

I Have an Attention Bridge to Sell You: Generalization Capabilities of Modular Translation Architectures

Timothee Mickus

Raúl Vázquez

Joseph Attieh

University of Helsinki
firstname.lastname@helsinki.fi

Abstract

Modularity is a paradigm of machine translation with the potential of bringing forth models that are large at training time and small during inference. Within this field of study, modular approaches, and in particular attention bridges, have been argued to improve the generalization capabilities of models by fostering language-independent representations. In the present paper, we study whether modularity affects translation quality; as well as how well modular architectures generalize across different evaluation scenarios. For a given computational budget, we find non-modular architectures to be always comparable or preferable to all modular designs we study.

1 Introduction

Machine Translation (MT) has historically been under two influences that seem *a prima facie* contradictory. One of the goals of MT research is to provide means of converting sentences from any language to any other. On the one hand, generalization capabilities hinge on our systems producing language agnostic representations. On the other hand, MT models ought to be apt at encoding the specifics of source languages (Belinkov et al., 2017). The former of these trends has deeply marked this field—the concept of an ‘interlingua’ runs through most of the history of MT research, from Richens (1956) to Lu et al. (2018). The latter has recently motivated the development of modular approaches, where network parameters are specifically tied to a specific language.

How can we reconcile these two seemingly paradoxical trends? One promising approach is the inclusion of fully-shared subnetworks in modular architectures, and especially *bridge* components: They have been argued to foster language-independent representations (Zhu et al., 2020) as well as zero-shot generalization capabilities (Liao et al., 2021). Our aim is to carefully assess whether

modular architectures in general and bridges do indeed foster greater generalization capabilities.

We therefore study six architectures, five of which modular, with a particular focus on how they generalize—both to unseen translation directions, and to novel domains. We find that modular systems still struggle to remain competitive with fully-shared MT systems in scenarios when not all translation directions are available—a conclusion that affects systems with and without fixed-size bridges equally. While encoder-sharing modular designs can rival or outperform non-modular settings in a wide range of scenarios, all other systems we study struggle in zero-shot and out-of-distribution conditions, strongly questioning that fully-shared sub-networks in modular MT systems can improve their generalization capabilities.

2 Related Work

The full span of multilingual NMT (MNMT) architectures rely in the implicit assumption that the systems leverage the multilingual data by creating a shared encoding space via sharing: from fully-shared models (Johnson et al., 2017), to fully-modular systems, where sharing occurs only at dataset level (Escolano et al., 2021). In this work, we assess those two extreme cases, focusing in the modular NMT systems that incorporate some parameter-sharing bridging layers. Lu et al. (2018) introduced an attentional neural interlingua, which processes language-specific encoder embeddings to produce language-agnostic representations. Zhu et al. (2020) proposed a language-aware interlingua that transforms the encoder representation to a shared semantic space, showcasing practical means of fostering the semantic consistency of translations. Vázquez et al. (2019) integrated a shared inner-attention mechanism, referred to as “attention bridge”, based on the work of Lin et al. (2017), to generate fixed-size sentence representations. Fur-

ther studies by Raganato et al. (2019) and Vázquez et al. (2020), whose work we specifically build upon, emphasized the advantages of using multiple attention heads on the semantic quality of the translation—as well as challenges, particularly with translating longer sentences. Boggia et al. (2023) explored the effects of sharing encoder parameters vs. increasing the number of languages in modular MNMT. More recently, Purason and Tättar (2022) used layers shared by language groups to enhance translation, Mao et al. (2023) proposed a variable-length bridge that uses a classification layer to predict its length, and in Pires et al. (2023) the encoder is built with interspersed fully-shared and language-specific layers.

3 Experimental Methodology

3.1 Model Variants

All the models we consider are Transformer-based (Vaswani et al., 2017), and implemented with the MAMMOTH library (Mickus et al., 2024).¹ An overview of the different modular architectures we consider is displayed in Figure 1. We ensure that all datapoints are processed by the same number of encoder and decoder layers (6 and 6 resp.).

Non-modular baseline. To provide a reasonable point of comparison with existing approaches, we consider a simple non-modular architecture where all parameters are shared across all translation directions. We note these fully-shared models as \mathcal{F} .

