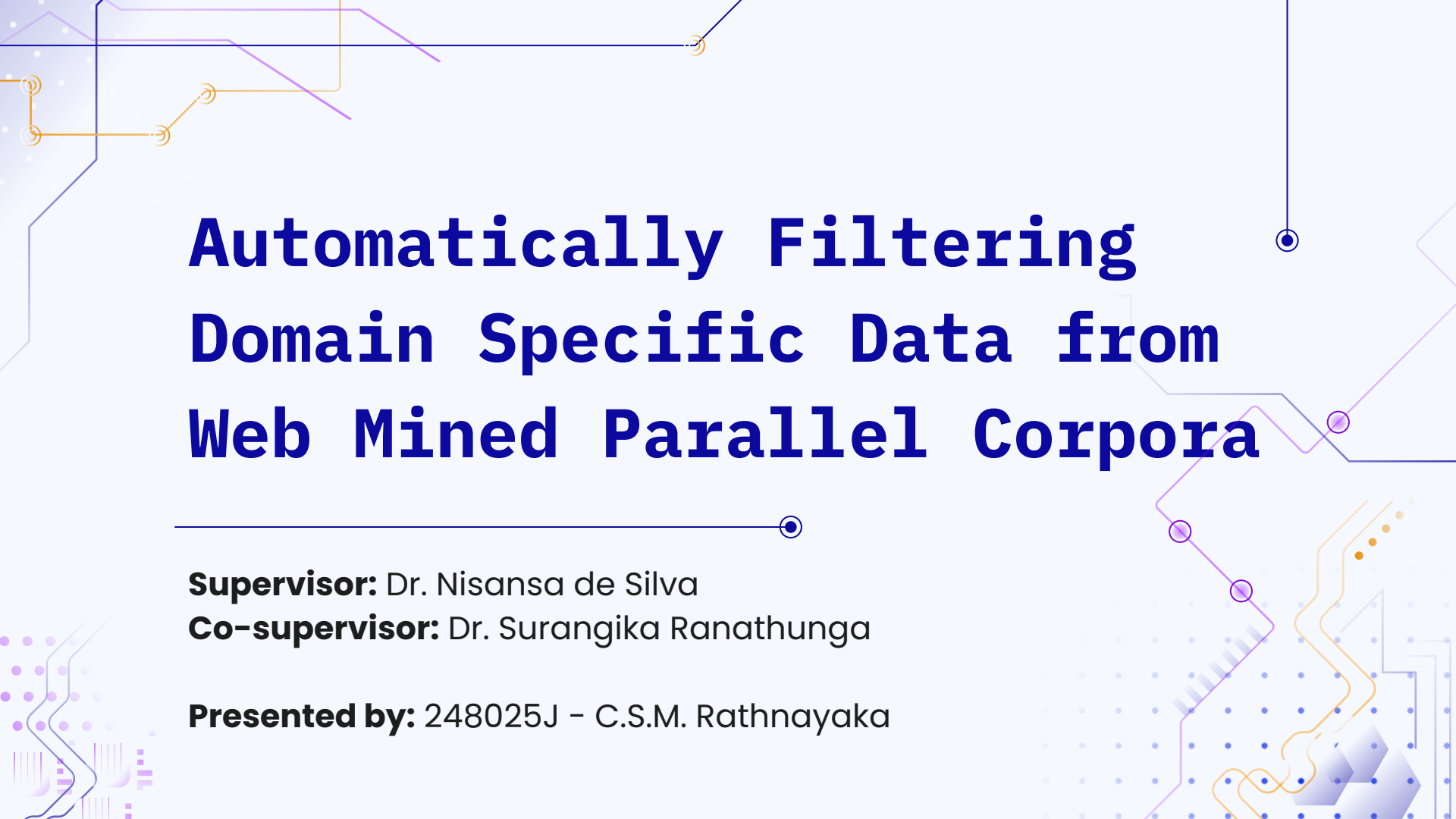


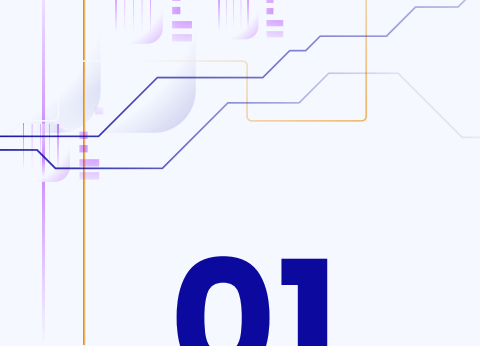


Automatically Filtering Domain Specific Data from Web Mined Parallel Corpora

Supervisor: Dr. Nisansa de Silva

Co-supervisor: Dr. Surangika Ranathunga

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



01

Introduction

Domain Adaptation



- **Training data** (Source Domain) does not exactly match with the **Testing data** (Target Domain)
- Lead to a drop in the model performance
- **Domain Adaptation** - The process of making a model perform well in new, unseen situations that differ from the environments it was originally trained on. [1]





02

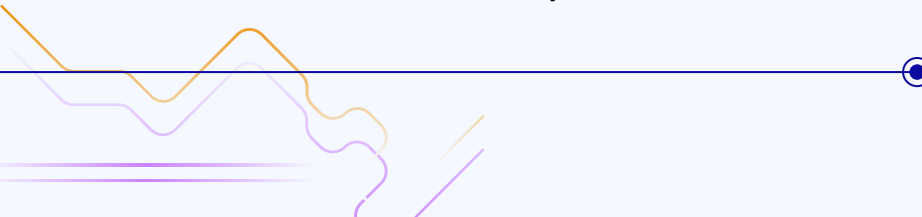
Research Problem & Objective



Problem

- Limited Domain-Specific Data for Low-Resource Languages
- Performance Gaps in NMT for Low-Resource Languages

Objective

- Generating Domain-Specific Datasets Without Human Intervention
 - Leveraging domain-specific data to enhance Neural Machine Translation performance.
- 



03

Related Work

Data Extraction

Data Selection for Language Models via Importance Resampling (DSIR) [2]

2023 - NeurIPS

The framework works by selecting important data points efficiently.

