Automatic Post-editing Shared Task 2022

(Findings of the 8th round)

Pushpak Bhattacharyya[1], Rajen Chatterjee[2], Markus Freitag[3], Diptesh Kanojia[4], Matteo Negri[5] and Marco Turchi[6]

[1]IIT Bombay, [2]Apple Inc., [3]Google, [4]University of Surrey,
[5]Fondazione Bruno Kessler, and [6]Zoom Video Communications

Shared Task Overview
(2022 Edition)

2022 Edition - Motivation

- Improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at the decoding stage;
- Cope with systematic errors of an MT system whose decoding process is not accessible;
- Provide professional translators with improved MT output quality to reduce (human) post-editing effort;
- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain.

2022 edition - Goals

- To extend the languages covered in our datasets;
- To further motivate post-MT efforts for automatic post-editing (APE);
- To encourage further research on low-resource Indian languages;
- To study and promote more fine-grained approaches for APE leading to better performance.

2022 edition - Task

Task: Automatic Post-editing for English-Marathi language pair.

Creation of APE model(s) which can identify and correct errors in the Machine Translation (MT) output using the gold-standard and synthetic data provided by task organizers.

The task formulation remains the same as from previous rounds.

Novelty:

Language Pair: Marathi is an Indo-Aryan language spoken by ~99 million speakers[1].

Data Domain: Multi-domain APE for healthcare, tourism/culture, and general/news.

[1] Ethnologue 2022 - Ethnologue has been an active research project since 1951 which maintains online archives of recognized languages list, and their statistics.

2022 edition - Data Breakdown

Participants provided with training and development data consisting of (source, target, human post-edit) triplets.

These sets respectively comprise of 18,000 and 1,000 instances, in which:

- The source (SRC) is an English (En) sentence;
- The target (TGT) is a Marathi (Mr) translation of the source produced by a generic, black-box NMT system unknown to participants.
    - Generated via a multilingual NMT system (Ramesh et al.,2022) is based on the Transformer architecture (Vaswani et al., 2017) and is trained on a total of 49 million sentence pairs where the En-Mr parallel corpus is 4.5 million sentence pairs. This parallel data is generic and covers many domains, including the three domains covered by the evaluation setting of this year: healthcare, tourism/culture and general/news.
- The human post-edit (PE) is a manually-revised version of the target, which was produced by native Marathi speakers.

Additionally, the participants were provided artificially-generated data, which:

- Consisted of 2 million synthetic triplets derived from *Anuvaad* En-Mr parallel corpus[1].

[1]https://github.com/project-anuvaad/anuvaad-parallel-corpus

2022 edition - Test Data

1,000 (source, target) pairs

- Similar in nature to the corresponding elements in the train/dev sets (*i.e.,* same domains, same NMT system).
- The human post-edits of the target elements were used to measure APE systems' performance both with automatic metrics (TER, BLEU) and via manual assessments.

Data Analysis and Evaluation
(2022 Edition)

Data Analysis

	Lang.	Domain	MT type	RR_SRC	RR_TGT	RR_PE	Basel. BLEU	Basel. TER	δ TER
2015	en-es	News	PBSMT	2.9	3.31	3.08	n/a	23.84	+0.31
2016	en-de	IT	PBSMT	6.62	8.84	8.24	62.11	24.76	-3.24
2017	en-de	IT	PBSMT	7.22	9.53	8.95	62.49	24.48	-4.88
2017	de-en	Medical	PBSMT	5.22	6.84	6.29	79.54	15.55	-0.26
2018	en-de	IT	PBSMT	7.14	9.47	8.93	62.99	24.24	-6.24
2018	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.38
2019	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.78
2019	en-ru	IT	NMT	18.25	14.78	13.24	76.20	16.16	+0.43
2020	en-de	Wiki	NMT	0.65	0.82	0.66	50.21	31.56	-11.35
2020	en-zh	Wiki	NMT	0.81	1.27	1.2	23.12	59.49	-12.13
2021	en-de	Wiki	NMT	0.73	0.78	0.76	71.07	18.05	-0.77
2022	en-mr	healthcare/ tourism/news	NMT	1.46	0.89	0.72	67.55	20.28	-3.49

NEW

Table 1: Data breakdown from the APE shared task since 2015: languages, domain, type of MT technology, repetition rate and initial translation quality (TER/BLEU of TGT). The last column (δ TER) indicates, for each evaluation round, the difference in TER between the baseline (*i.e.,* the “do-nothing” system) and the top-ranked submission.

Complexity Indicators: Repetition Rate

Complexity Indicators help identify the challenging nature of the task.