Fully modular baseline. A second natural point of comparison is a modular system without bridge; e.g. Escolano et al. (2021). Such models, noted \mathcal{N} below, contain one 6-layer Transformer encoder and one 6-layer Transformer decoder per language, which are then selected for predictions depending on the desired language pair.

Semi-modular approaches. All other remaining architectures we will discuss contain both language specific and language-independent parameters. A simple means of achieving this consist in using a single shared encoder for all source languages (abbrv. \mathcal{E}), which would allow to leverage training signals from all source languages so as to provide more robust encoder representations. Conversely, one can consider employing a single shared decoder for all target languages (abbrv. \mathcal{D}) in the hopes of bolstering generation capabilities.

¹Configuration files available at github.com/Helsinki-NLP/mammoth/tree/main/examples/ab-neg/.

Bridges. We also consider models with a “bridge” layer, i.e., where all parameters are language specific aside from the last Transformer layer in the encoder. Such models have been explored by e.g. Boggia et al. (2023). These models are noted \mathcal{T} , and contain 5-layer language-specific Transformer encoders, followed by a shared Transformer layer serving as a bridge—i.e. they are \mathcal{N} -type modular systems where the parameters of the last layers of each encoder are tied.

Fixed-size attention bridges. An alternative proposed by Vázquez et al. (2020) consists in using fixed-size attention bridge (FSAB) designs. FSAB models, noted \mathcal{L} , resemble \mathcal{T} models except for the fact that the fully-shared Transformer layer bridge is replaced by the structured embedding architecture proposed by Lin et al. (2017):

$$\mathbf{Y} = \text{softmax} \left(\mathbf{W}_Q \text{ReLU}(\mathbf{W}_K \mathbf{X})^\top \right) \cdot \mathbf{X} \quad (1)$$

with \mathbf{X} the input matrix of the shared layer. Models of the \mathcal{L} architecture contain language-specific encoders comprising 5 Transformer layers, followed by one FSAB layer shared across all languages.

3.2 Datasets

We use two MT datasets: the United Nations Parallel Corpus (Ziemski et al., 2016, UNPC), which contains documents in six UN languages (Arabic, Chinese, English, French, Russian, and Spanish); and OPUS100 (Zhang et al., 2020), an English-centric multilingual corpus derived from Tiedemann (2012) spanning 100 languages. We ignore all OPUS translation directions not present in UNPC. Since the UNPC contains over 10M paired sentences across six languages (Arabic, English, Spanish, French, Russian, Mandarin Chinese), we consider the entire released data, rather than the fully aligned sub-corpus, and hold out 10% of the data for any evaluation and/or experiments. We ensure that sentences are unique to a split, i.e., if a pair of sentences (s_1, s_2) is present in the test split, then any pair (s_1, s_3) involving either of these sentence will also be assigned to the test split. Out of these 10%, we randomly select 25k sentences per language pairs to use as test sets. The remaining 90% examples are used for training, with 10k sentences per language pairs set aside for validation.

Test splits for generalization. We assess generalization capabilities in two common setups: zero-shot translation directions and out-of-distribution

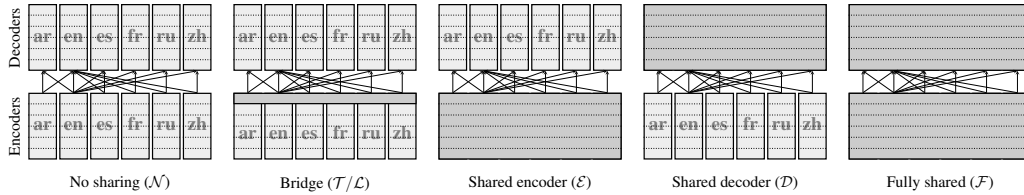


Figure 1: Overview of considered architectures, focusing on **EN** setting (using English as a pivot). Layers shaded in dark gray are shared across all languages; layers shaded in light gray are specific to a source or target language.

(OOD) examples. To evaluate out-of-distribution performances, we simply train models on one dataset (UNPC or OPUS) and evaluate it on the other (resp. OPUS or UNPC). Since bridge components are argued to be useful for unseen translation directions (Liao et al., 2021), we experiment with different language pivots to artificially create zero-shot translation directions. We construct three distinct UNPC training sets: (i) one using all 30 translation directions available in the UNPC, (“**All**”); (ii) one using all 10 directions involving English as a source or target (“**EN**”); and (iii) one using all 10 directions involving Arabic as a source or target (“**AR**”). This allows us to evaluate our models in both English-centric and non-English-centric contexts as well as in a zero-shot setting.² Hence, we refer to **EN** or **AR** being pivot languages, when an experiment is centered around that language.