Distance Weighted Cosine Similarity Measure for Text Classification [7]

2013 - IDEAL

Weighted Cosine Similarity - An improved Cosine Similarity

Divergence measures based on the Shannon entropy[9]

1991 - IEEE

JSD

SoftDedup: an Efficient Data Reweighting Method for Speeding Up Language Model Pre-training [5]

2024

Commonness Score
Instead of removing duplicates, reduces the importance (or sampling weight) of data that appears very frequently

SemDeDup: Data-efficient learning at web-scale through semantic deduplication [3]

2023

a method which leverages embeddings from pre-trained models to identify and remove "semantic duplicates"

Robust Guidance for Unsupervised Data Selection: Capturing Perplexing Named Entities for Domain Specific Machine Translation [4]

2024

identifies data that should be selected, by assessing the perplexity of named entity tokens translated by a given pre-trained model

LESS: Selecting Influential Data for Targeted Instruction Tuning [6]

2024 ICML

Related to Instruction Tuning in LLMs

Improved sqrt-cosine similarity measurement [8]

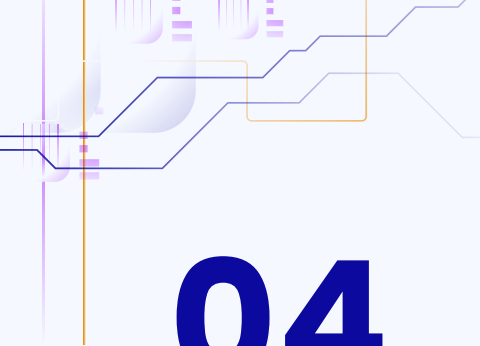
2017 - Big Data(J)

An improved Cosine Similarity

Explored Methodologies

- **Extracting domain-relevant sentences from the web-mined parallel corpus**
 - Basic Methods
 - Cosine Similarity with BOW²
 - Jensen-Shannon Divergence (JSD) [9]
 - Cosine Similarity with XLM-R [10]
 - BERT Classifier [11]
 - Advanced Methods
 - DSIR Method [2]
 - Weighted Cosine Similarity [7]
 - Commonness Method [5]
- **Translating Monolingual Domain Specific Corpus**
- **Combining Both Translating & Extracting Methods**

²https://en.wikipedia.org/wiki/Bag-of-words_model



04

Proposed Methodology

Datasets

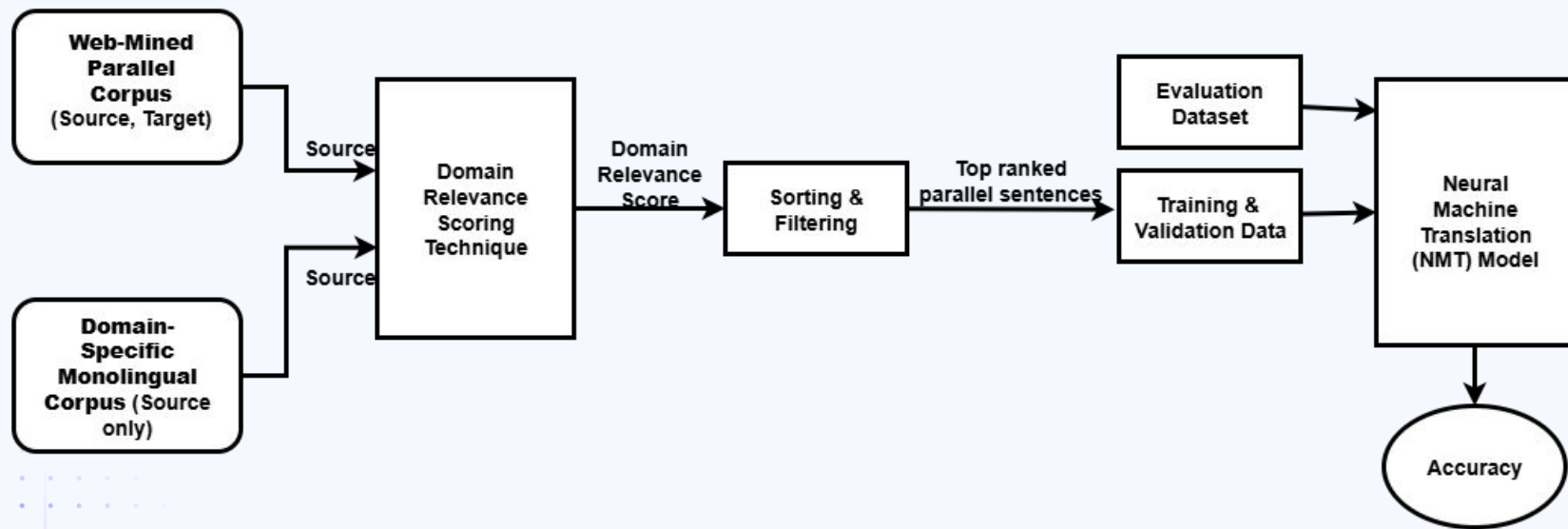
Webmined Parallel Corpus: CCMatrix [12] (English-Sinhala, English-Tamil)

Domain Specific Corpora(English):

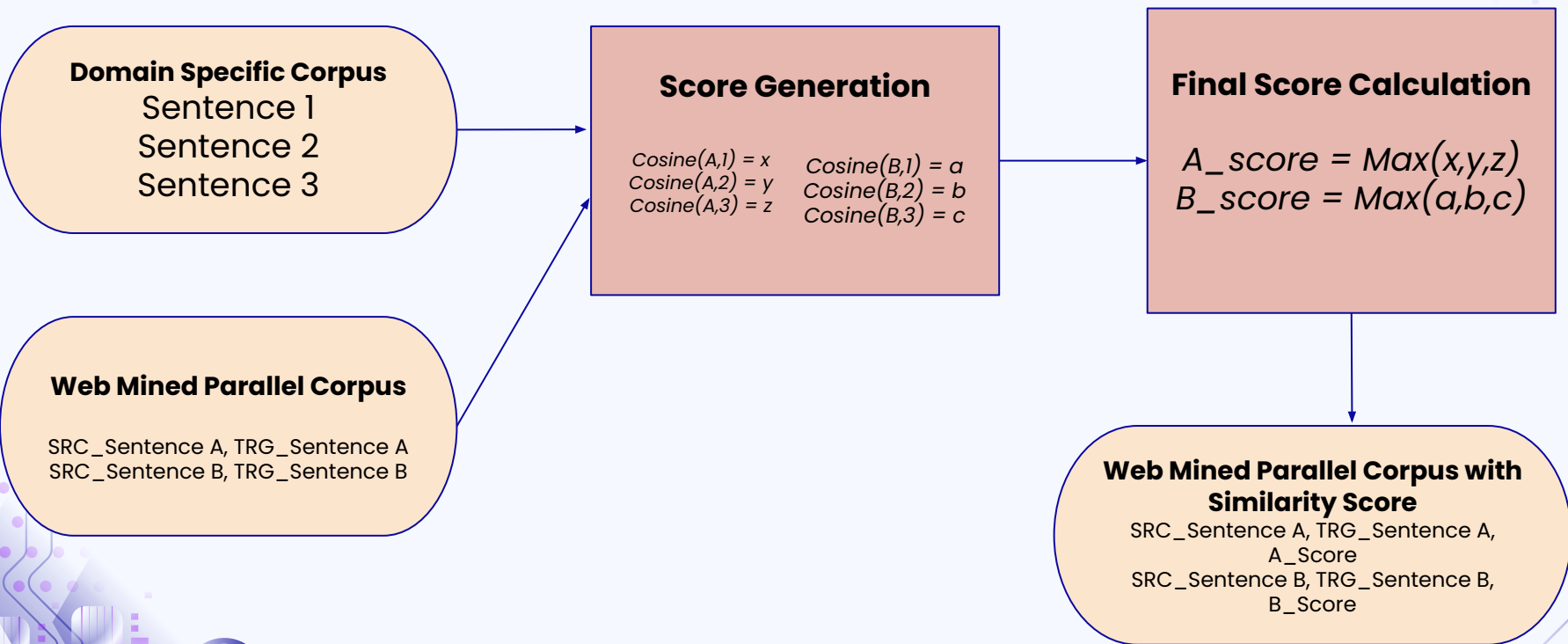
- Bible¹ (Dataset derived from Bible texts)
- Govt [13] (Dataset sourced from official government publications)
- Opensubtitles [14] (Dataset extracted from movie subtitles)
- Maths [15] (Dataset composed of mathematics questions and answers.)

