

DoDS-IITPKD: LoRA Fine-Tuning and LLM Post-Editing for Low-Resource Indic Machine Translation

G. L. John Salvin Amisha Ramesan Abigairl Nyasha Chigwededza*
Shrikant Tryambak Budde* Swapnil Hingmire

Mehta Family School of Data Science and Artificial Intelligence,
Indian Institute of Technology Palakkad, Kerala, India

{142503002, 142514002}@smaail.iitpkd.ac.in,
abbychigwededza@gmail.com, shreebudde24@gmail.com,
swapnilh@iitpkd.ac.in

Abstract

We describe the DoDS-IITPKD submissions to the WMT 2026 Shared Task on Low-Resource Indic Language Translation. We cover six English-centric pairs in both directions: Assamese, Mizo, Khasi, and Manipuri in Category 1, and Bodo and Kokborok in Category 2. Each system adapts a frozen multilingual backbone with low-rank adapters (LoRA with DoRA and rank-stabilised scaling). NLLB-200-3.3B handles Mizo, Khasi, and Kokborok, and IndicTrans2-1B handles Assamese, Manipuri, and Bodo. For the two languages outside the NLLB tokenizer (Khasi and Kokborok) we use same-script surrogate language tags. We add self back-translation for the Indic-English directions and, for some systems, a line-aligned LLM post-editing pass. On the official test sets, our systems obtain the best BLEU of any submission for Bodo to English (36.10) and English to Kokborok (7.36), and the best on-time BLEU for Kokborok to English (20.28). LLM post-editing gives large gains in the Indic-English direction (up to +8.20 BLEU) and almost none in the English-Indic direction.

1 Introduction

The WMT 2026 Low-Resource Indic Language Translation task asks for bidirectional English $\leftrightarrow$ X systems for several under-resourced languages of North-East India (Pakray et al., 28-29 October 2026). These languages differ in data volume, script, and pretrained-model coverage, so we route each language to the backbone that covers it best and apply one shared parameter-efficient recipe. NLLB-200-3.3B (NLLB Team et al., 2022) handles Mizo, Khasi, and Kokborok, while IndicTrans2-1B (Gala et al., 2023) handles Assamese, Manipuri, and Bodo. Both are adapted with rank-64 LoRA (Hu et al., 2021) using DoRA (Liu et al.,

2024) and rsLoRA (Kalajdzievski, 2023), optionally augmented with self back-translation and a line-aligned LLM post-editing pass.

This work builds on our WMT 2025 submission (Khongthaw et al., 2025), which fine-tuned distilled 600M and 200M models with rank-16 LoRA (Pakray et al., 2025). We scale both backbones, upgrade the adapters, handle the languages outside the NLLB tokenizer with same-script surrogate tags, and add LLM post-editing. On the official test sets our systems obtained the highest BLEU of any submission for Bodo to English and English to Kokborok, the highest on-time BLEU for Kokborok to English, and the second-highest for Manipuri (Bengali) to English. We release our code and system outputs.¹

2 Related Work

Massively multilingual models such as NLLB-200 (NLLB Team et al., 2022) and IndicTrans2 (Gala et al., 2023) provide strong starting points for low-resource pairs. Parameter-efficient methods adapt them cheaply: LoRA (Hu et al., 2021) trains low-rank updates, DoRA (Liu et al., 2024) stabilises them by decoupling magnitude and direction, and rsLoRA (Kalajdzievski, 2023) rescales for higher ranks. Back-translation (Sennrich et al., 2016) is the standard augmentation for low-resource NMT (Edunov et al., 2018). Small professionally translated corpora such as SMOL (Caswell et al., 2025) and bilingual lexica such as GATITOS (Jones et al., 2023) give outsized gains for the least-resourced languages. Prior editions of this task (Pal et al., 2023; Kakum et al., 2023; Pakray et al., 2024, 2025) show that fine-tuned pretrained models dominate and that systems using external data usually top the merged leaderboards.

* Work done at IIT Palakkad.

