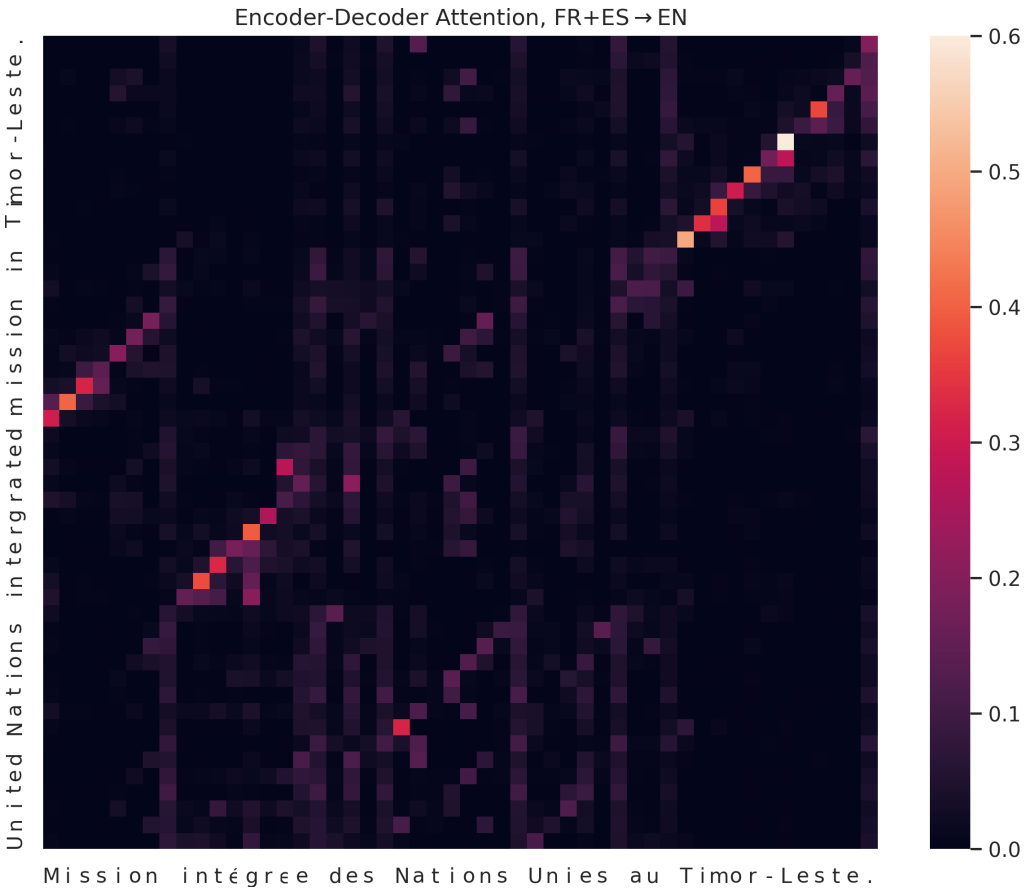


(a) transformer trained on FR→EN, testing with FR as input.