¹ <https://huggingface.co/datasets/bible-nlp/biblenlp-corpus>

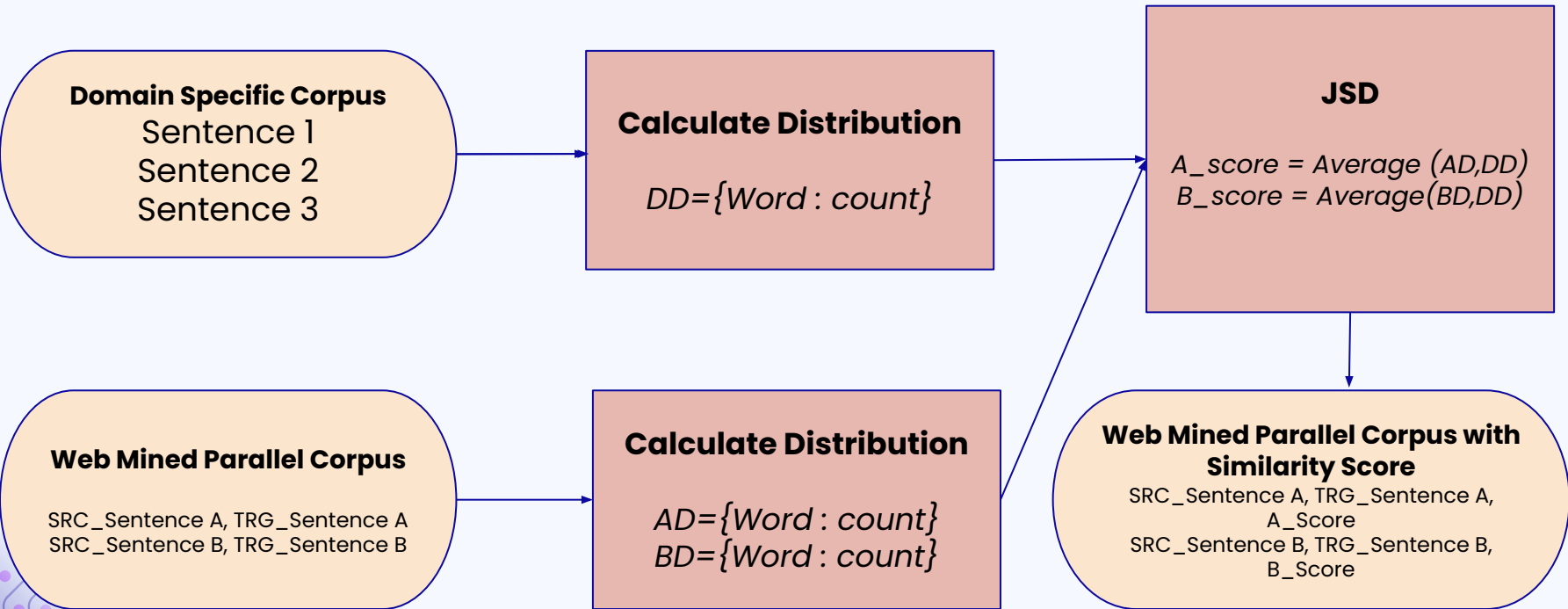
Method 1: Extracting domain-relevant sentences from the web-mined parallel corpus



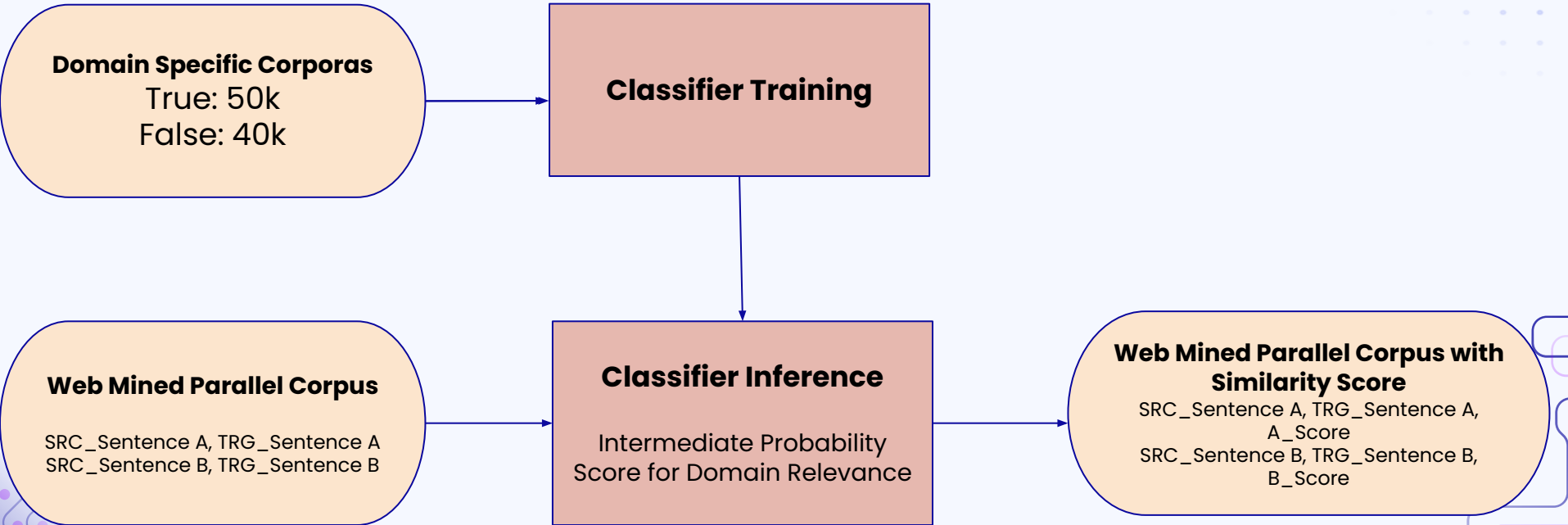
Generating Similarity Scores: Process Overview (Cosine Similarity)



