SLPG_FJWU at WMT 2026: Domain Adapted Augmentation for Low Resource Arabic-Asian Language Translation

Shomaiza Zafar[1,5], Sadaf Abdul Rauf[1,2], Sheema Firdous[1,3], Muhammad Saad Munir[4] and Nadeem Fakhar[5]

[1] Speech and Language Processing Group (SLPG), FJWU-IIUI, Islamabad, Pakistan
[2] Department of Computer Science, International Islamic University, Islamabad, Pakistan
[3] The College of Engineering and Physical Sciences, Aston University, Birmingham, United Kingdom
[4] Aitchison College, Lahore, Pakistan
[5] Department of Computer Science, Fatima Jinnah Women University, Pakistan

{shomaizazafar786,sadaf.abdulrauf,sheemafirdous400}@gmail.com

Abstract

This paper describes the SLPG_FJWU submission to the WMT 2026 Low-Resource Arabic-Asian Language Translation shared task, covering three language directions: Urdu→Arabic, Arabic→Urdu, and Arabic→English. For each direction, we submitted a primary and a contrastive system. Given the Urdu-Arabic as a low resource language pair, we built a supplementary silver standard corpus by mining CC-100 Urdu monolingual text with sentence embedding based domain filtering and machine translation, adding 27,172 sentence pairs on top of the 20,299 official WMT pairs. For Arabic-English, we augmented the WMT data with News Commentary. The contrastive systems outperformed our primary systems in every direction, with BLEU gains ranging from 1.6 to 3.2 points, and on the official leaderboard our contrastive systems ranked 1st in Arabic→Urdu, 3rd in Urdu→Arabic, and 5th in Arabic→English by BLEU.

1 Introduction

This paper describes the SLPG_FJWU submission to the WMT 2026 Low-Resource Arabic-Asian Language Translation shared task (Laskar et al., 2026). We explore domain adapted corpus augmentation for low resource Neural Machine Translation (NMT), focusing specifically on the Arabic-Urdu language pair. To address the limited availability of domain specific parallel resources, we augment the training corpus using Large Language Model(LLM) powered monolingual data mining followed by domain aware filtering. We further explored the potential of a pretrained multilingual model to transfer this data into translation quality across three directions.

NMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) has witnessed great success over the years (Vaswani et al., 2017; Zhang and Zong, 2020). These systems leverage large scale parallel corpora to learn continuous representations of source and target languages, allowing them to capture complex linguistic structure, contextual dependencies, and semantic relationships more effectively than conventional Statistical Machine Translation (SMT) systems (Koehn et al., 2003; Cho et al., 2014). However, the effectiveness of NMT systems is closely tied to the availability of large scale, high quality parallel corpora, creating a substantial barrier for low resource language pairs (Koehn and Knowles, 2017; Fadaee et al., 2017), including Arabic paired with under resourced Asian languages.

As Arabic continues to play an expanding role in global education, business, and communication, developing effective translation frameworks for Arabic centric multilingual interactions has become increasingly important, especially for low resource Asian languages where the availability of high quality parallel corpora remains limited (El Idrysy et al., 2025). The Urdu-Arabic language pair represents a particularly challenging low resource translation scenario. Despite sharing a common (Perso-Arabic) writing script and substantial historical and linguistic interactions, the availability of high quality parallel resources remains extremely limited. For instance, the official training data provided by the WMT 2026 shared task comprises only 20,299 parallel sentence pairs, limiting the ability of NMT models to effectively learn cross-lingual representations.

