Automatically Filtering Domain Specific Data from Web Mined Parallel Corpora

Supervisor: Dr. Nisansa de Silva

Co-supervisor: Dr. Surangika Ranathunga

Presented by: 248025J - C.S.M. Rathnayaka

01

Research Problem & Objectives

Problem

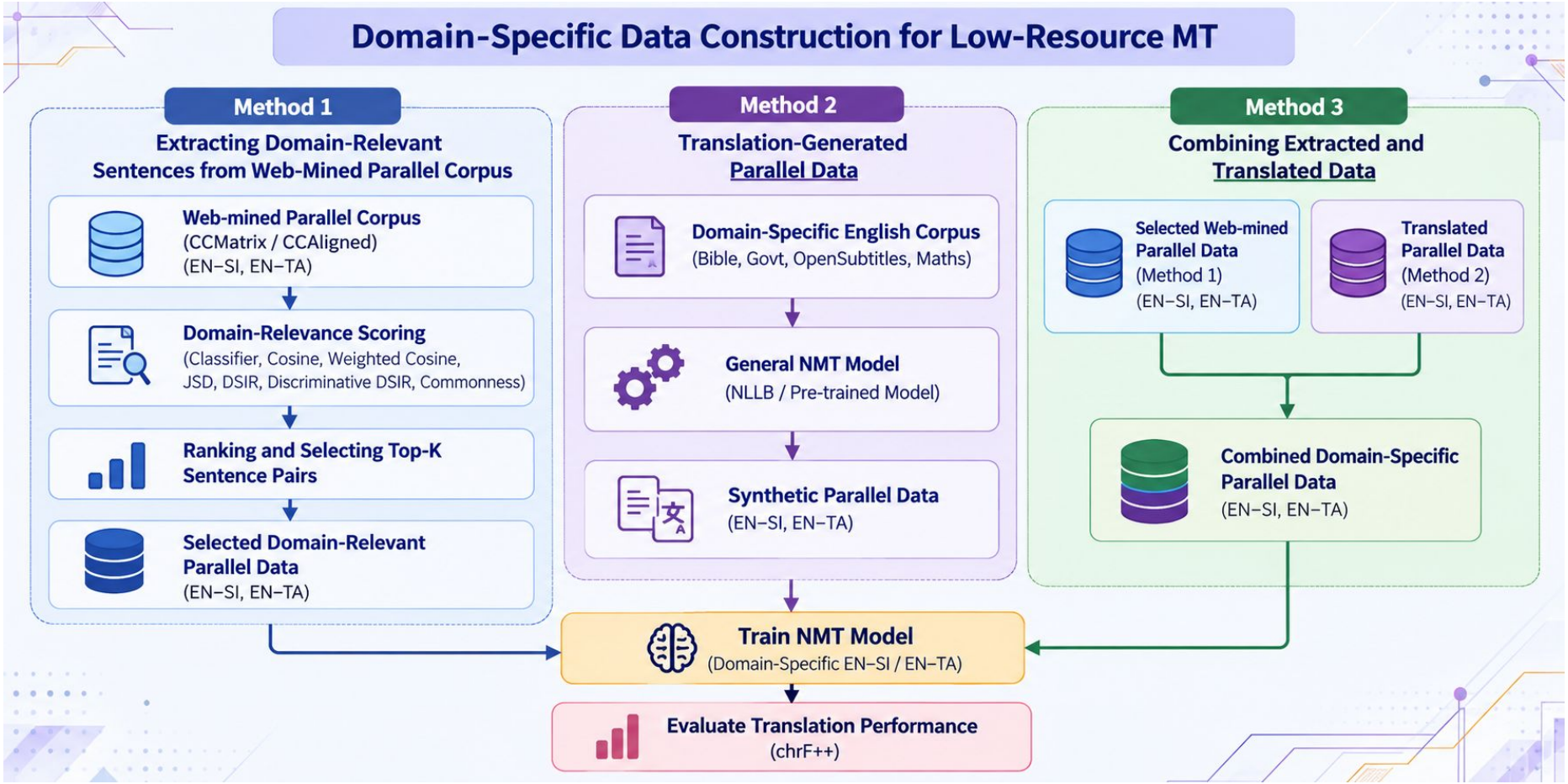
- Domain-specific NMT requires domain-relevant parallel data.
  - Translation models perform better when trained on data representative of the target domain.[1]
- Large web-mined parallel corpora are available
  - These corpora are noisy, mixed-domain, and contain varying-quality sentence pairs.
- How can we automatically identify useful domain-relevant parallel sentence pairs from large, noisy web-mined corpora?