Generating Similarity Scores: Process Overview (JSD)

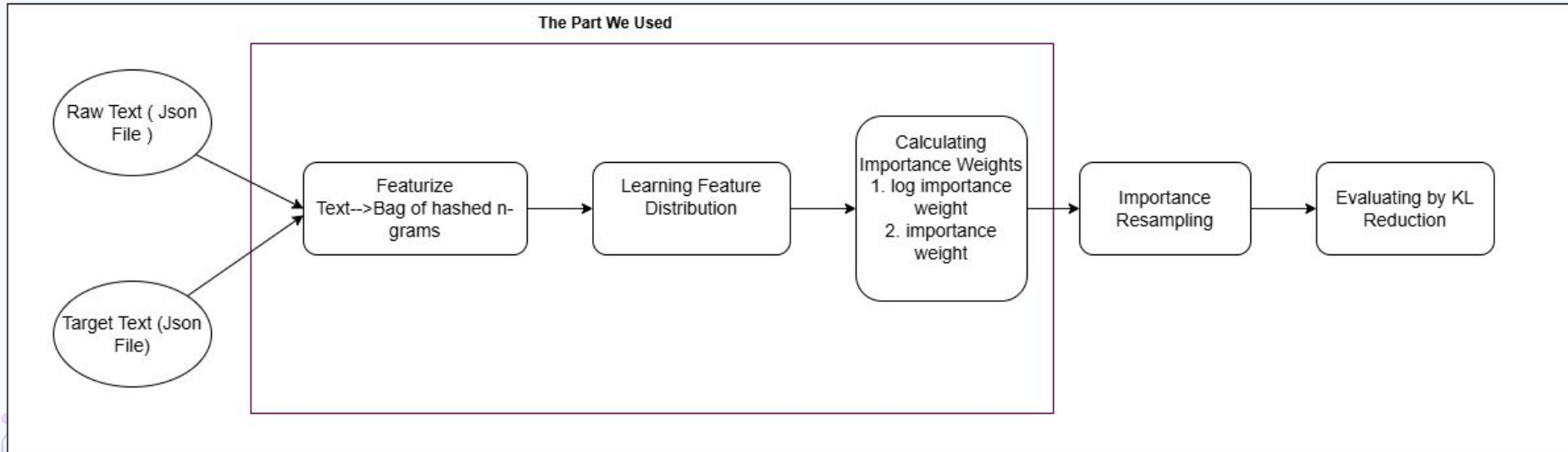


Generating Similarity Scores: Process Overview (Classifier)



Data Selection for Language Models via Importance Resampling (DSIR)[2]

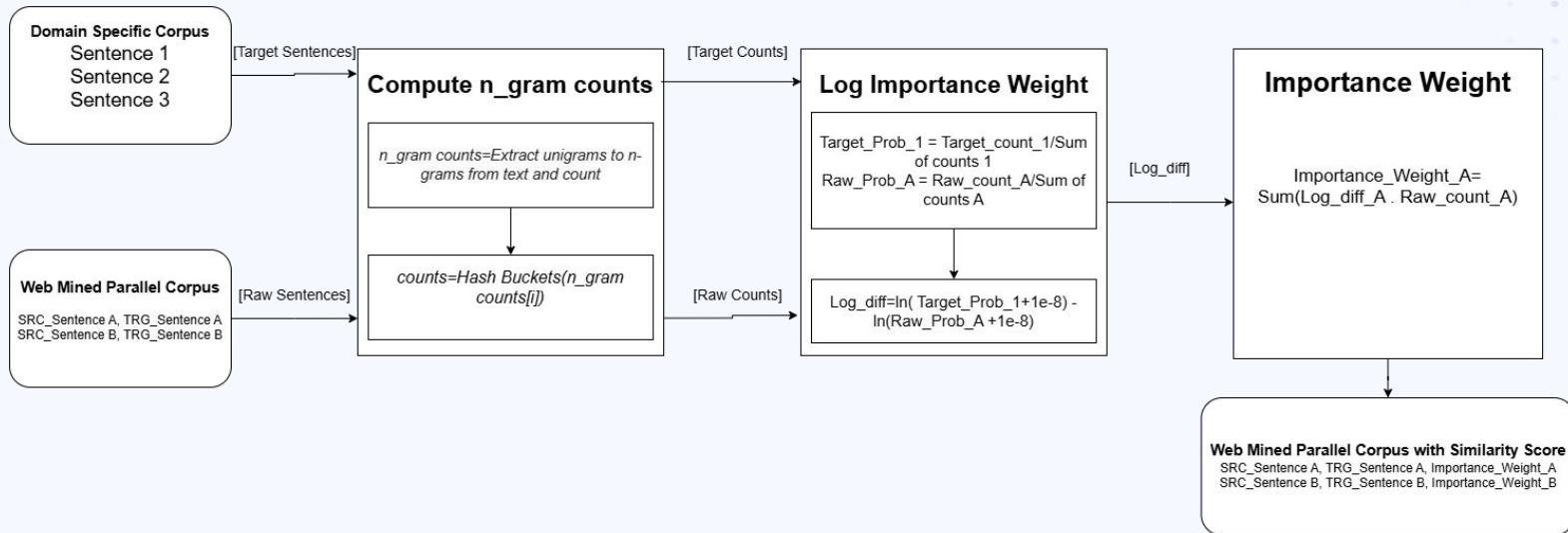
The Method in the Paper



[2] 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

Data Selection for Language Models via Importance Resampling (DSIR)[2] [Ctd...]