Despite the substantial gains achieved by NMT in resource rich settings, its performance remains highly sensitive to the availability of parallel training data, with low resource language pairs such as Urdu-Arabic remaining highly constrained by

Proceedings of the Eleventh Conference on Machine Translation, pages 2351–2359
October 28-29, 2026 ©2026 Association for Computational Linguistics

data scarcity, thereby limiting translation quality (Koehn et al., 2003; Abouammoh et al., 2016). To address these limitations, prior research has widely explored a range of data centric strategies, including corpus mining for domain adaptation (Chu and Wang, 2018; Firdous and Rauf, 2023; Naz et al., 2020), transfer learning (Pan and Yang, 2010; Zoph et al., 2016; e Hira et al., 2019), and back-translation (Sennrich et al., 2016; Edunov et al., 2018; Haq et al., 2020) etc. Corpus mining for low resource pairs has evolved from classifier and SMT based extraction over comparable corpora (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009) to embedding based similarity mining (Bouamor and Sajjad, 2018; Artetxe and Schwenk, 2019; Feng et al., 2022) (Section 2). We focused on leveraging monolingual data through domain adapted filtering, combined with transfer learning through fine-tuning a large pretrained multilingual model.

For this submission, we explore the potential of sentence embedding based domain filtering for mining parallel data from Urdu monolingual text. Urdu sentences are mined from CC-100 (Conneau et al., 2020) as detailed in Section 3. For learning sentence embeddings of the mined monolingual data, rather than training encoders from scratch, we leverage a pretrained multilingual encoder for sentence filtering in our Urdu-Arabic pipeline. We used LEALLA-large, a lightweight system developed by (Mao and Nakagawa, 2023), to compute the language agnostic low-dimensional sentence embeddings for each sentence and the WMT DevTest anchor. Sentences were filtered based on their similarity to the anchor set, and then translated into Arabic to form a silver-standard training corpus. For Arabic-English, we merge the official WMT data with News Commentary (Tiedemann, 2012), to expand the training pool. Our main contributions include:

- Presenting a domain0 adaptation methodology for sentence selection using LEALLA sentence embeddings, adapted from comparable-corpus retrieval (Firdous and Rauf, 2023) to a purely monolingual setting
- Applying transfer learning through fine-tuning NLLB-200 (Costa-jussà et al., 2022) across all three translation directions

The paper is structured as follows: Section 3 elaborates the data collection and augmentation pipeline, Section 4 describes model selection and training, Section 5 outlines the results and discussion, followed by the conclusion in Section 6.

2 Related Work

From comparable corpus mining to embedding-based extraction. Extracting parallel sentences from comparable, non-parallel corpora was treated as a classification problem by (Munteanu and Marcu, 2005) who trained a maximum entropy classifier to extract translation pairs from non-parallel news text, later (Abdul-Rauf and Schwenk, 2009; Abdul Rauf and Schwenk, 2011) used SMT and Information Retrieval pipeline to achieve the same objective. These pipelines were later replaced by embedding based similarity: Bouamor and Sajjad (2018) combined multilingual sentence embeddings with NMT and a classifier, Artetxe and Schwenk (2019) proposed margin based scoring over multilingual sentence embeddings and this approach was scaled to produce the WikiMatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b) parallel corpora. Heffernan et al. (2022) introduced a teacher student distillation approach to train LASER3, a set of encoders targeting individual low resource languages, which the NLLB Team (NLLB Team et al., 2022) then adopted at scale to mine parallel data for over 200 languages, including many with little to no prior parallel corpora.

Domain adapted corpus filtering Firdous and Rauf (2023) used LEALLA embeddings to filter comparable biomedical corpora by domain anchor similarity where the domain anchor is an embedding matrix of in domain sentences, they improved BLEU by more than 2.3 points through domain centric filtering. Our pipeline (Section 3.2) follows the same embed-then-threshold-filter approach but adapts it from a comparable corpus to a purely monolingual setting: Urdu CC-100 sentences and WMT DevTest sentences are embedded with LEALLA (Mao and Nakagawa, 2023), a knowledge distilled, 256-dimensional version of LaBSE (Feng et al., 2022), sentences below a similarity threshold are discarded and the corresponding Arabic side is produced by machine translation.