Objectives

- Design and implement an automated method to **select domain-relevant sentence pairs** from large, noisy web-mined parallel corpora
- Determine the appropriate quantity and quality of selected domain-specific data to improve low-resource NMT performance while minimizing noise and redundancy.

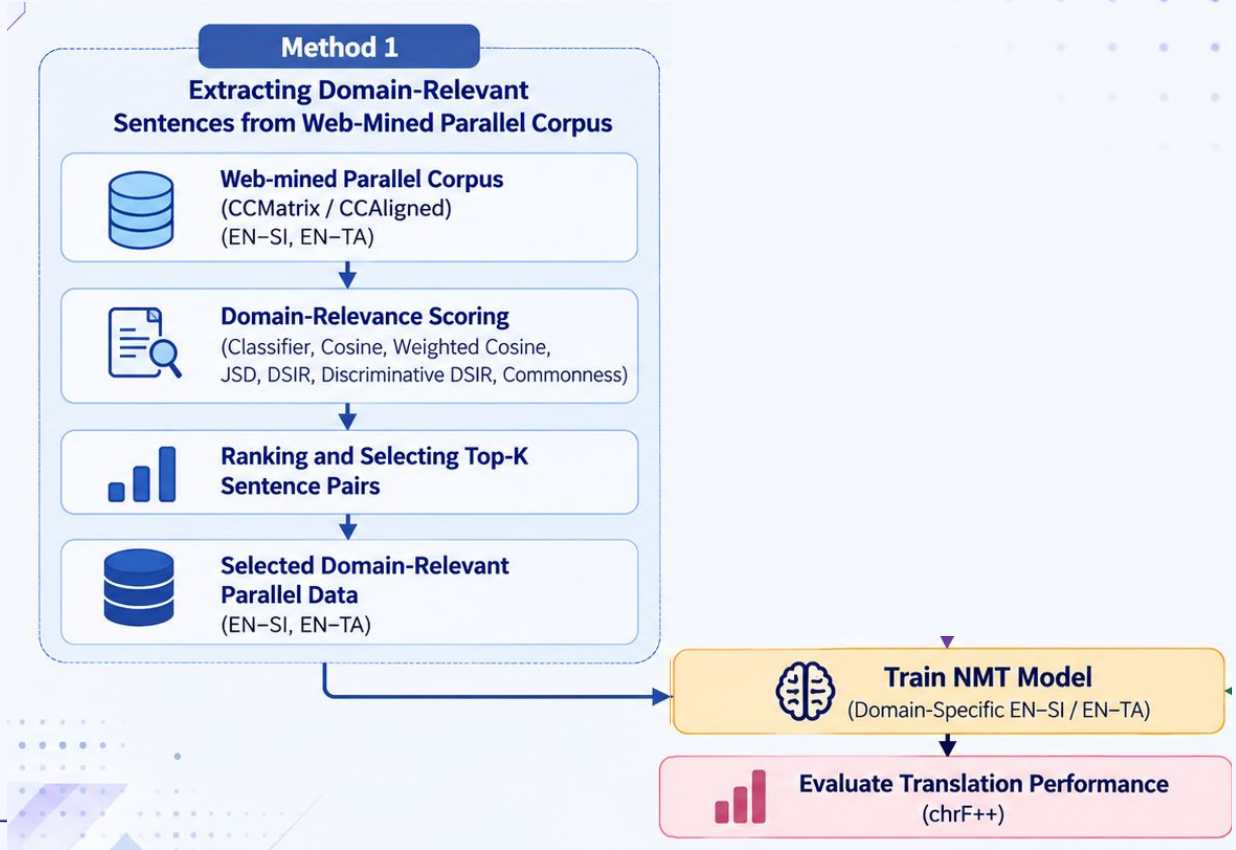
02

Overall Research Framework


Domain-Specific Data Construction for Low-Resource MT
Method 1
Extracting Domain-Relevant Sentences from Web-Mined Parallel Corpus
Web-mined Parallel Corpus
(CCMatrix / CCAligned)
(EN–SI, EN–TA)
Domain-Relevance Scoring
(Classifier, Cosine, Weighted Cosine, JSD, DSIR, Discriminative DSIR, Commonness)
Ranking and Selecting Top-K Sentence Pairs
Selected Domain-Relevant Parallel Data
(EN–SI, EN–TA)
Method 2
Translation-Generated Parallel Data
Domain-Specific English Corpus
(Bible, Govt, OpenSubtitles, Maths)
General NMT Model
(NLLB / Pre-trained Model)
Synthetic Parallel Data
(EN–SI, EN–TA)
Method 3
Combining Extracted and Translated Data
Selected Web-mined Parallel Data
(Method 1)
(EN–SI, EN–TA)
Translated Parallel Data
(Method 2)
(EN–SI, EN–TA)
Combined Domain-Specific Parallel Data
(EN–SI, EN–TA)
Train NMT Model
(Domain-Specific EN–SI / EN–TA)
Evaluate Translation Performance
(chrF++)

03

Method 1: Domain-Relevant Data Extraction


Method 1
Extracting Domain-Relevant Sentences from Web-Mined Parallel Corpus
Web-mined Parallel Corpus
(CCMatrix / CCAligned)
(EN–SI, EN–TA)