Training conditions. To enable zero-shot translation (Vázquez et al., 2019; Artetxe and Schwenk, 2019, cf.), we train our models on auto-encoding tasks for all 6 languages. UNPC models are trained on monolingual data derived from the UNPC, and likewise OPUS models are trained on OPUS monolingual data. We train three seeds of all six model variants (\mathcal{F} , \mathcal{N} , \mathcal{E} , \mathcal{D} , \mathcal{T} , \mathcal{L}) on the four training sets (**UNPC-All**, **UNPC-EN**, **UNPC-AR**, **OPUS-EN**) under a strictly controlled computational budget: All models are exposed to the same number of datapoints and are trained with 6 AMD MI250X GPUs. We use the hyperparameters of Boggia et al. (2023) aside from batch accumulation, set to 8. We use $k = 50$ in \mathcal{L} models as Vázquez et al. (2020).

4 Results

The primary metric used for evaluating the performance of our models is BLEU (Papineni et al., 2002; Post, 2018).³ Results are shown in Table 1.

²Since OPUS100 is English-centric, only one variant of this dataset is considered for training.

³While COMET (Rei et al., 2020) would in principle be preferable, computing it for all translation directions in every

Choice of architecture. A clear trend emerges from our results: Across the board, the encoder-shared models \mathcal{E} are found to be the most successful, followed by the fully-shared, non-modular models \mathcal{F} . The latter only prevails upon the former in Arabic-centric scenario. At times, these architectures outrank other models considered by large margins of up to 7.5 BLEU points. While fully modular \mathcal{N} models or FSAB-based \mathcal{L} models perform well in the **EN**-centric scenario, these are not overwhelmingly better than \mathcal{F} .

Choice of pivot language. We experiment with different pivot languages, **EN** and **AR**, to understand their influence on the results. Our observations indicate that the choice of a pivot language can significantly impact the outcomes: The results with **AR** are always below the corresponding scores with **EN** on translation directions studied during training, whereas **AR** models yield generally higher performance in zero-shot conditions than their **EN** counterparts. Furthermore, we find tentative evidence that the behavior in **EN** and **AR** differs from that of **All**: In the latter case, we find a more limited impact of the architecture being used, with score varying at most by ± 4.2 BLEU points; whereas we observe a spread of up to ± 7.3 BLEU points for the former. As one would expect, being exposed to all translation directions during training (**All**) allows to improve performances averaged all translation directions. If we restrict ourselves to directions a model was exposed to during training, we find that **EN** models often outperform **All** models; whereas **AR** models are more in line with the values we see for **All**. This would suggest that there is a difficulty inherent to the translation directions considered; focusing only on directions that involve English may inflate performances.

Translation directions (seen vs. unseen). Expanding on what we already briefly touched on, we systematically find performances in zero-shot model in our study is prohibitively costly.