Repetition Rate (RR):

- Measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1...4) and combining them using the geometric mean.
- The very low RR values (*i.e.,* 1.46, 0.89, and 0.72 respectively for the SRC, TGT and PE elements) seem to confirm that repetition rate is a scarcely reliable complexity indicator.
- Values close to those observed in rounds were the top-ranked submissions achieved both very large (2020) and very small (2021) gains over the baseline.

Complexity Indicators: MT Quality

Complexity Indicators help identify the challenging nature of the task.

MT Quality of TGT:

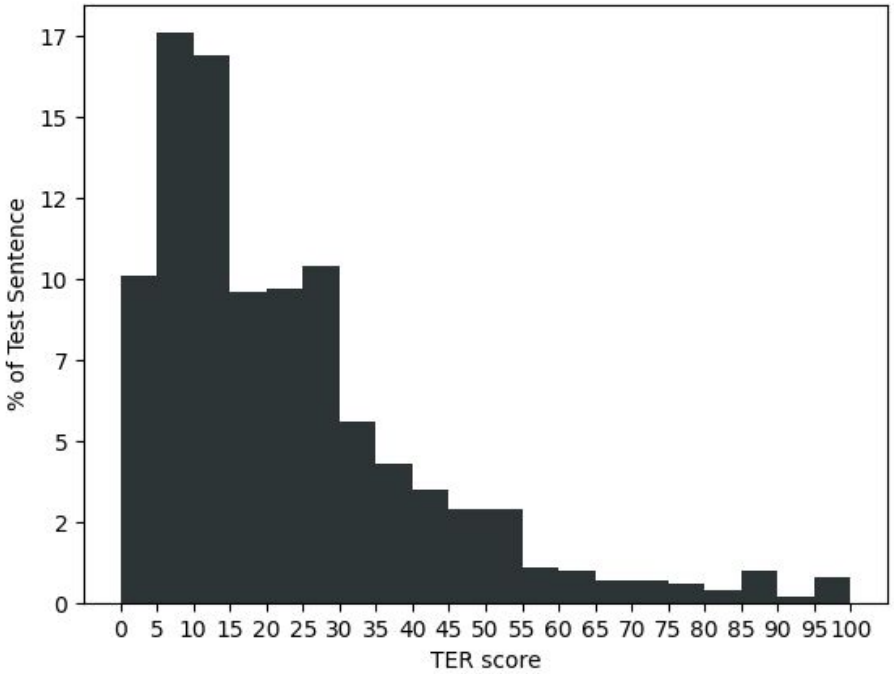
- Measured using TER and BLEU.
- In principle, higher quality of the original translations leaves the APE systems with smaller room for improvement.
- The quality of the initial translations (20.28 TER / 67.55 BLEU) places this round among those of medium-high difficulty (20.0 < TER < 25.0)
- The δ TER of this year (-3.49) [Table 1] also falls in this range, confirming the correlation between the quality of the initial translations and the actual potential of APE.

Complexity Indicators: TER Distribution

Given the TER Distribution on test set this year:

The APE 2022 test set can be considered of medium-high difficulty compared to the past rounds.

As shown in the figure, the TER distribution is quite skewed towards lower values (about 45% of the samples fall in the 15 < TER < 45 interval) but only 10% of the items can be considered as perfect or near-perfect translations (i.e., 0<TER<5).



Evaluation Metrics

- Automatic evaluation was carried out after tokenizing the data using *sacremoses.*
- Computing the distance between the automatic post-edits produced by each system for the target elements of the test set, and the human corrections of the same test items.
- Case-sensitive TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) were respectively used as primary and secondary evaluation metrics.
- The official systems' ranking is hence based on the average TER calculated on the test set.

System Submissions and Results
(2022 Edition)

Baseline Approach

- The official baseline results - TER and BLEU scores calculated by *comparing the raw MT output with human post-edits*.
- This corresponds to the score achieved by a *"do-nothing"* APE system that leaves all the test targets unmodified.
- For each submitted run, the statistical significance of performance differences with respect to the baseline was calculated with the bootstrap test (Koehn, 2004).

System Submissions

Each participating team was allowed to submit at most 2 system outputs (explicitly indicating *primary* submission).

- In the case that none of the submissions is marked as primary, the latest submission was considered the primary submission.

Submissions invited via email with a file naming pattern:

- **INSTITUTION-NAME_METHOD-NAME_SUBTYPE**, where:
  - **INSTITUTION-NAME** is an acronym/short name for your institution, e.g. "UniXY"
  - **METHOD-NAME** is an identifier for your method, e.g. "pt_1_pruned"
  - **SUBTYPE** indicates whether the submission is primary or contrastive with the two alternative values: PRIMARY, CONTRASTIVE.

Participants also invited to submit a short paper to WMT (optional).