Synthetic and mined training data. Our silver standard generation uses forward translation the counterpart to back translation (Sennrich et al., 2016; Edunov et al., 2018) where a source to tar-

get system generates the synthetic target side from monolingual source data. The domain filtered Urdu sentences are translated into Arabic using Google Translate (Section 3.2) to form pseudo parallel training pairs. This choice of direction is an active design decision: Abdul-Rauf et al. (2016) show that domain appropriate data selection is needed regardless of which side is synthesized, and both back-translation (Haq et al., 2020) and forward-translation strategies have proven effective for low resource MT. Synthetic pairs improve translation quality provided the synthetic side does not introduce systematic noise that the model then learns to reproduce (Edunov et al., 2018).

Transfer learning for low resource NMT. Transfer learning addresses data scarcity by reusing knowledge from a resource rich task or language for a resource poor target (Pan and Yang, 2010; Zoph et al., 2016). Our pipeline (Section 4) follows this direction by fine-tuning NLLB-200 (Costa-jussà et al., 2022), a massively multilingual pre-trained model, on our domain augmented corpora across all three translation directions.

3 Dataset Preparation

3.1 Official WMT Training Data

The training data from WMT 2026 (Laskar et al., 2026) includes a parallel corpus for each language pair. For the Urdu-Arabic pair, the official training corpus consists of 20,299 sentence pairs in the formal news domain, while the development corpus and DevTest corpora consist of 900 and 500 sentence pairs, respectively. For Arabic–English, the availability of abundant parallel texts has resulted in a substantially larger official training corpus. Our internal WMT only training corpus includes 21,000 sentence pairs (see Section 3.3).

3.2 Domain Adapted Corpus Construction for Urdu-Arabic

The primary contribution of the current study is a domain adapted auxiliary corpus for Urdu-Arabic, where CC-100 Urdu monolingual data is extracted and filtered on the basis of relevance to the WMT news domain, and then the target side is generated in Arabic. Unlike volume based augmentation of data, the proposed pipeline is based on previous research, where they concluded that indiscriminate extraction results in dilution of the domain specific training signal, while a smaller domain specific corpus is more effective than a larger corpus that is not domain specific (Firdous and Rauf, 2023; Shi, 2025). The proposed pipeline has four stages, summarized in Figure 1 and Table 1: corpus preparation, embeddings, domain anchored similarity filtering, and silver standard target generation.

1. Corpus preparation: CC-100's Urdu subset (Conneau et al., 2020) contains around 27.5 million lines of web-crawled texts. The crawled corpus cannot be directly used for embedding computation due to substantial text duplication, inconsistent line-ending encodings, and the presence of multi-sentence lines. After resolving these issues through corpus cleaning and normalization, 21,753,881 sentences were retained, from which a two million sentence subset was selected using reservoir sampling. Since the sampling is performed independently of sentence position, each sentence has an equal probability of being selected, ensuring that the resulting subset is not biased by its location within the corpus.

2. Embeddings: All two million sampled Urdu sentences were subsequently encoded with LEALLA-large (Mao and Nakagawa, 2023), a knowledge distilled multilingual sentence encoder that compresses the representational capacity of LaBSE (Feng et al., 2022) into a much lighter 256-dimensional embedding space through sequence level knowledge distillation (Kim and Rush, 2016). We chose LEALLA specifically for its compute accuracy trade-off at this scale, as done by Firdous and Rauf (2023) for comparable corpus filtering. The WMT DevTest Urdu sentences were embedded with the same model to act as domain anchors, fixed reference points in the embedding space, representing the target distribution of the news and current affairs domain, against which every sentence is scored.