¹<https://github.com/John-salvin/wmt26-indicmt-dods-iitpkd>

3 Data

We enter six pairs in both directions: Assamese, Mizo, Khasi, and Manipuri (Bengali script) in Category 1, and Bodo and Kokborok in Category 2. All primary training uses the official WMT 2026 parallel data (Pakray et al., 28-29 October 2026), which continues the earlier editions of this task (Pal et al., 2023; Kakum et al., 2023; Pakray et al., 2024, 2025). Contrastive systems additionally use public corpora: BPCC-Human (AI4Bharat, 2023), Google SMOL (Caswell et al., 2025), and GATITOS (Jones et al., 2023) parallel data, IndicNECorp 1.0 (Pal et al., 2023) monolingual and parallel data, and Tatoeba (Tiedemann, 2020) English–Khasi. External data is filtered for empty fields, extreme length ratios, and overlap with the official train and test sets. Evaluation is on the organizers’ blind test sets.

4 System

4.1 Shared adaptation

Every direction keeps the base model frozen and trains low-rank adapters on the attention projections and, for the No Language Left Behind (NLLB) model, it trains low-rank adapters on the feed-forward projections. The adapted forward pass through a weight W_0 is

$$h = W_0x + \frac{\alpha}{\sqrt{r}} BAx, \quad (1)$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are the only trainable parameters. In this case d is the original hidden dimension of the model layer and r is the LoRA rank. The adapter learns a low-rank correction to the frozen pretrained weight rather than replacing it, enabling the model to specialize to each translation direction while updating only a small fraction of the parameters. We enable DoRA, which splits the update into magnitude and direction, and rsLoRA, which supplies the rank-stabilised $\alpha/\sqrt{r}$ scaling in Equation 1. Adapters are direction-specific, so we train many directions cheaply and can warm-start one direction from another. Training proceeds in up to three stages: supervised fine-tuning on official data (Stage 1), back-translation retraining (Stage 2), and an extended stage that adds external corpora for contrastive runs (Stage 3). Figure 1 shows the shared pipeline; Table 1 gives the language-tag assignment and Table 2 the hyperparameters.

Language	Script	Backbone / tag	Status
Mizo	Latin	NLLB lus_Latn	native
Khasi	Latin	NLLB lus_Latn	surrogate
Kokborok	Bengali	NLLB ben_Beng	surrogate
Assamese	Bengali	IT2 asm_Beng	native
Manipuri	Bengali	IT2 mni_Beng	native
Bodo	Devanagari	IT2 brx_Deva	native

Table 1: Language-tag assignment. IT2 denotes IndicTrans2-1B. English uses eng_Latn.

4.2 NLLB-200: Mizo, Khasi, Kokborok

The base is facebook/nllb-200-3.3B (NLLB Team et al., 2022), which requires each translation to be conditioned on a target-language token corresponding to a language in its FLORES-200 vocabulary. Of our three target languages, only Mizo has a native FLORES-200 language code. Khasi and Kokborok are absent from the tokenizer, so we condition the decoder on a same-script surrogate code, Mizo’s lus_Latn for Khasi and Bengali ben_Beng for Kokborok, and let the adapter learn the true target language from parallel data (Table 1). The surrogate seeds the correct script and lexical prior without adding tokenizer entries, unlike our WMT 2025 system, which registered a new token.

4.3 IndicTrans2: Assamese, Manipuri, Bodo

We use the IndicTrans2 1B distilled models (Gala et al., 2023), indictrans2-en-indic-1B for en→X and indictrans2-indic-en-1B for X→en. All three targets are native, so no surrogate is needed. Preprocessing uses the IndicProcessor from IndicTransToolkit. We restrict LoRA to the attention projections and leave the shared embedding and lm_head untouched, which we found necessary to avoid wrong-script output on low-resource directions.

4.4 Back-translation

The official corpora are small, so we use back-translation to generate additional synthetic parallel data which improves the model’s ability to generalize in low-resource settings. For the Indic-English directions we back-translate the target side of the official data with the reverse-direction adapter and retrain on the official data (upsampled) mixed with the synthetic pairs (Sennrich et al., 2016). We use plain beam-search back-translation, the more dependable choice at this data scale (Edunov et al., 2018). For the English-Indic directions back-