Sample Target_Counts = $[[0,1,2,0,4,1,6,0,\dots],[2,3,5,7,8,2,3,0,1,\dots],\dots]$

Discriminative DSIR

Method

Web Mined Parallel Corpus with Similarity Score

SRC_Sentence A, TRG_Sentence A,
A_Score
SRC_Sentence B, TRG_Sentence B,
B_Score

Importance Weight(IW)

$IW_A = A_score / 1 - A_score$
 $IW_B = B_Score / 1 - B_Score$

Final Score

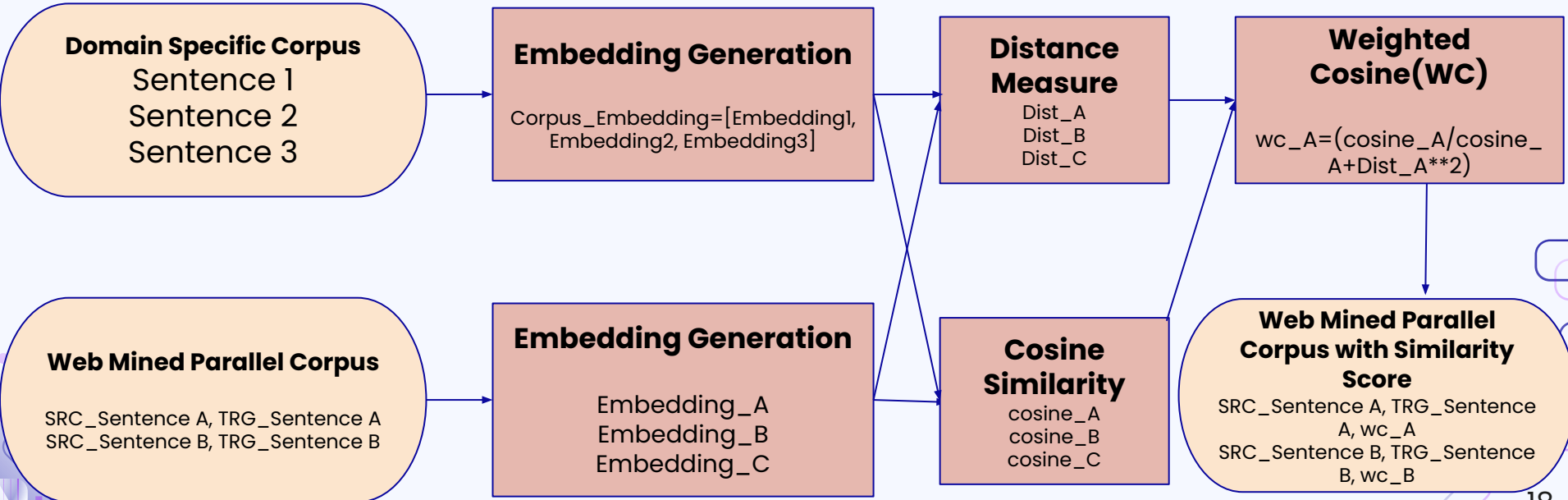
$A'_Score = A_Score * IW_A$
 $B'_Score = B_score * IW_B$

- Since we got better results for Cosine Similarity, We initiated this method with Cosine(XLMR)

Web Mined Parallel Corpus with Similarity Score

SRC_Sentence A, TRG_Sentence A,
A'_Score
SRC_Sentence B, TRG_Sentence B,
B'_Score

Weighted Cosine Similarity[7]

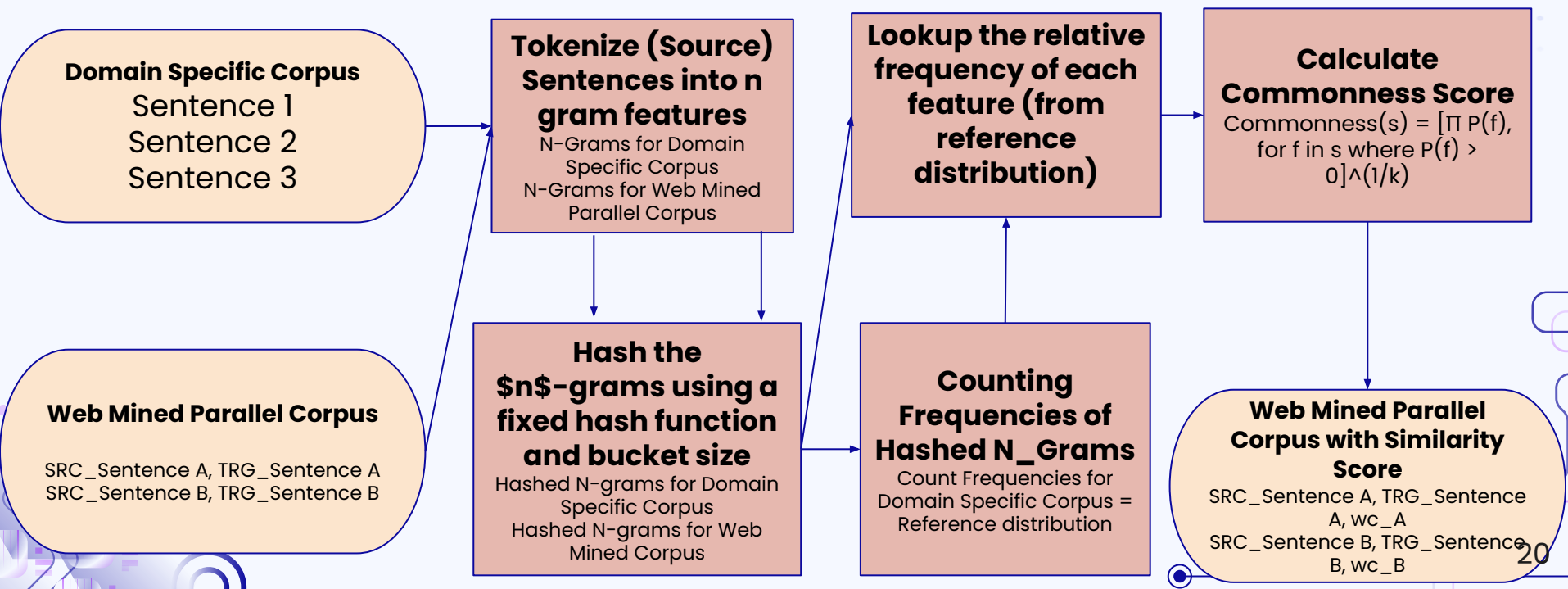


Weighted Cosine Similarity[7] [Ctd...]