3. Domain anchored similarity filtering: For each sentence in the two million sentence sample, we computed cosine similarity to the DevTest anchor embeddings. Figure 2 shows the yield of this filtering step across a range of thresholds, from 30 sentences at ≥ 0.80 to 128,933 sentences at 0.50. We retained only those sentences with scores above 0.55, selected as an operating point on the yield curve rather than tuned via grid search against held-out validation data: yield falls sharply above 0.55 (5,630 at 0.60, 1,295 at 0.65), too small to meaningfully augment the 20,299 official pairs, while lowering to 0.50 raises yield $\sim$4.7$\times$ to 128,933

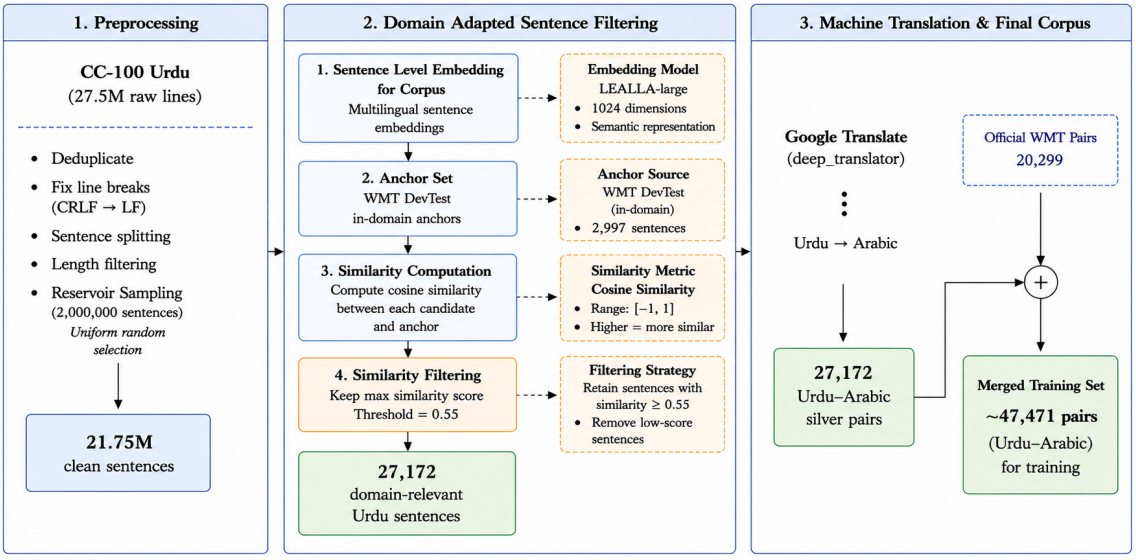

Figure 1: Pipeline for constructing the Urdu-Arabic training corpus from CC-100 (news domain).

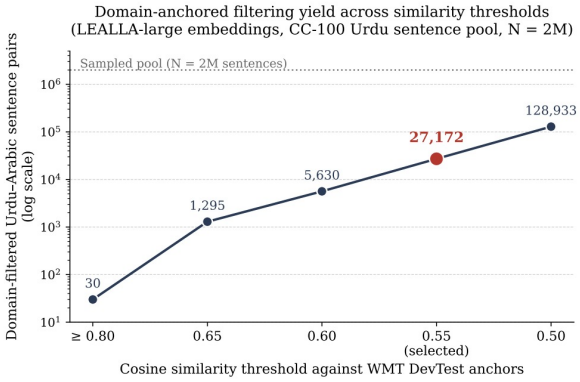

Figure 2: Domain anchored filtering yield across similarity thresholds (LEALLA large embeddings, CC-100 Urdu sentence pool, $N = 2$M). The selected threshold (0.55) is highlighted.

sentences at the cost of weaker similarity to the news and current affairs anchor. This balances corpus scale against domain selectivity, consistent with prior findings that a smaller, domain matched corpus can outperform a larger corpus that is not domain specific (Firdous and Rauf, 2023; Shi, 2025). The result is 27,172 domain relevant sentences, 0.12% of the 21.75-million-sentence cleaned pool rather than merely being Urdu text, these sentences are selected specifically for relevance to the shared task's target domain. Section 5 discusses how the resulting improvement relates to domain alignment rather than data volume.