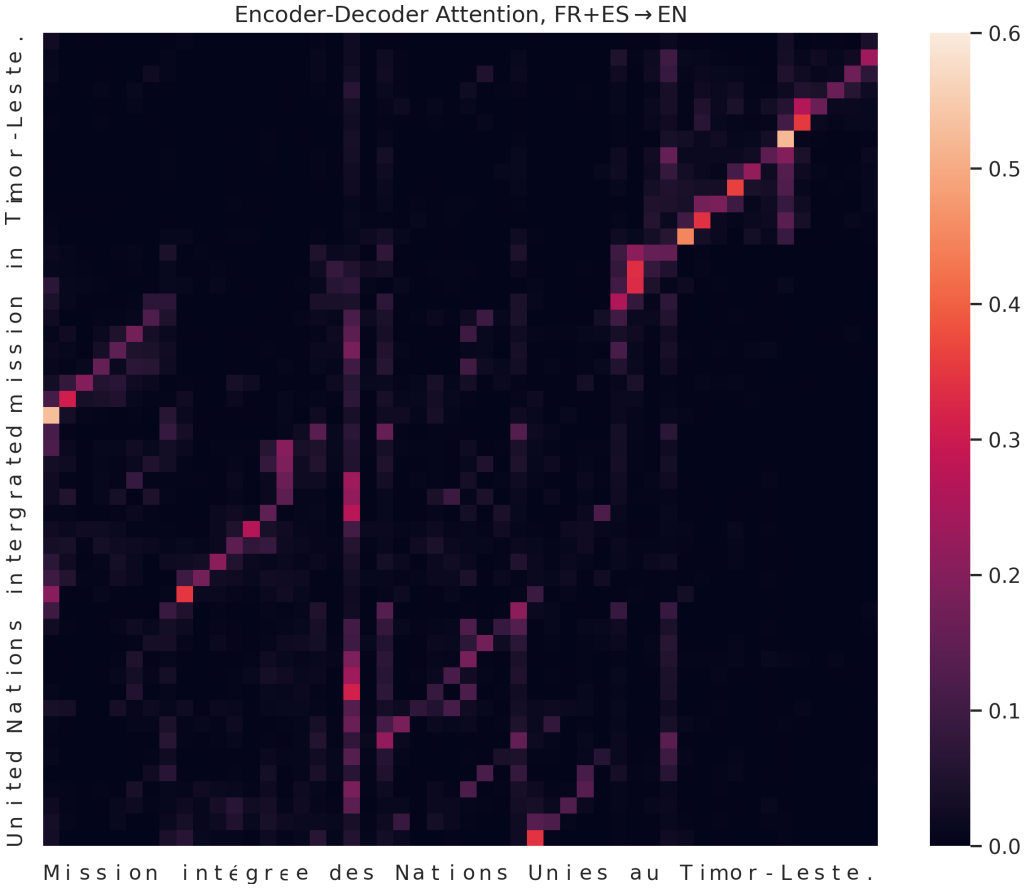


(b) convtransformer trained on FR→EN, testing with FR as input.

Figure 4: Example alignments produced by character-level models trained on FR→EN.

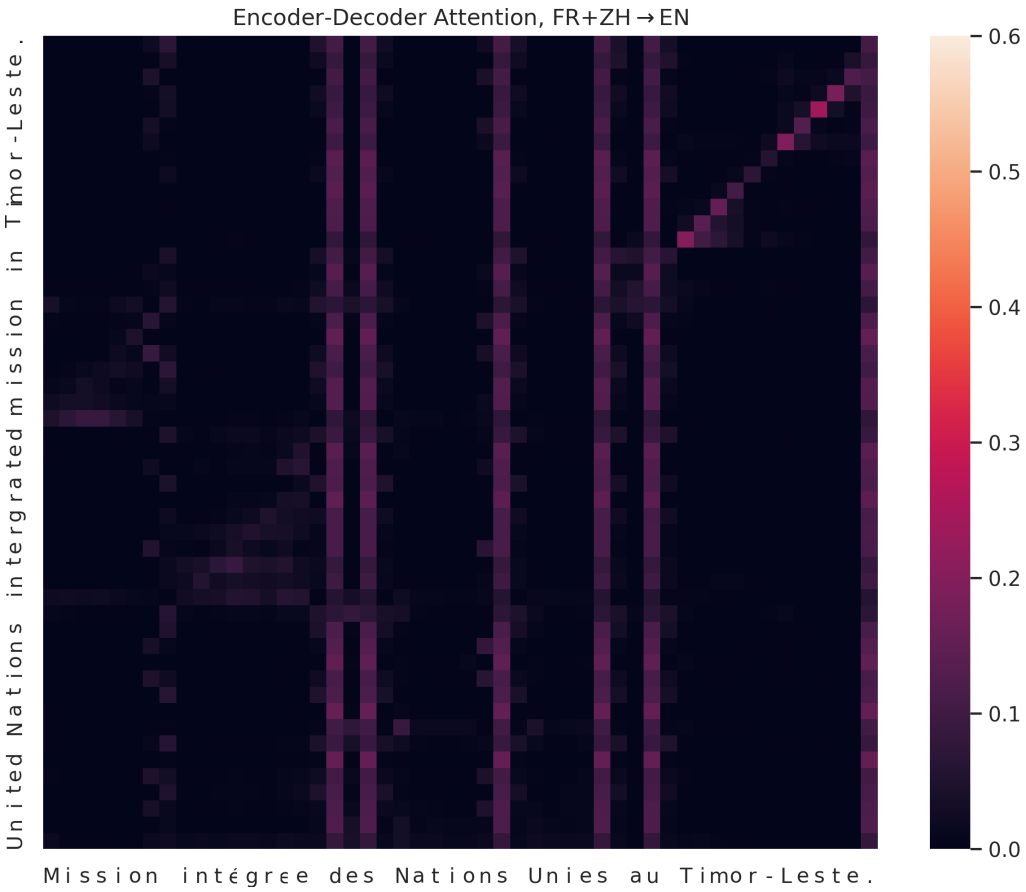


(a) transformer trained on FR+ES→EN, testing with FR as input.