Domain-Relevance Scoring
(Classifier, Cosine, Weighted Cosine, JSD, DSIR, Discriminative DSIR, Commonness)
Ranking and Selecting Top-K Sentence Pairs
Selected Domain-Relevant Parallel Data
(EN–SI, EN–TA)
Train NMT Model
(Domain-Specific EN–SI / EN–TA)
Evaluate Translation Performance
(chrF++)

Baseline: Cross-Lingual Similarity-Based Filtering (LASER) [2]

Purpose: Identify high-quality parallel sentence pairs from the web-mined corpus based on the similarity between the source (English) and target (Sinhala/Tamil) sentences using LASER embeddings.

1 Web-mined Parallel Sentence Pair

English (source)

The government announced new policies for education.

Sinhala / Tamil (target)

ජජය අධ්‍යාපනය සඳහා නව ප්‍රතිපත්ති ප්‍රකාශ කළා.

(Sinhala example)

2 Sentence Embeddings using LASER [3]

LASER
Multilingual Sentence Encoder

English sentence embedding → e_{en} (d dimensions)

Target sentence embedding → $e_{si/ta}$ (d dimensions)

3 Compute Cosine Similarity

Cosine Similarity (LASER)

$$s = \frac{e_{en} \cdot e_{si/ta}}{\|e_{en}\| \|e_{si/ta}\|}$$

-1 0 1

Higher similarity score indicates a better aligned translation pair.

4 Rank and Filter Sentence Pairs

Sentence Pair	Similarity Score	
Pair 1	0.82	Keep (high score)
Pair 2	0.76	
Pair 3	0.71	
...	...	
Pair n	0.12	Filter out (low score)

5 Selected Parallel Data (Baseline Dataset)



High-quality parallel sentence pairs selected using source–target cosine similarity (LASER).

Key Points:

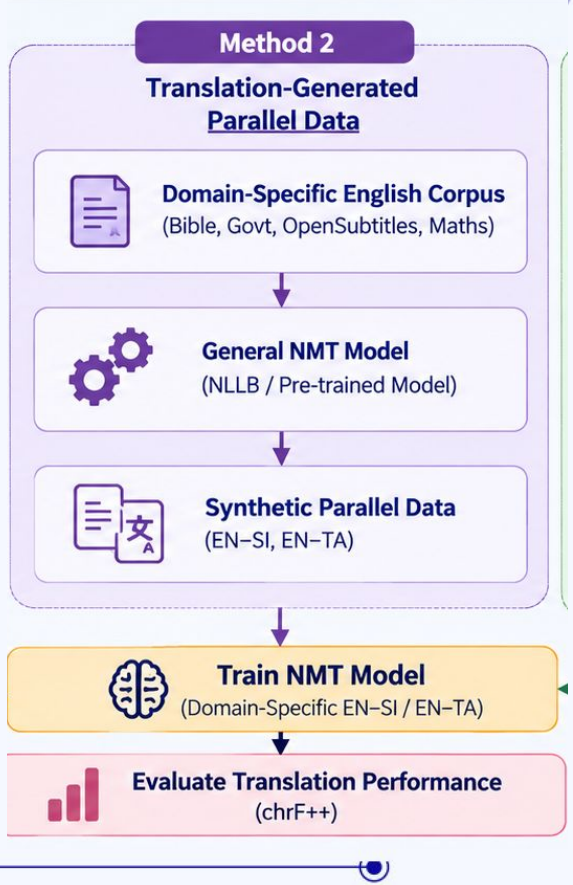
- Uses **LASER** (Language-Agnostic SEntence Representations) to obtain multilingual sentence embeddings for both source and target sentences.
- Filters parallel sentence pairs based on **cosine similarity** between the English and Sinhala/Tamil sentences.
- This baseline focuses on **parallel-data quality** (translation/alignment quality), whereas the proposed methods focus on **domain relevance** of the source-side English sentence.

Domain-Relevance Scoring Methods

Category	Methods
Classification-Based	Classifier
Similarity-Based	Cosine Similarity, Weighted Cosine Similarity
Divergence - Based	Jensen–Shannon Divergence(JSD)
Importance / Relevance-Based	Data Selection for Language Models via Importance Resampling (DSIR), Discriminative DSIR, Commonness