4. Silver standard target generation: The 27,172 domain filtered Urdu sentences were translated into Arabic using the Google Translate API [1] via the `deep_translator` package, producing machine generated Arabic translations for authentic, domain relevant Urdu source text. The resulting pairs are silver standard but are domain relevant and formatted consistently with the official WMT data. Combined with the 20,299 official pairs, this gives a ~47,471-pair training set for the Urdu-Arabic contrastive system (see Table 1).

Quality estimation: To assess the translation quality of the silver standard pairs independently of the domain similarity filtering already applied in Step 3, we additionally scored all 27,172 Urdu-Arabic pairs with CometKiwi (Rei et al., 2022), using the `wmt22-cometkiwi-da` checkpoint. CometKiwi is a reference-free quality estimation (QE) model that predicts a translation quality score in $[0, 1]$ directly from the source sentence and its candidate translation, without requiring a human reference, making it well suited to a silver standard corpus for which no human Arabic reference exists. The corpus achieved a mean CometKiwi score of 0.795 (median = 0.822, SD = 0.085). As shown in Figure 3, 86.4% of the pairs scored above 0.7 and only 0.7% scored below 0.5, indicating that the domain anchored mining pipeline (Steps 1–3) predominantly yields translation pairs of usable quality, consistent with the downstream training gains reported in Section 5.

[1] Google Cloud, Cloud Translation API, https://cloud.google.com/translate/docs

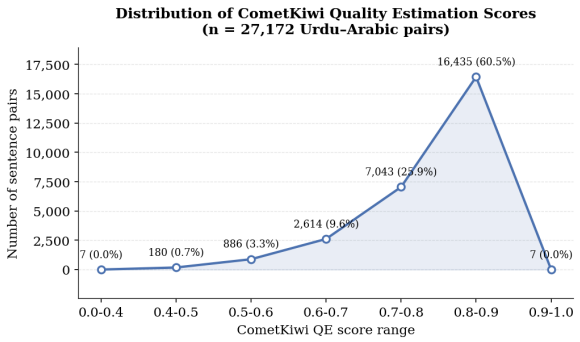

Figure 3: Distribution of CometKiwi reference-free quality estimation scores over the 27,172 mined Urdu-Arabic silver standard pairs.

Pair / System	Data Sources	Pairs
Ur↔Ar Primary	WMT only	20,299
Ur↔Ar Contrastive	WMT + CC-100 (mined)	47,471
Ar→En Primary	WMT only	21,000
Ar→En Contrastive	WMT + News Commentary	111,136

Table 1: Training data composition across all submitted systems.

3.3 Domain Consistent Corpus Expansion for Arabic-English

For Urdu-Arabic (Section 3.2), no domain matched parallel resource existed, so we built a mining pipeline to create one from scratch. Arabic-English did not require this: News Commentary (Tiedemann, 2012) is a parallel corpus of editorial and opinion journalism whose topics and domain closely match the WMT training data, and is itself already an existing, in-domain parallel corpus. We combined News Commentary with the official WMT Arabic-English training data and deduplicated on the Arabic side to remove any sentences already present in the WMT set, yielding a combined pool of roughly 111,136 parallel sentence pairs. This augmented corpus is used as training data for the contrastive system, while the WMT-only data serves the primary system. This reflects a strategy driven by data availability: mining is necessary only where no domain matched parallel resource exists, while an existing in-domain corpus such as News Commentary can be directly incorporated when available.

4 System Description: Model Selection

All four submitted systems fine-tune NLLB-200 (No Language Left Behind; Costa-jussà et al., 2022), a Transformer based (Vaswani et al., 2017) multilingual translation model from Meta-AI, trained to translate directly between 200 languages with meaningful coverage of low resource languages that most earlier multilingual systems largely ignored. This made it a natural starting point for a submission spanning both a well resourced pair (Arabic-English) and a low resource one (Urdu-Arabic).