Multiple extensions were provided to all participants given the challenging nature of the task.

Submissions Received

ID	Participating team
IITB	Computation for Indian Language Technology - IIT Bombay, India (Deoghare and Bhattacharyya, 2022)
IIIT-Lucknow	IDIAP Research Institute, Switzerland
LUL	Samsung Research and Communication University of China, China (Xiaoying et al., 2022)

Table 2: Submissions received from these three* teams.

*The IIIT-Lucknow team did not produce a system description and is left out of our analysis.

LUL (Samsung Research and Communication University of China)

Transformers-based APE system built using *fairseq* (Ott et. al., 2019)

Approach:

- Data Augmentation - generating synthetic triplets
  - In-house MT system
    - Translate text drawn from various sources.
  - External System (Google Translate)
    - Back-translate the post-edits in APE train set.
- Mixture of experts'
  - Using domain-specific adapters added to the decoder the the base APE model.

IITB (Computation for Indian Language Technology Lab at IIT Bombay)

Transformers-based APE system using a multi-source approach (Chatterjee et. al., 2017)

Approach:

- Two encoders to generate representations for SRC and MT, w/ a single decoder.
- Curriculum-learning strategy
  - Incrementally done using synthetic data, and then fine-tuning on real APE data.
- Uses LaBSE to filter low-quality synthetic triplets.
- Additionally, uses sentence-level quality estimation model to avoid overcorrection where the data was acquired from the newly release En-Mr subtask data from the QE Shared task.

APE Shared Task Results

		TER	BLEU
en-mr	IITB_APE_QE_combined_PRIMARY.tsv	**16.79**	**72.92**
	LUL_HyperAug_Adaptor_CONTRASTIVE	**19.06**	**69.96**
	LUL_HyperAug_Finetune_PRIMARY	**19.36**	**69.66**
	baseline (MT)	20.28	67.55
	IIIT-Lucknow_adversia-machine-translation_PRIMARY.txt	57.14	23.43
	IIIT-Lucknow_adversia-machine-translation_CONTRASTIVE.txt	99.81	3.16

Table 3: Results for the WMT22 APE English-Marathi shared task – average TER (↓), BLEU score (↑) Statistically significant improvements over the baseline are marked in bold.

Submission Analysis

Systems	Modified	Improved	Deteriorated	Prec.
IITB_APE_QE_combined_PRIMARY	452 (45.2%)	287 (63.49%)	126 (27.87%)	69.49
LUL_HyperAug_Adaptor_CONTRASTIVE	491 (49.1%)	261 (53.15%)	150 (30.54%)	63.5
LUL_HyperAug_Finetune_PRIMARY	537 (53.7%)	269 (50.09%)	189 (35.19%)	58.73
IIIT-Lucknow_adversia-machine-translation_PRIMARY	999 (99.9%)	46 (0.46%)	929 (92.99%)	0.47
IIIT-Lucknow_adversia-machine-translation_CONTRAS.	1000 (100%)	9 (0.09%)	987 (98.7%)	0.09
Average	69.6 (49.3)	31.4 (55.6)	57.0 (31.2)	38.4 (63.9)

Table 4: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2022 English-Marathi sub-task. The “Prec.” column shows systems’ precision as the ratio between the number of improved sentences and the number of modified instances for which improvement/deterioration is observed (i.e., Improved + Deteriorated)

Conclusion and Future Direction

Conclusion

- 8th round of the APE shared task conducted in 2022.
- Language pair focus on English - Marathi with domain focus on:
  - Healthcare
  - Tourism/Culture
  - General/News
- Human evaluation carried out but unreliable outcome.
- Discussion on complexity indicators - medium/high difficulty APE task.
- Two systems able to improve over the *“do-nothing”* baseline.
  - Error reductions upto -3.49 TER and +5.37 BLEU.
  - Confirms viability of the APE task for downstream improvements of *“black-box”* NMT systems.

Future Direction

- New test sets ready for future En-Mr APE Shared task for 2023 and 2024 editions.
- We invite submissions for the 2023 APE Shared Task. :)

References

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. Transactions of the Association for Computational Linguistics, 10:145–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, pages 223–231.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sourabh Deoghare and Pushpak Bhattacharyya. 2022. IIT Bombay’s WMT-22 automatic post-editing shared task submission. In Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi. Association for Computational Linguistics.

Huang Xiaoying, Lou Xingrui, Zhang Fan, and Tu Mei. 2022. LUL’s WMT-22 automatic post-editing shared task submission. In Proceedings of the Seventh Con- ference on Machine Translation, Abu Dhabi. Associ- ation 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: FBK’s participation in the WMT 2017 APE shared task. In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pages 630–638, Copenhagen, Denmark. Association for Computational Linguistics.