		Translation directions	\mathcal{N}	\mathcal{F}	\mathcal{E}	\mathcal{D}	\mathcal{T}	\mathcal{L}	
test on UNPC	train on UNPC	All (seen)	26.6 ± 0.5	28.2 ± 1.1	29.0 ± 0.2	24.8 ± 0.2	26.6 ± 0.1	26.4 ± 0.2	
		all	18.1 ± 0.3	24.6 ± 1.3	23.1 ± 1.7	15.7 ± 1.8	16.8 ± 0.4	17.9 ± 0.1	
		AR seen	26.4 ± 0.1	27.1 ± 0.9	26.6 ± 0.7	22.0 ± 1.0	24.9 ± 0.1	26.2 ± 0.0	
		unseen	13.9 ± 0.3	23.4 ± 1.4	21.4 ± 2.3	12.5 ± 2.2	12.8 ± 0.6	13.7 ± 0.1	
		all	19.3 ± 0.4	22.8 ± 2.9	23.9 ± 1.0	17.9 ± 0.4	19.3 ± 0.1	19.4 ± 0.1	
		EN seen	34.5 ± 0.2	33.1 ± 2.6	35.9 ± 0.1	31.6 ± 0.9	34.0 ± 0.2	34.6 ± 0.3	
		unseen	11.7 ± 0.6	17.6 ± 3.0	17.9 ± 1.4	11.0 ± 1.0	11.9 ± 0.1	11.8 ± 0.1	
		train on OPUS	all	16.4 ± 0.2	20.6 ± 0.5	20.9 ± 0.4	13.6 ± 1.2	16.8 ± 0.2	16.3 ± 0.1
		EN seen	30.8 ± 0.2	30.5 ± 0.5	31.1 ± 0.3	23.7 ± 1.5	30.7 ± 0.3	30.6 ± 0.3	
		unseen	9.1 ± 0.3	15.6 ± 0.5	15.8 ± 0.5	8.5 ± 1.0	9.9 ± 0.1	9.1 ± 0.1	
test on OPUS	train on UNPC	All (seen)	17.6 ± 0.2	19.1 ± 0.8	19.7 ± 0.2	16.3 ± 0.2	17.5 ± 0.2	17.5 ± 0.3	
		all	12.3 ± 0.2	16.5 ± 0.8	15.4 ± 1.0	10.3 ± 1.4	11.4 ± 0.2	12.1 ± 0.1	
		AR seen	17.7 ± 0.1	18.4 ± 0.8	17.9 ± 0.6	13.8 ± 1.0	16.6 ± 0.1	17.6 ± 0.1	
		unseen	9.2 ± 0.3	15.5 ± 0.8	13.9 ± 1.3	8.4 ± 1.6	8.5 ± 0.2	9.0 ± 0.1	
		all	13.4 ± 0.3	15.9 ± 2.0	17.0 ± 0.5	12.6 ± 0.2	13.3 ± 0.1	13.5 ± 0.2	
		EN seen	19.7 ± 0.1	19.8 ± 1.4	21.0 ± 0.0	18.5 ± 0.7	19.2 ± 0.1	19.8 ± 0.2	
		unseen	8.2 ± 0.3	12.7 ± 2.5	13.6 ± 0.9	7.7 ± 0.7	8.4 ± 0.3	8.2 ± 0.3	
		train on OPUS	all	15.0 ± 0.2	17.8 ± 0.3	17.9 ± 0.3	12.4 ± 1.0	15.5 ± 0.2	14.8 ± 0.1
		EN seen	25.1 ± 0.2	24.6 ± 0.3	25.1 ± 0.2	19.7 ± 1.6	24.9 ± 0.2	24.9 ± 0.3	
		unseen	6.6 ± 0.2	12.1 ± 0.3	11.8 ± 0.5	6.3 ± 0.7	7.7 ± 0.2	6.4 ± 0.1	

Table 1: Summary of performances, with **best** and **second best** values highlighted (avg. of 3 seeds ± std. dev.), and broken down according to whether the translation direction was seen during training or not (i.e., zero shot).

conditions to remain firmly below what we observe for translation directions observed during training. This holds across pivot languages and architectures. We do not observe that bridges (\mathcal{T} or \mathcal{L}) provide benefits in terms of zero-shot performances over fully modular systems (\mathcal{N}). Instead, it would appear that sharing the encoder (\mathcal{E} or \mathcal{F}) is beneficial—although it is uncertain that this is due to greater generalization capabilities rather than overall improved performances, the improvement brought about by \mathcal{E} and \mathcal{F} models is more substantiated in zero-shot settings (with a gap of at least 4.1 BLEU points in zero-shot settings, whereas \mathcal{F} can be outperformed by \mathcal{L} and/or \mathcal{N} for training directions).

In-distribution vs. out-of-distribution. Comparing performances in-distribution and out-of-distribution does not suggest that bridges meaningfully improve generalization capabilities. Performances of \mathcal{T} and \mathcal{L} models are in line with what we observe for the bridge-less \mathcal{N} models.