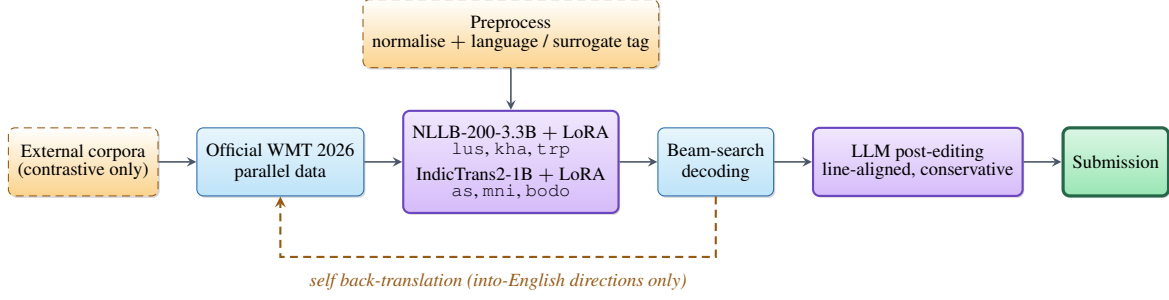


Figure 1: Shared adaptation pipeline. A frozen multilingual backbone is adapted with low-rank adapters (LoRA with DoRA and rsLoRA), optionally augmented with self back-translation and post-edited by an LLM. Each language is routed to the backbone that covers it best; all other stages are shared.

Setting	NLLB-200-3.3B	IndicTrans2-1B
rank r / α	64 / 16	64 / 32
DoRA / rsLoRA	yes / yes	yes / yes
dropout	0.05	0.05
target modules	attn + FFN	attn only
learning rate	2e-4, 1e-4	5e-5
epochs (S1/S2/S3)	5 / 2 / 3	5 / 2 / 3
schedule	warmup 0.03	cosine, LS 0.1

Table 2: LoRA hyper-parameters. Common to both: bf16, AdamW, effective batch 8×4 , max length 256, gradient checkpointing. Learning rates are Stage-1 then Stages 2–3 (NLLB).

translation did not help on improving the results on development set (we used the gold test data from WMT25 (Pakray et al., 2025) as the development set). So the primary systems were trained only on the official parallel data without using synthetic back-translated data.

4.5 LLM post-editing

The outputs with weak or gibberish translation were selected and for these selected runs, we post-edit outputs with Claude Opus 4.8 (Anthropic, 2026) under a strict line-aligned protocol: the model receives the English side and the system hypothesis and must return exactly one edited line per input, keeping the hypothesis unchanged when it cannot confidently improve it. The generation uses deterministic decoding (temperature=0) to ensure reproducibility. After each batch, we automatically verify that the number of output lines matches the input before merging the edited batches into the final submission. Post-edited runs are labelled *contrastive2* for the NLLB and *en*→*X* IndicTrans2 directions, and *contrastive* for the *X*→*en* IndicTrans2 directions.

Submissions. For each direction the primary uses official data with self back-translation, and up to two contrastive runs add external corpora and/or LLM post-editing.

5 Results

Table 3 reports the official evaluation of all our runs: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), chrF++ (Popović, 2015), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020), computed by the organizers on the blind test sets.

Our systems obtained the highest BLEU of any submission for Bodo to English (36.10) and English to Kokborok (7.36), and the highest on-time BLEU for Kokborok to English (20.28), ahead of the best on-time competitor (16.67) and far above the best WMT 2025 system in that direction (2.41 BLEU; Pakray et al., 2025). In Manipuri (Bengali) to English our contrastive run (40.84) ranked second only to the leading team, and our primary (38.84) was the second-best primary. Our primaries were also the second-best primary for English to Bodo (16.07) and among the top three for Assamese to English (24.37), English to Khasi (20.85), and Bodo to English (31.17).