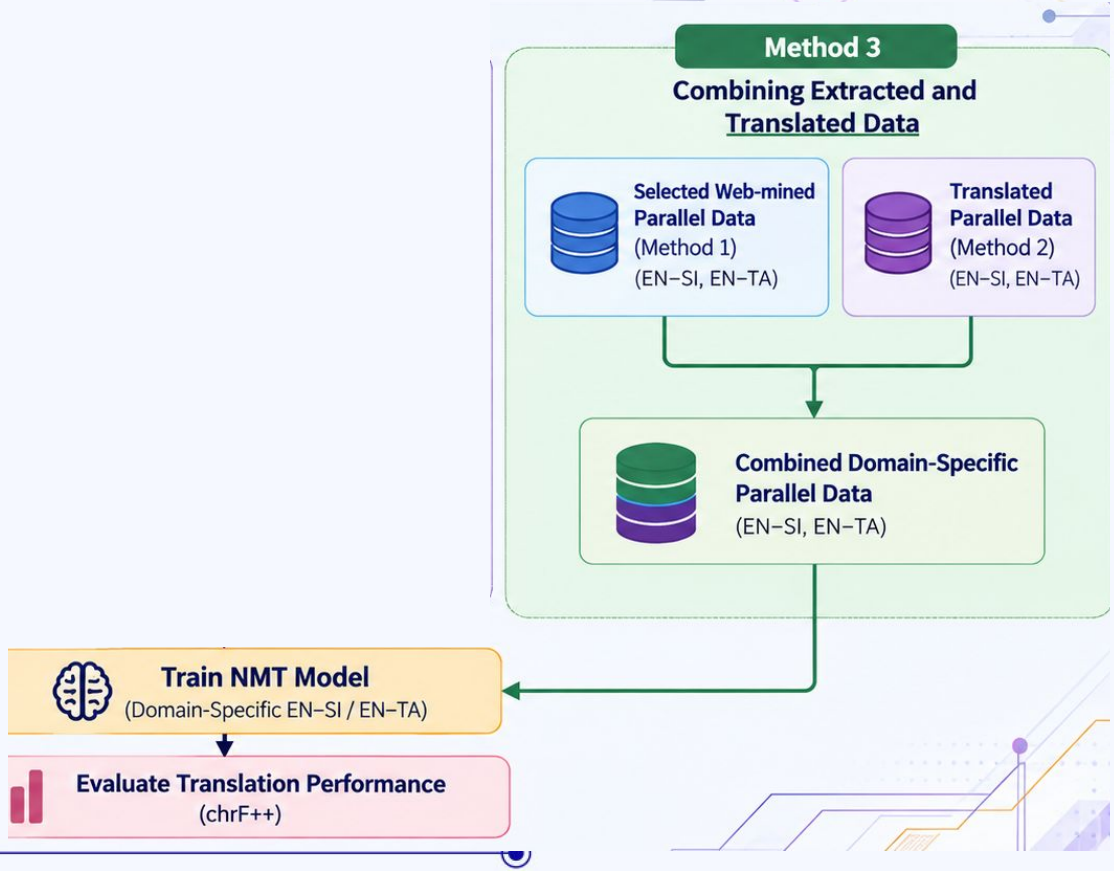
04

Method 2: Translation Generated Data


Method 2
Translation-Generated
Parallel Data
Domain-Specific English Corpus
(Bible, Govt, OpenSubtitles, Maths)
General NMT Model
(NLLB / Pre-trained Model)
Synthetic Parallel Data
(EN–SI, EN–TA)
Train NMT Model
(Domain-Specific EN–SI / EN–TA)
Evaluate Translation Performance
(chrF++)

05

Method 3: Combining Both Approaches


Method 3
Combining Extracted and Translated Data
Selected Web-mined Parallel Data (Method 1) (EN–SI, EN–TA)
Translated Parallel Data (Method 2) (EN–SI, EN–TA)
Combined Domain-Specific Parallel Data (EN–SI, EN–TA)
Train NMT Model (Domain-Specific EN–SI / EN–TA)
Evaluate Translation Performance (chrF++)

06

Datasets & Experimental Setup

Datasets

Language Pairs:

- English - Sinhala
- English - Tamil

Webmined Parallel Corpus:

- CCMatrix Original Dataset[10]
- CCMatrix Cleaned Dataset[11]
- CCAligned Cleaned Dataset[11]

Domain Specific Corpora(English):

- Bible [12](Dataset derived from Bible texts)
- Govt(SiTa) [13](Dataset sourced from official government publications)
- Opensubtitles [14](Dataset extracted from movie subtitles)
- Maths [15](Dataset composed of mathematics questions and answers.)

Evaluation - Basic Configurations

Evaluation Datasets:

- Bible
- Govt
- Opensubtitles
- Maths [16]

Evaluation Metric

- Chrf++ [17]
- Bleu [18]
- SpBleu [19]
- Chrf[20]

Evaluation Task:

- Vanilla NMT - **Fairseq [21]**
  .
- General Purpose NMT - **NLLB [22]**

07

Results

Domain-Relevance Scoring Methods (k=25k, Cleaned CCMatrix)

Domain	En–Si: Best Domain-Relevance Method	chrF++	En–Ta: Best Domain-Relevance Method	chrF++
Bible	Disc. DSIR + Cosine XLMR	**$18.8_{+4.7}$**	Disc. DSIR + Weighted Cosine XLMR / XLM-R Cosine	**$15.5_{+1.1}$**
SiTa	Disc. DSIR + Cosine XLMR	**$23.8_{+1.4}$**	DSIR n-gram (n=2)	**$10.6_{+1.8}$**
OpenSubtitles	Disc. DSIR + Cosine XLMR / JSD	**$34.0_{+11.8}$**	DSIR n-gram (n=2) / XLM-R Cosine	**$38.4_{+17.4}$**
Math	XLM-R Cosine	**$33.0^{*}_{-12.0}$**	XLM-R Cosine	**$34.0_{+3.1}$**

Discriminative DSIR-based approaches showed the strongest or near-strongest performance across most experimental settings.