- As the Distance Measure, We tried Euclidean and JSD

[7] 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

Commonness Method



Evaluation - Basic Configurations

Evaluation Datasets:

- Bible
- Govt
- Opensubtitles
- Maths [18]

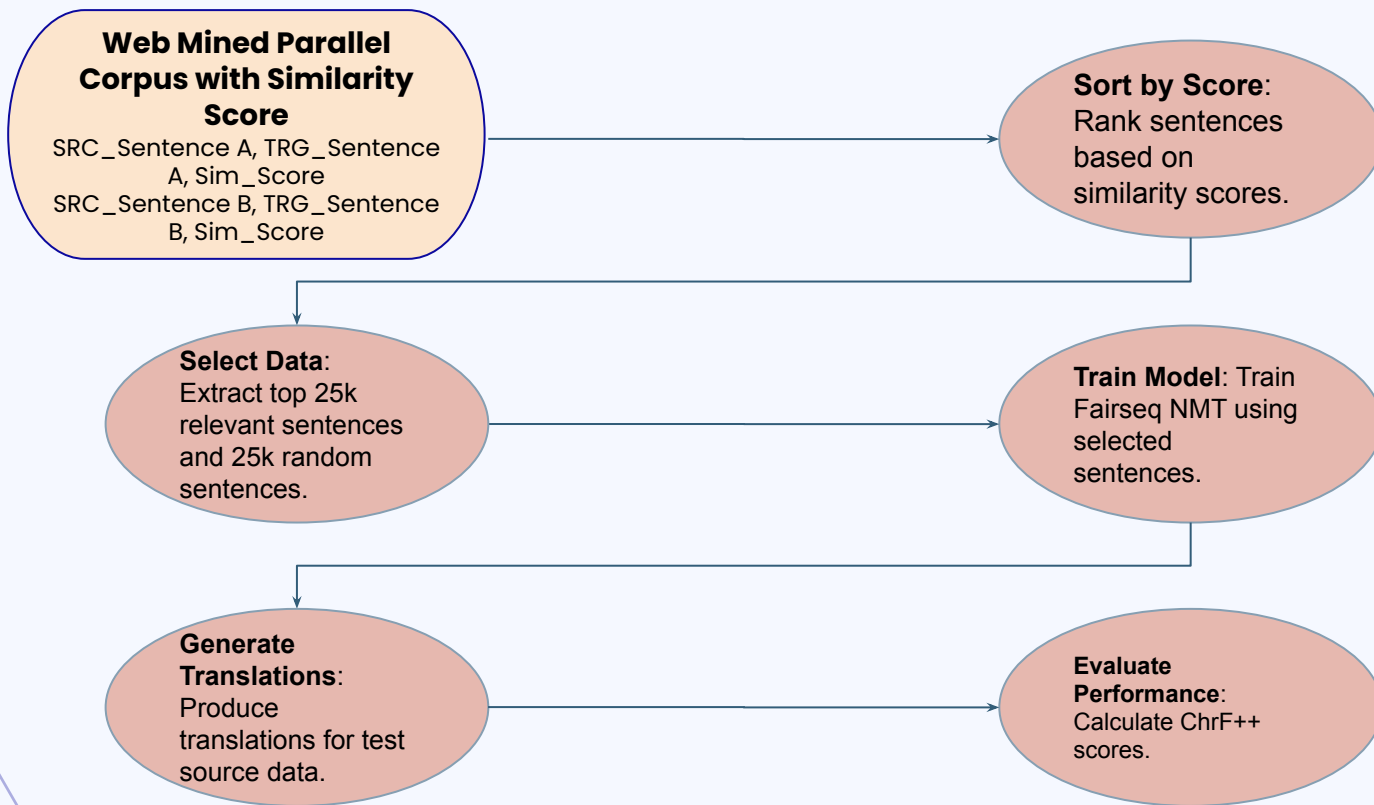
Evaluation Metric

- Chrf++ [19]

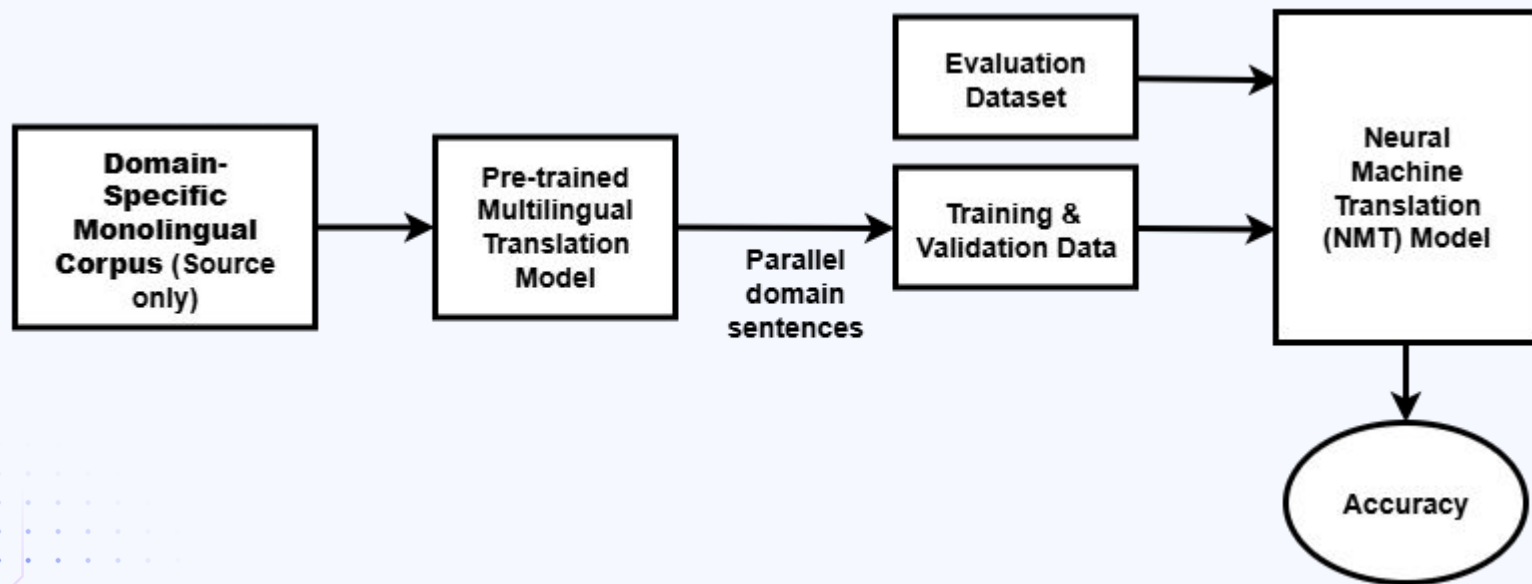
Evaluation Task:

- NMT - **Fairseq** [16].