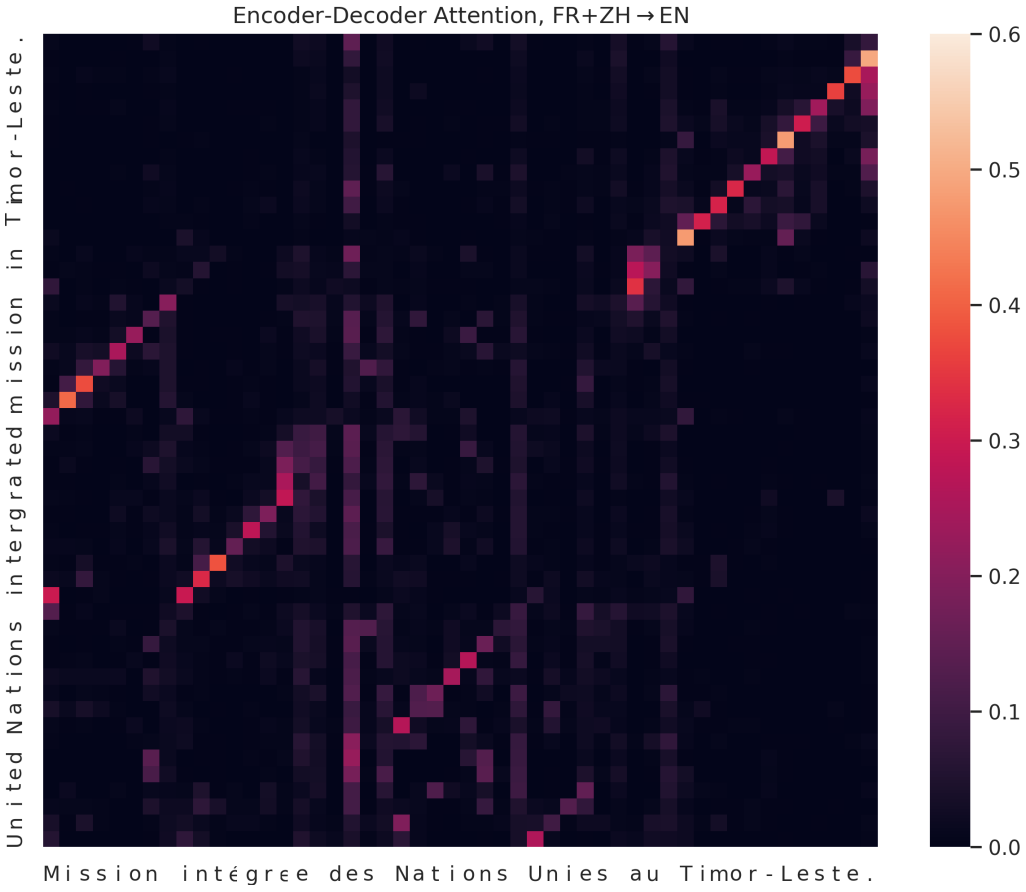


(b) convtransformer trained on FR+ES→EN, testing with FR as input

Figure 5: Example alignments produced by character-level models trained on FR+ES→EN.