For each direction, we submit a primary system trained only on the official WMT data, and a contrastive system trained on the corresponding augmented corpus (as described in Section 3), the CC-100 derived corpus for Urdu-Arabic, and the News Commentary combined corpus for Arabic-English. Fine-tuning NLLB-200 on this data constitutes the transfer learning component of our submission, directly instantiating the shared task's stated transfer learning objective (Laskar et al., 2026) alongside the corpus level contributions described in Section 3. Since naive domain fine-tuning of pretrained multilingual NMT systems can degrade general domain performance if not handled carefully (Grosso et al., 2022), the primary systems serve not only as a data constrained baseline but as a stable reference point: they show what is achievable using only the official WMT data, and provide the point of comparison for judging how much the augmented data actually helped the contrastive systems (Section 5).

Direction	System	BLEU	chrF	COMET	Rank
Ur→Ar	Primary	14.04	48.83	80.22	5th/6
Ur→Ar	**Contrastive**	**16.99**	**51.47**	**82.84**	**3rd/6**
Ar→Ur	Primary	24.58	50.10	74.53	3rd/6
Ar→Ur	**Contrastive**	**27.80**	**52.54**	**76.26**	**1st/6**
Ar→En	Primary	28.37	55.50	80.96	5th/11
Ar→En	**Contrastive**	**30.00**	**57.04**	**81.15**	**5th/10**

Table 2: Internal DevTest scores (BLEU / chrF / COMET) and official WMT 2026 leaderboard rank, by direction and system.

5 Results and Discussion

Our internal DevTest score is reported along with WMT 2026 leaderboard rank in Table 2. The internal score has been calculated using sacreBLEU (Post, 2018), which computes BLEU (Papineni et al., 2002) and chrF (Popović, 2015); COMET (Rei et al., 2020) is reported alongside them, while the rank has been published by the organizers of the shared task (Laskar et al., 2026).

Across all translation directions, our contrastive

system consistently outperformed the primary system in both official metrics and internal evaluation metrics. The improvement is most significant in the Urdu-Arabic direction. CC-100 filtering according to the domain adds 2.95 BLEU / 2.64 chrF for Urdu→Arabic and 3.22 BLEU / 2.44 chrF for Arabic→Urdu. The additional 27,172 sentences constitute only a tiny fraction of the 21.75-million-sentence clean CC-100 set. We attribute this improvement primarily to domain aligned filtering rather than mining volume, consistent with prior findings that data quality and domain fit, rather than synthetic data volume, drive gains in low resource NMT (Shi, 2025).

For Arabic-English, News Commentary augmentation consistently gives us a boost of 1.63 BLEU / 1.54 chrF, keeping our official rank at 5th place among competing teams. This is one of the harder pairs to move: Arabic-English already has a strong baseline (28.37 BLEU for the primary system alone), therefore, any further gains must come from true domain alignment rather than filling an obvious data gap. The consistent gains achieved by the augmented corpus across both official and internal evaluation sets demonstrate that fine-tuning NLLB-200 on a training corpus matched in both domain and language, expanded fivefold, constitutes a meaningful contribution, rather than an improvement confined solely to the low resource Urdu–Arabic setting.

CHrF is consistently higher than BLEU across all systems, reflecting its partial matching for characters and its ability to deal with the rich morphology of Arabic and Urdu. Both chrF and BLEU produce consistent system rankings across all six configurations, reinforcing the relative performance trends observed under the official evaluation metrics and confirming the robustness of the evaluation. COMET scores show the same pattern: contrastive systems outperform primary systems across all three directions, with the largest gains in Urdu→Arabic (+2.61 COMET) and Arabic→Urdu (+1.72 COMET). The Arabic-English COMET gap (+0.19) is smaller than its BLEU/chrF gap, suggesting the augmentation's benefit there is more modest under semantic evaluation than surface overlap metrics alone indicate. The results further demonstrate that the proposed pipeline comprising domain specific corpus construction from CC-100, embedding driven data filtering, and subsequent fine-tuning of NLLB-200 plays a central role in improving translation performance. The proposed approach yields its most substantial gains for the most severely under resourced language pairs, while also delivering measurable improvements for the comparatively resource rich Arabic-English pair, indicating that its effectiveness extends beyond low resource translation scenarios.