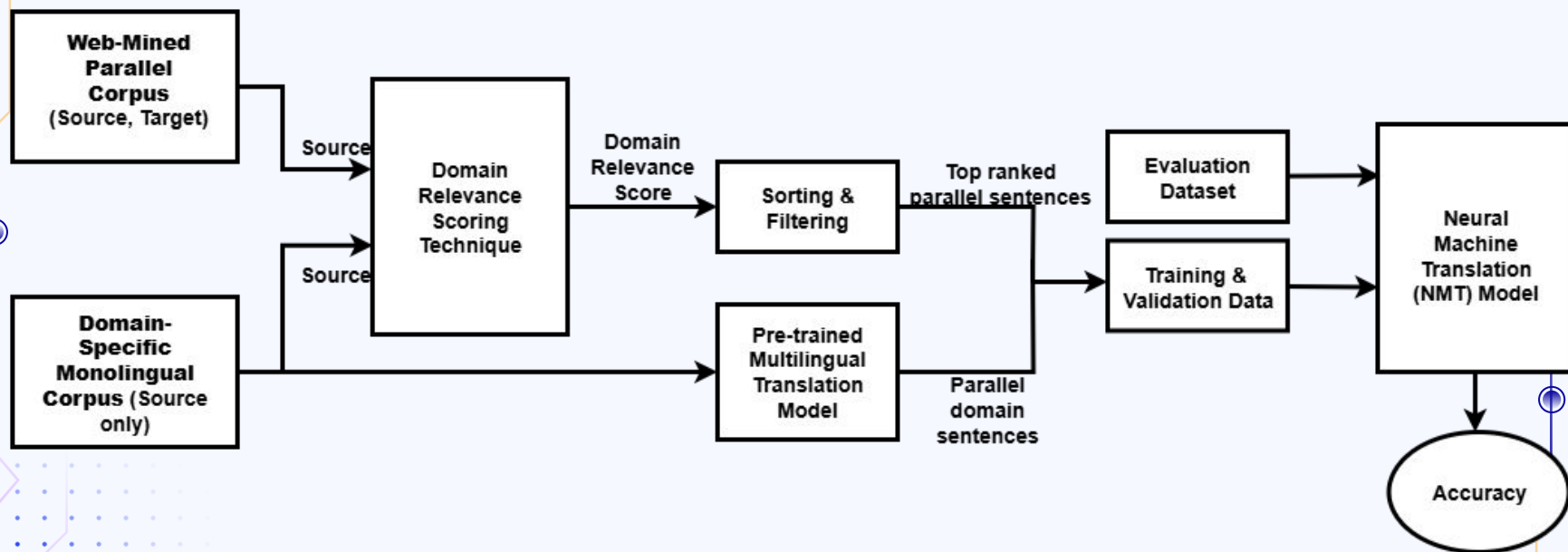
Evaluation - Process



Method 2: Translating Monolingual Domain Specific Corpus



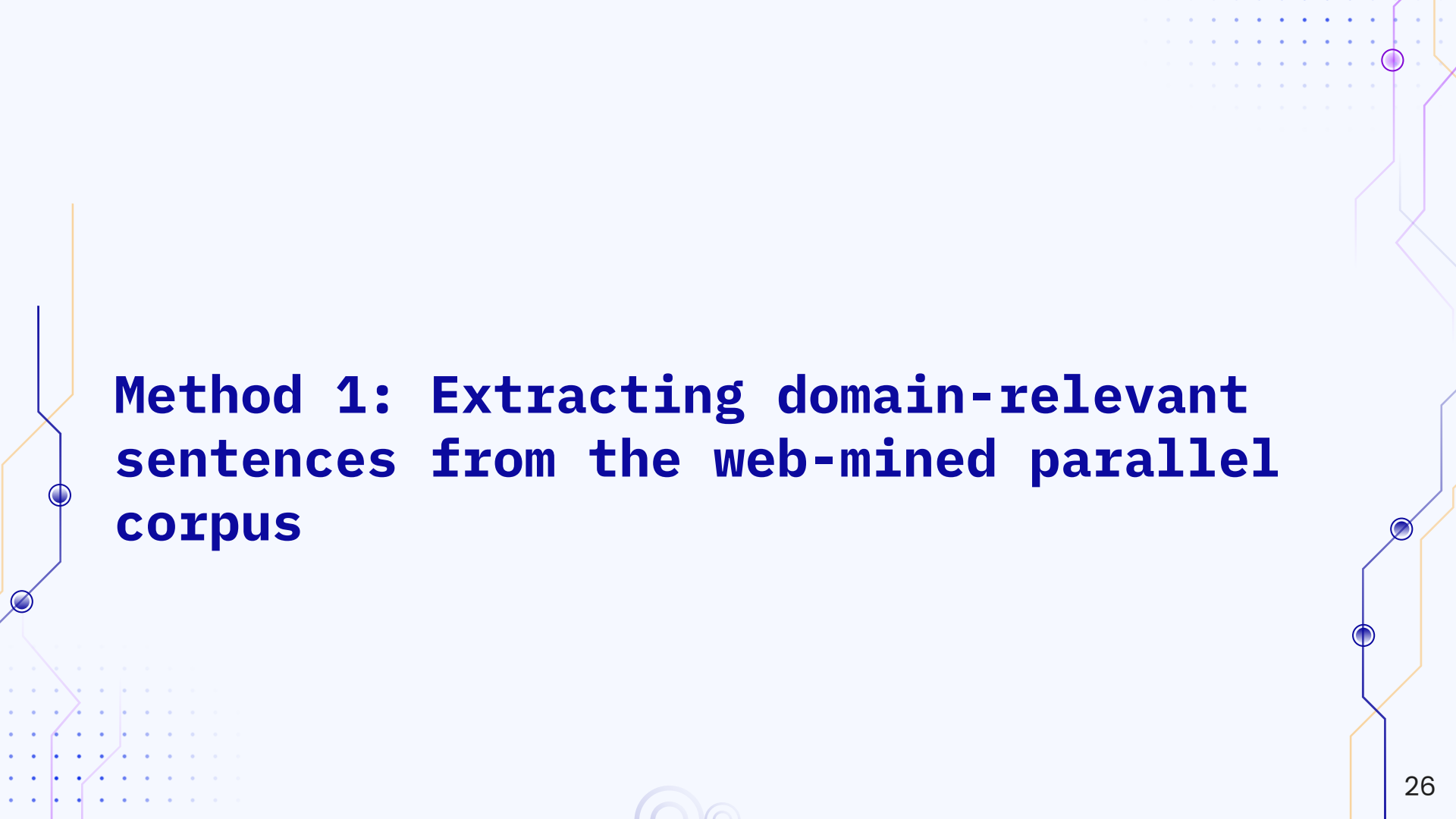
Method 3: Combining Both Translating & Extracting Methods