5 Statistical modeling

SHAP analysis & predictors importance Are our observations statistically significant? To establish which factors are at play, we rely on SHAP (Lundberg and Lee, 2017), a library and algorithm to derive heuristics for Shapley values (Shapley, 1953). We fit a gradient boosting decision tree

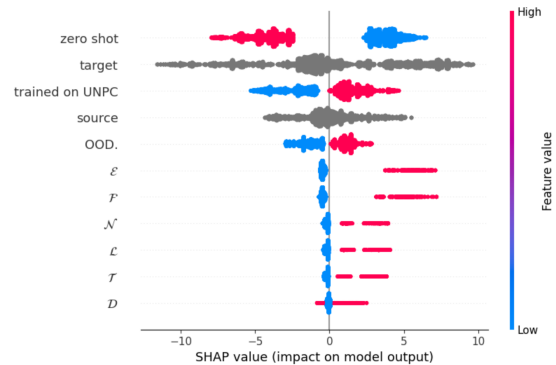


Figure 2: Overview of SHAP values, sorted by mean absolute value. Grey: categorical predictors; red: binary predictors where the value is true; blue, where it is false.

regression model with CatBoost (Prokhorenkova et al., 2018) to explain the BLEU scores obtained on specific language pairs and datasets by all the models we trained. We use as predictors (i) the source language (categorical); (ii) the target language (categorical), (iii) whether the model was trained on UNPC (binary); (iv) whether this translation direction in zero-shot (binary); (v) whether this test corresponds to an out-of-distribution setting (binary); as well as (vi–xi) which architecture is used (binary predicates for each of \mathcal{N} , \mathcal{F} , \mathcal{E} , \mathcal{D} , \mathcal{T} , and \mathcal{L}).

Figure 2 provides a general overview of the results of this analysis. The exact evaluation

	coef	std err	t	P> t
<i>Intercept</i>	11.8312	0.211	56.153	0.000
has bridge	4.4287	0.248	17.851	0.000
shares enc	5.3919	0.248	21.733	0.000
zero shot	-7.5567	0.139	-54.277	0.000
OOD	1.9472	0.140	13.917	0.000
from EN	1.9792	0.178	11.099	0.000
from ES	2.7001	0.204	13.235	0.000
from FR	1.5625	0.182	8.578	0.000
from RU	0.5932	0.182	3.257	0.001
from ZH	-2.6625	0.182	-14.617	0.000
to EN	9.1120	0.178	51.098	0.000
to ES	7.7651	0.204	38.061	0.000
to FR	4.0147	0.182	22.040	0.000
to RU	2.0937	0.182	11.494	0.000
to ZH	-5.2907	0.182	-29.046	0.000
trained on UNPC	4.3278	0.125	34.705	0.000
has bridge×zero shot	-4.7733	0.283	-16.876	0.000
has bridge×OOD	0.4169	0.283	1.472	0.141
shares enc×zero shot	0.3607	0.283	1.275	0.202
shares enc×OOD	0.9785	0.283	3.455	0.001

Table 2: OLS coefficients and significance. Intercept: \mathcal{N} -type, not OOD, not zero-shot, from & to AR.

conditions—i.e. the training and testing corpora and the specific language pairs seen at training and during the test at hand all, corresponding to predictors (i–v)—have a strong impact on the observed BLEU scores. We also see that using models of type \mathcal{F} and \mathcal{E} more strongly and more positively impacts the BLEU scores we observe than any other model type. In short, we find that most modular models fail to bring about results comparable with what we see for our non-modular baseline \mathcal{F} , with the sole exception of encoder-sharing \mathcal{E} .

OLS model & predictors interaction. Is there evidence that some modular architectures (and bridges in particular) enhance generalization capabilities? While SHAP values provide independent coefficients for each factor, this question is at its core one of interrelation—and is thus best studied through models able to capture potential interactions between predictors. To that end, we fit a simple ordinary least squares (OLS) linear model to predict the BLEU scores of our models using as predictors (i) whether the architecture contains a bridge (i.e., models of type \mathcal{T} or \mathcal{L}); (ii) whether it shares the encoder across source languages (i.e., models of type \mathcal{F} or \mathcal{E}); (iii) whether the model is tested in zero-shot; (iv) whether it is tested in an OOD setting; (v) whether the model was trained on UNPC; (vi & vii) the source and target languages; (viii–xi) the interactions between modular design (i.e., predictors i & ii) and performances in gener-

alization conditions (viz. predictors iii & iv).⁴

Our model achieves a R^2 of 0.763. Predictor coefficients and significance are listed in Table 2. As expected, modular design and training & test conditions (predictors i–vii) are always significant. Zero shot performances are linked to the strongest negative coefficient in our model; likewise, translating from or to ZH also turns out to degrade performance somewhat compared to the intercept (AR). Looking at interactions, we find that models with a bridge require a clear *negative* correction in zero-shot scenarios, *opposite* to what has been argued by Liao et al. (2021). Models of type \mathcal{F} and \mathcal{E} require a positive correction in OOD settings, suggesting they distinguish themselves further from other modular architectures. This statistical modeling suggests that bridge-based architectures significantly decrease generalization capabilities, as opposed to other modular (\mathcal{E}) and non-modular (\mathcal{F}) designs—in contrast with much of the discourse about their benefit for language independence and usefulness in zero-shot conditions (Raganato et al., 2019; Zhu et al., 2020; Vázquez et al., 2020).