The clearest finding is that LLM post-editing is strongly asymmetric, and the three metrics agree (Table 4) with this. This reflects the LLM’s much stronger generation capability in English than in the low-resource Indic target languages. In Indic-English it helped in every direction and on every metric, rising from +0.54 BLEU, +0.29 chrF++, and +0.09 COMET for Assamese to +8.20 BLEU, +11.82 chrF++, and +10.78 COMET for Kokborok. The character-level and neural metrics move at least as much as BLEU, so the gains are not an artefact of n-gram overlap. Out of English the ef-

Direction	System	BLEU↑	METEOR↑	TER↓	chrF++↑	BERTScore↑	COMET↑
en→as	primary	14.73	38.19	66.18	49.07	85.74	81.59
	contrastive	15.08	39.13	65.48	49.77	86.05	81.95
	contrastive2	14.73	38.20	66.16	49.07	85.74	81.60
as→en	primary	24.37	59.55	67.94	55.63	94.77	85.79
	contrastive	24.91	59.70	67.77	55.92	94.81	85.88
	constrained	24.21	59.44	68.19	55.50	94.76	85.77
en→lus	primary	18.75	45.36	69.19	48.56	81.47	64.74
	contrastive	18.23	45.09	68.89	47.98	81.56	65.01
	contrastive2	18.77	45.43	69.08	48.57	81.48	64.75
lus→en	primary	27.65	63.24	62.39	59.15	95.01	79.43
	contrastive	30.18	62.60	57.38	58.74	95.27	79.37
	contrastive2	31.60	65.29	55.72	60.80	95.72	81.53
en→kha	primary	20.85	44.57	63.08	44.04	79.54	65.58
	contrastive	18.99	44.27	58.60	43.30	80.53	62.31
	contrastive2	21.05	45.03	61.27	44.46	79.81	65.85
kha→en	primary	16.99	40.92	82.49	41.48	91.91	65.66
	contrastive	19.14	43.25	77.40	43.29	92.63	68.15
	contrastive2	24.63	55.01	66.92	51.83	94.55	76.93
en→mni	primary	8.65	18.43	86.47	39.63	84.67	67.61
	contrastive	5.23	18.45	86.74	39.60	84.72	67.78
	contrastive2	8.66	18.43	84.51	39.75	84.72	67.65
mni→en	primary	38.84	73.61	47.90	68.52	95.69	82.74
	contrastive	40.84	74.57	45.79	69.49	96.33	84.09
	constrained	37.75	73.49	49.01	68.00	95.75	82.77
en→bodo	primary	16.07	36.23	72.30	49.88	84.95	62.30
	contrastive	16.78	36.80	72.54	50.26	84.91	62.22
	contrastive2	16.33	36.52	71.95	49.95	84.95	62.33
bodo→en	primary	31.17	68.24	52.78	60.23	95.00	81.22
	contrastive	36.10	71.51	47.47	63.93	95.88	83.78
	constrained	31.41	67.26	52.90	60.00	95.02	81.39
en→trp	primary	3.86	17.44	95.34	27.33	71.19	58.99
	contrastive	7.36	28.40	81.52	35.04	75.14	60.72
	contrastive2	7.35	28.37	81.55	35.03	75.14	60.73
trp→en	primary	3.04	17.32	151.93	22.78	84.99	44.42
	contrastive	12.08	35.70	94.12	36.65	90.86	65.47
	contrastive2	20.28	50.42	75.11	48.47	93.67	76.25

Table 3: Official WMT 2026 test-set results for all DoDS-IITPKD runs, as released by the organizers. Bold marks the best BLEU per direction among our runs. *primary*: official data with self back-translation; *contrastive/contrastive2*: add external corpora and/or LLM post-editing; *constrained*: official data with back-translation, no external corpora or post-editing.

fect is negligible under all three metrics, at most 0.45 BLEU, 0.70 chrF++, and 0.35 COMET in absolute value. The editor is fluent in English and can repair adequacy and fluency errors when English is the target, but has little competence in the low-resource target languages, so under our conservative protocol it leaves those outputs essentially unchanged.

Back-translation was applied only to the Indic-English directions, where it helped; English-Indic primaries kept their official-only adapters. External corpora helped most where the official data was smallest: they raised English to Kokborok from

3.86 to 7.36 and Kokborok to English from 3.04 to 12.08 before post-editing, while adding little or regressing elsewhere, for example English to Manipuri (Bengali) fell from 8.65 to 5.23.