Extracted Data vs. Translation-Generated Synthetic Data(EnSi)

Data Type	Size	Bible	Govt (SiTa)	OpenSubtitles	Math
Extracted	25K	18.8	23.8	34.0	33.0
Extracted	50K	49.3	57.0	47.3	42.8
Translated	25K	50.9	46.0	30.0	45.3
Translated	~50K	27.5	30.8	46.0	48.9
Translated	100K	–	–	53.1	54.5

Neither approach is universally best;

Extracted Data Size for Combination Experiment

Domain	Base Data	+25K	+50K	+75K	+100K	+125K
Bible	47K	27.5	**28.7**	28.1	24.3	26.2
SiTa / Govt	53K	30.8	**35.3**	35.1	34.3	32.9
OpenSubtitles (50K base)	50K	51.2	**54.6**	52.7	54.5	53.5
OpenSubtitles (100K base)	100K	54.8	**56.2**	54.8	55.7	53.9
Math (50K base)	50K	52.7	54.7	**56.2**	55.0	—
Math (100K base)	100K	54.0	54.8	55.1	**56.6**	—

Selected for combination experiment: 50K extracted

Top 50k Results

Domain	EN→SI CCMatrix	EN→SI CCAligned	EN→TA CCMatrix	EN→TA CCAligned
Bible	**49.3**$_{+3.3}$ — Disc. DSIR + Weighted Cosine (LASER3)	**50.6**$_{+2.8}$ — Disc. DSIR + Weighted Cosine (LASER3)	**38.0**$_{-6.7}$ — Disc. DSIR + Weighted Cosine (XLM-R)	**46.3**$_{+1.8}$ — DSIR 2-gram
SiTa	**57.0**$_{-0.1}$ — Disc. DSIR + Commonness	**65.9**$_{-0.3}$ — Cosine Similarity (LASER3)	**36.1**$_{-8.8}$ — Cosine Similarity (LASER3)	**63.0**$_{-3.2}$ — Disc. DSIR + Commonness (LASER3)
OpenSubtitles	**47.3**$_{+1.0}$ — DSIR 2-gram	**48.4**$_{+0.2}$ — DSIR + Commonness	**62.1**$_{+2.7}$ — Disc. DSIR + Commonness (LASER3)	**64.2**$_{+1.2}$ — Disc. DSIR + Commonness
Math	**42.8**$_{-8.2}$ — Cosine Similarity (LASER3)	**56.3**$_{-0.4}$ — Disc. DSIR + Cosine (LASER3)	**41.2**$_{-17.4}$ — Disc. DSIR + Weighted Cosine (LASER3)	**67.9**$_{+1.1}$ — Disc. DSIR + Commonness (LASER3)

Translated Data size for Combination Experiment

Training Data	Bible	SiTa / Govt	OpenSubtitles	Math
Baseline (Top 50K)	46.0	57.1	48.3	51.0
Translated 25K	50.9	46.0	30.0	45.3
Extracted 50K	49.3	57.0	44.3	42.8
Translated 25K + Extracted 50K	**55.4**	**63.4**	53.6	**55.2**
Translated 50K + Extracted 50K	28.7	35.3	54.6	54.7
Translated 100K + Extracted 50K	—	—	**56.2**	54.8

Progress Since Last Review

✓ Extended experiments to **CCAligned**
✓ Completed **dataset-size optimization** (25K–125K)
✓ Completed **translation-generated synthetic data** experiments
✓ Compared **extracted vs. translated** domain-specific data
✓ Completed **combined-data experiments**
✓ Identified **50K extracted + 25K translated** as the strongest overall configuration
✓ Completed final **EN–SI / EN–TA** experimental evaluation
✓ Completed analysis, thesis writing and revisions
✓ **Thesis ready for submission**

Reference

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

[2] Surangika Ranathunga, Nisansa De Silva, Velayuthan Menan, A. Fernando, and Charitha Rathnayake, “Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora,” ACL Anthology pp. 860–880, Mar. 2024, Accessed: Sep. 13, 2024. [Online]. Available: https://aclanthology.org/2024.eacl-long.52/

[3] Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics,pages 1477–1490,Dubrovnik, Croatia. Association for Computational Linguistics

[4] J. Devlin, M.-W. Chang, K. Lee, K. Google, and A. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019. Available: https://arxiv.org/pdf/1810.04805