6 Conclusions

In this work, we study the claim that bridge layers in modular architectures foster greater generalization capabilities. Given a carefully controlled computational budget, bridge architectures never clearly outperform bridge-less architectures, be they modular or not. In particular, we find non-modular architectures exhibit strong competitiveness, as they are only outperformed by modular architectures with language independent encoders and modular language-specific decoders. Additionally, we note that training conditions, such as the translation direction accessible to a model during training, have a significant impact.

These results suggest that current modular architectures, especially those using bridging layers, have limited potential insofar MT is concerned. In most cases, a default non-modular transformer fares better or just as well than the most effective modular system. Our study focused on modular architectures in a small-scale, well controlled experimental protocol; we leave questions such as whether these remarks carry on at a larger scale, both of model parameter counts and number of languages concerned, for future work.

⁴We ignore datapoints from type \mathcal{D} models since we are not aware of specific claims with respect to this architecture.

Acknowledgements

 This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113). We also thank the CSC-IT Center for Science Ltd., for computational resources and NVIDIA AI Technology Center (NVAITC) for the expertise in distributed training.

References

- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. [What do neural machine translation models learn about morphology?](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.
- Michele Boggia, Stig-Arne Grönroos, Niki Loppi, Timothee Mickus, Alessandro Raganato, Jörg Tiedemann, and Raúl Vázquez. 2023. [Dozens of translation directions or millions of shared parameters? comparing two types of multilinguality in modular machine translation](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 238–247, Tórshavn, Faroe Islands. University of Tartu Library.
- Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. [Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Junwei Liao, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2021. [Improving zero-shot neural machine translation on language-specific encoders- decoders](#). In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. [A structured self-attentive sentence embedding](#). In *International Conference on Learning Representations*.
- Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. [A neural interlingua for multilingual machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Zhuoyuan Mao, Haiyue Song, Raj Dabre, Chenhui Chu, and Sadao Kurohashi. 2023. [Variable-length neural interlingua representations for zero-shot neural machine translation](#).
- Timothee Mickus, Stig-Arne Grönroos, Joseph Attieh, Michele Boggia, Ona De Gibert, Shaoxiong Ji, Niki Andreas Loppi, Alessandro Raganato, Raúl Vázquez, and Jörg Tiedemann. 2024. [MAMMOTH: Massively multilingual modular open translation @ Helsinki](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 127–136, St. Julians, Malta. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Telmo Pessoa Pires, Robin M. Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. [Learning language-specific layers for multilingual machine translation](#).
- 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, Brussels, Belgium. Association for Computational Linguistics.
- Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. [Catboost: unbiased boosting with categorical features](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Taido Purason and Andre Tättar. 2022. [Multilingual neural machine translation with the right amount of sharing](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 91–100, Ghent, Belgium. European Association for Machine Translation.
- Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2019. [An evaluation of](#)

- language-agnostic inner-attention-based representations in machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepLANLP-2019)*, pages 27–32, Florence, Italy. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. **COMET: A neural framework for MT evaluation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Richard H. Richens. 1956. **Preprogramming for mechanical translation**. *Mechanical Translation*, 3(1):20–25.
- Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.
- Jörg Tiedemann. 2012. **Parallel data, tools and interfaces in OPUS**. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- 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*, volume 30. Curran Associates, Inc.
- Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann. 2020. **A systematic study of inner-attention-based sentence representations in multilingual neural machine translation**. *Computational Linguistics*, 46(2):387–424.
- Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. **Multilingual NMT with a language-independent attention bridge**. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepLANLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. **Improving massively multilingual neural machine translation and zero-shot translation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.
- Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. **Language-aware interlingua for multilingual neural machine translation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.
- 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’16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).