6 Conclusion

We described the DoDS-IITPKD submissions to the WMT 2026 Low-Resource Indic Language Translation task: one parameter-efficient recipe (rank-64 LoRA with DoRA and rsLoRA) over two frozen backbones and twelve directions, same-script surrogate tags for languages outside the NLLB tokenizer, direction-aware self back-

Direction	BLEU↑			chrF++↑			COMET↑		
	base	PE	Δ	base	PE	Δ	base	PE	Δ
as→en	24.37	24.91	+0.54	55.63	55.92	+0.29	85.79	85.88	+0.09
lus→en	30.18	31.60	+1.42	58.74	60.80	+2.06	79.37	81.53	+2.16
kha→en	19.14	24.63	+5.49	43.29	51.83	+8.54	68.15	76.93	+8.78
mni→en	38.84	40.84	+2.00	68.52	69.49	+0.97	82.74	84.09	+1.35
bodo→en	31.41	36.10	+4.69	60.00	63.93	+3.93	81.39	83.78	+2.39
trp→en	12.08	20.28	+8.20	36.65	48.47	+11.82	65.47	76.25	+10.78
en→as	15.08	14.73	-0.35	49.77	49.07	-0.70	81.95	81.60	-0.35
en→lus	18.75	18.77	+0.02	48.56	48.57	+0.01	64.74	64.75	+0.01
en→kha	20.85	21.05	+0.20	44.04	44.46	+0.42	65.58	65.85	+0.27
en→mni	8.65	8.66	+0.01	39.63	39.75	+0.12	67.61	67.65	+0.04
en→bodo	16.78	16.33	-0.45	50.26	49.95	-0.31	62.22	62.33	+0.11
en→trp	7.36	7.35	-0.01	35.04	35.03	-0.01	60.72	60.73	+0.01

Table 4: Effect of LLM post-editing (PE) on three metrics. *base* is the strongest non-post-edited DoDS-IITPKD run in that direction, selected by BLEU; its chrF++ and COMET are the official scores of that same run, so all three metrics compare an identical pair of systems. All values are official WMT 2026 test-set scores.

translation, and a line-aligned LLM post-editing pass. Our systems ranked first for Bodo to English and English to Kokborok, and first among on-time systems for Kokborok to English. The main practical lesson is that LLM post-editing is a cheap, training-free layer that helps most when translating into a language the editor knows well, and that self back-translation should be applied per direction rather than per pair.

Acknowledgments

We thank the organizers of the WMT 2026 Shared Task on Low-Resource Indic Language Translation for the datasets and evaluation, and the Department of Data Science, IIT Palakkad, for computing resources on the Madhava HPC cluster.

References

- AI4Bharat. 2023. Bharat Parallel Corpus Collection (BPCC). <https://huggingface.co/datasets/ai4bharat/BPCC>. Released with IndicTrans2.
- Anthropic. 2026. Claude Opus 4.8. <https://www.anthropic.com>.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. **SMOL: Professionally translated parallel data for 115 under-represented languages**. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. **Understanding back-translation at scale**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. **IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages**. *Transactions on Machine Learning Research*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. **LoRA: Low-rank adaptation of large language models**. *arXiv preprint arXiv:2106.09685*.
- Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. **Bilex rx: Lexical data augmentation for massively multilingual machine translation**. *arXiv preprint arXiv:2303.15265*.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. **Neural machine translation for limited resources English–Nyishi pair**. *Sādhanā*, 48(4):237.
- Damjan Kalajdzievski. 2023. **A rank stabilization scaling factor for fine-tuning with LoRA**. *arXiv preprint arXiv:2312.03732*.

- Ontiwell Khongthaw, G.I. Salvin, Shrikant Budde, Abigail Chigwededza, Dhruvadeep Malkar, and Swapnil Hingmire. 2025. [DoDS-IITPKD:submissions to the WMT25 low-resource Indic language translation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1248–1252, Suzhou, China. Association for Computational Linguistics.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. [DoRA: Weight-decomposed low-rank adaptation](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Partha Pakray, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. [Findings of WMT 2025 shared task on low-resource Indic languages translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. [Findings of WMT 2024 shared task on low-resource Indic languages translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 28-29 October 2026. Findings of the WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Workshop on Machine Translation (WMT) under EMNLP 2026*, Budapest, Hungary.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. [Findings of the WMT 2023 shared task on low-resource Indic language translation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. 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.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.
- 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.
- Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Jörg Tiedemann. 2020. [The Tatoeba Translation Challenge – realistic data sets for low resource and multi-lingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.