[5] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale.” Available: https://arxiv.org/pdf/1911.02116

[6] Li, B., Han, L. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. In: Yin, H., et al. Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41278-3_74

[7] Lin, J. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 1991, 37, 145–151

[8] S. Xie, S. Santurkar, T. Ma, and P. Liang, “Data Selection for Language Models via Importance Resampling.” Accessed: Sep. 12, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6b9aa8f418bde2840d5f4ab7a02f663b-Paper-Conference.pdf

[9] N. He et al., “SoftDedup: an Efficient Data Reweighting Method for Speeding Up Language Model Pre-training.” Accessed: Sep. 12, 2024. [Online]. Availablehttps://arxiv.org/pdf/2407.06654

Reference [Ctd..]

[10] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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.

[11] Aloka Fernando, Nisansa de Silva, Menan Velayuthan, Charitha Rathnayake, and Surangika Ranathunga.2025. Improving the quality of web-mined parallel corpora of low-resource languages using debiasing heuristics. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28252–28269, Suzhou, China. Association for Computational Linguistics.

[12] Shravan Nayak, Surangika Ranathunga, Sarubi Thillainathan, Rikki Hung, Anthony Rinaldi, Yining Wang, Jonah Mackey, Andrew Ho, and En-Shiun Annie Lee. 2023. Leveraging auxiliary domain parallel data in intermediate task fine-tuning for low-resource translation. arXiv preprint arXiv:2306.01382.

[13] Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. arXiv preprint arXiv:2011.02821 (2020).

[14] Surangika Ranathunga, Nisansa de Silva, Aloka Fernando, Kavindu Warnakulasuriya, Isuru Wijesiri,Menan Velayuthan, Charitha Rathnayaka, Thivaharan Varatharajan, Sajeevi Silva, Piumi Kandanaarachchi, and 1 others. 2026. Ensita-a trilingual multi-domain parallel dataset and benchmark for domain-specific machine translation. arXiv preprint arXiv:2609.29511

[15] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. arXiv preprint arXiv:2402.14830.

[16] O. Gamage et al., "A Multilingual Dataset (MultiMWP) and Benchmark for Math Word Problem Generation," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 1838-1848, 2025, doi: 10.1109/TASLPRO.2025.3552936.

[17] Popović, Maja. "chrF++: words helping character n-grams." Proceedings of the second conference on machine translation. 2017.

Reference [Ctd..]

[18] Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318

[19]Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, Angela Fan; The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. Transactions of the Association for Computational Linguistics 2022; 10 522–538. doi: https://doi.org/10.1162/tacl_a_00474

[20] Popović, Maja. "chrF: character n-gram F-score for automatic MT evaluation." Proceedings of the tenth workshop on statistical machine translation. 2015.

[21] M. Ott et al., “FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling,” 2019. Available: https://aclanthology.org/N19-4009.pdf

[22] "Scaling neural machine translation to 200 languages." Nature 630, no. 8018 (2024): 841-846.

[23] D. Saunders, “Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey,” Journal of Artificial Intelligence Research, vol. 75, pp. 351–424, Sep. 2022, doi: https://doi.org/10.1613/jair.1.13566.

[24] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, A. Morcos, and M. Ai, “SemDeDup: Data-efficient learning at web-scale through semantic deduplication.” Accessed: Sep. 12, 2024. [Online]. Available: https://arxiv.org/pdf/2303.09540

[25] S. Ji, H. Raja Sinulingga, and D. Kwon, “Robust Guidance for Unsupervised Data Selection: Capturing Perplexing Named Entities for Domain-Specific Machine Translation.” Accessed: Sep. 12, 2024. [Online]. Available: https://arxiv.org/pdf/2402.19267

[26] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen, “LESS: Selecting Influential Data for Targeted Instruction Tuning.” Accessed: Sep. 12, 2024. [Online]. Available: https://arxiv.org/pdf/2402.04333

[27] Sohangir, S., Wang, D. Improved sqrt-cosine similarity measurement. J Big Data 4, 25 (2017). https://doi.org/10.1186/s40537-017-0083-6

Thanks !