6 Conclusion

This paper presented the SLPG_FJWU submission to WMT 2026 across three translation directions, Urdu→Arabic, Arabic→Urdu, and Arabic→English. Our primary contribution is a silver standard, domain adapted corpus for Urdu-Arabic, extending the data preparation method of Firdous and Rauf (2023) from comparable to monolingual data mining and yielding 27,172 domain matched parallel sentences. For Arabic-English, combining the official WMT data with the News Commentary corpus produced a substantially larger, domain consistent training pool. Both augmentation strategies, paired with NLLB-200 fine-tuning as the transfer learning component of our submission, outperformed all constrained primary systems, achieving official leaderboard placements of 1st in Arabic→Urdu, 3rd in Urdu→Arabic, and 5th in Arabic→English.

For the severely low resource Urdu-Arabic pair, these gains were achieved using a domain filtered corpus that is only a small fraction of the size of the underlying monolingual pool, pointing to domain alignment as a key contributing factor; isolating its effect from mining volume through a controlled ablation remains an important next step. Reference-free CometKiwi scoring of the silver standard pairs (Section 3.2) further indicates that the mined corpus is predominantly of usable translation quality, lending additional support to this attribution. More broadly, we expect this construction methodology to transfer to other Arabic-Asian pairs facing similar resource constraints within the shared task's scope.

Future work includes a controlled ablation comparing the domain filtered corpus against an equally sized random, unfiltered sample to isolate the effect of domain alignment from mining volume, and exploring back translation from Arabic monolingual data as a further augmentation strategy for Urdu→Arabic.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4):341–375.

Sadaf Abdul-Rauf, Holger Schwenk, Patrik Lambert, and Mohammad Nawaz. 2016. Empirical use of information retrieval to build synthetic data for SMT domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):745–754.

Murad Abouammoh, Kashif Shah, and Ahmet Aker. 2016. Creation of comparable corpora for English-Urdu, Arabic, Persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4193–4196, Portorož, Slovenia. European Language Resources Association (ELRA).

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. corr abs/1806.00258 (2018). *arXiv preprint arXiv:1806.00258*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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*.

Noor e Hira, Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, and Raheel Nawaz. 2019. Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 156–163, Florence, Italy. 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.

Fatima Zahra El Idrysy, Soufiane Hourri, Ikram El Miqdadi, Assia Hayati, Yassine Namir, Bassma Ncir, and Jamal Kharroubi. 2025. Unlocking the language barrier: A journey through arabic machine translation. *Multimedia Tools and Applications*, 84(14):14071–14104.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Sheema Firdous and Sadaf Abdul Rauf. 2023. Biomedical parallel sentence retrieval using large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 263–270, Singapore.

Mathieu Grosso, Pirashanth Ratnamogan, Alexis Mathey, William Vanhuffel, and Michael Fotso Fotso. 2022. Robust domain adaptation for pre-trained multilingual neural machine translation models. In *Proceedings of the Second Workshop on Multilingual Representation Learning (MMNLU, EMNLP 2022 Workshop)*. ArXiv:2210.14979.

Sami Ul Haq, Sadaf Abdul Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document level nmt of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*,

pages 442–446. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*, pages 127–133.

Sahinur Rahman Laskar, Firoj Alam, Bishwaraj Paul, Irfan Ahmad, Maya Silvi Lydia, and Pankaj Kundan Dadure. 2026. Findings of the WMT 2026 shared task on low-resource Arabic-Asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Sumbal Naz, Sadaf Abdul-Rauf, Sami Ul Haq, et al. 2020. Fjwu participation for the wmt20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 849–856.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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.

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.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, 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.

Zhiqiang Shi. 2025. Data augmentation for low-resource neural machine translation: A systematic analysis. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 510–522, Mumbai, India. The Asian Federation of Natural Language Processing and Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

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 30*, pages 5998–6008.

Jiajun Zhang and Chengqing Zong. 2020. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.