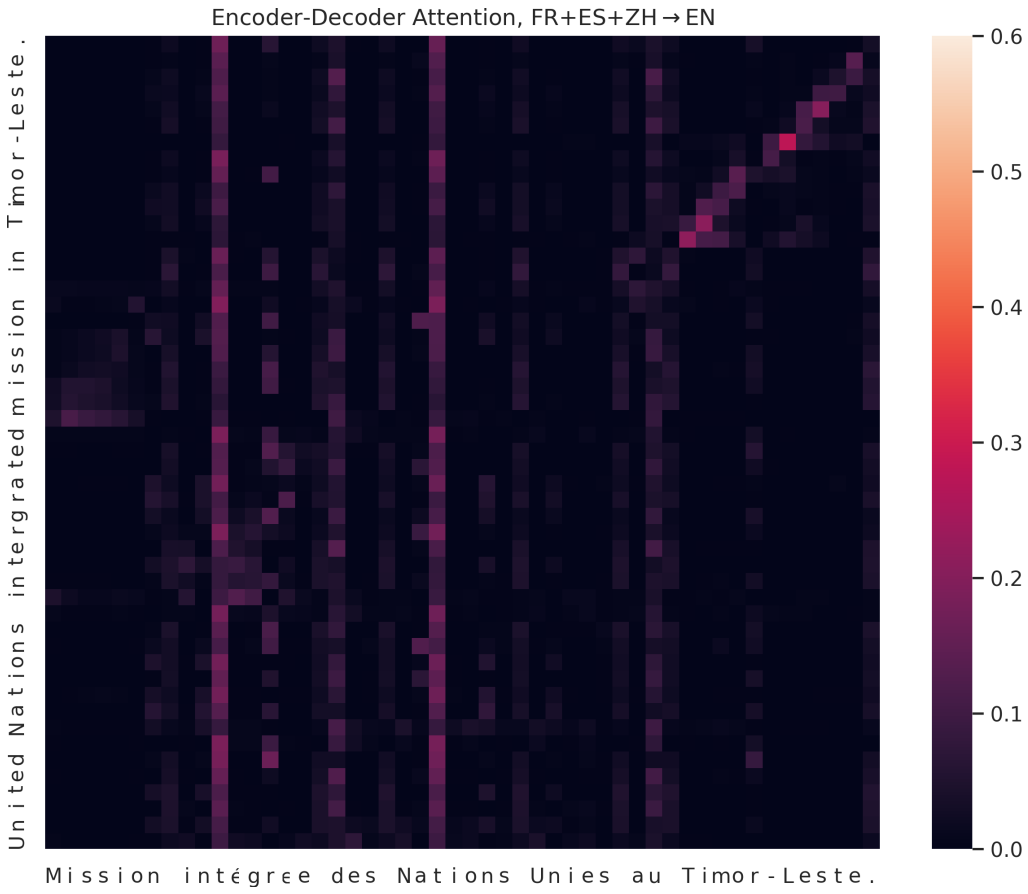


(a) transformer FR+ZH→EN, test on FR

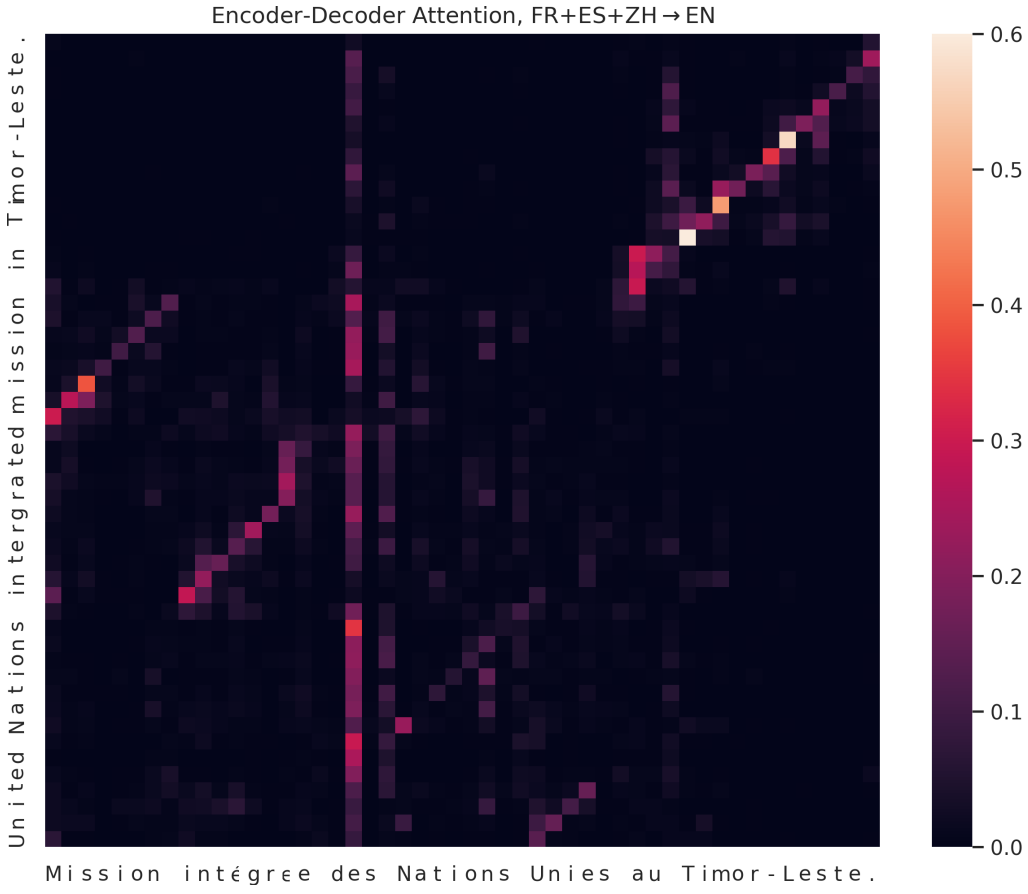


(b) convtransformer FR+ZH→EN, test on FR

Figure 6: Example alignments produced by character-level models trained on FR+ZH→EN.



(a) transformer FR+ES+ZH→EN, test on FR



(b) convtransformer FR+ES+ZH→EN, test on FR

Figure 7: Example alignments produced by character-level models trained on FR+ES+ZH→EN.