05

Results



Method 1: Extracting domain-relevant sentences from the web-mined parallel corpus

Results for English-Sinhala Corpus

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Bible	14.1
		random25k	Bible	7
Bible	Cosine Similarity BOW	top25k	Bible	13.8
		random25k	Bible	7.6
	JSD	top25k	Bible	15.9
		random25k	Bible	8.8
	Cosine Similarity XLMR	top25k	Bible	17
		random25k	Bible	8.6
	Classifier	top25k	Bible	9.9
		random25k	Bible	6.8
	Weighted Cosine Similarity (Euclidean)	top25k	Bible	18.1
		random25k	Bible	8.3
	Weighted Cosine Similarity (JSD)	top25k	Bible	1.9
		random25k	Bible	7.8

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Bible	14.1
		random25k	Bible	7
Bible	DSIR_N_GRAM_2 (Default)	top25k	Bible	14.5
		random25k	Bible	15.3
	DSIR_N_GRAM_3	top25k	Bible	15.7
		random25k	Bible	8.6
	DSIR_N_GRAM_4	top25k	Bible	15.6
		random25k	Bible	8.2
	DSIR_N_GRAM_5	top25k	Bible	15.8
		random25k	Bible	6
	Discriminative_DSIR	top25k	Bible	18.8
		random25k	Bible	7.3
	Discriminative_DSIR + Weighted Cosine	top25k	Bible	16.4
		random25k	Bible	7.2
	Commonness Scoring	top25k	Bible	15.8
		random25k	Bible	8
	Commonness Scoring + DSIR	top25k	Bible	15.8
		random25k	Bible	8.3
	Commonness Scoring + Discriminative DSIR	top25k	Bible	18.3
		random25k	Bible	8.1

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Govt	22.4
		random25k	Govt	4.6
Govt	Cosine Similarity BOW	top25k	Govt	11
		random25k	Govt	6.2
	JSD	top25k	Govt	19.4
		random25k	Govt	7.4
	Cosine Similarity XLMR	top25k	Govt	23
		random25k	Govt	6
	Classifier	top25k	Govt	10.6
		random25k	Govt	5.7
	Weighted Cosine Similarity (Euclidean)	top25k	Govt	21.5
		random25k	Govt	6
	Weighted Cosine Similarity (JSD)	top25k	Govt	2.7
		random25k	Govt	6.9
	DSIR_N_GRAM_2 (Default)	top25k	Govt	19.9
		random25k	Govt	5
	DSIR_N_GRAM_3	top25k	Govt	19.6
		random25k	Govt	6.4

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Govt	22.4
		random25k	Govt	4.6
Govt	DSIR_N_GRAM_4	top25k	Govt	21.8
		random25k	Govt	4.7
	DSIR_N_GRAM_5	top25k	Govt	15.8
		random25k	Govt	6
	Discriminative_DSIR	top25k	Govt	23.8
		random25k	Govt	5.3
	Discriminative_DSIR + Weighted Cosine	top25k	Govt	22.7
		random25k	Govt	6.9
	Commonness Scoring	top25k	Govt	21.3
		random25k	Govt	5.2
	Commonness Scoring + DSIR	top25k	Govt	21.3
		random25k	Govt	5.2
	Commonness Scoring + Discriminative DSIR	top25k	Govt	23.6
		random25k	Govt	6

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	OpenSubtitles	22.2
		random25k	OpenSubtitles	9.7
OpenSubtitles	JSD	top25k	OpenSubtitles	34
		random25k	OpenSubtitles	29
	Cosine Similarity XLMR	top25k	OpenSubtitles	31
		random25k	OpenSubtitles	25
	Classifier	top25k	OpenSubtitles	24
		random25k	OpenSubtitles	16
	Weighted Cosine Similarity (Euclidean)	top25k	OpenSubtitles	21
		random25k	OpenSubtitles	26
	DSIR_N_GRAM_2 (Default)	top25k	OpenSubtitles	28
		random25k	OpenSubtitles	25
	DSIR_N_GRAM_3	top25k	OpenSubtitles	25
		random25k	OpenSubtitles	26

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	OpenSubtitles	22.2
		random25k	OpenSubtitles	9.7
OpenSubtitles	DSIR_N_GRAM_4	top25k	OpenSubttitles	26
		random25k	OpenSubttitles	26
	DSIR_N_GRAM_5	top25k	OpenSubttitles	29
		random25k	OpenSubttitles	25
	Discriminative_DSIR	top25k	OpenSubttitles	34
		random25k	OpenSubttitles	25
	Discriminative_DSIR + Weighted Cosine	top25k	OpenSubttitles	21
		random25k	OpenSubttitles	25
	Commonness Scoring	top25k	OpenSubttitles	25
		random25k	OpenSubttitles	27
	Commonness Scoring + DSIR	top25k	OpenSubttitles	28
		random25k	OpenSubttitles	26
	Commonness Scoring + Discriminative DSIR	top25k	OpenSubttitles	26
		random25k	OpenSubttitles	25

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Maths	42
		random25k	Maths	13
Maths	JSD	top25k	Maths	29
		random25k	Maths	18
	Cosine Similarity XLMR	top25k	Maths	33
		random25k	Maths	12
	Classifier	top25k	Maths	21
		random25k	Maths	12
	Weighted Cosine Similarity (Euclidean)	top25k	Maths	18.9
		random25k	Maths	13.7
	DSIR_N_GRAM_2 (Default)	top25k	Maths	23
		random25k	Maths	20
	DSIR_N_GRAM_3	top25k	Maths	24
		random25k	Maths	19

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Maths	42
		random25k	Maths	13
OpenSubtitles	DSIR_N_GRAM_4	top25k	Maths	27
		random25k	Maths	17
	DSIR_N_GRAM_5	top25k	Maths	25
		random25k	Maths	18
	Discriminative_DSIR	top25k	Maths	31
		random25k	Maths	18
	Discriminative_DSIR + Weighted Cosine	top25k	Maths	12.9
		random25k	Maths	18
	Commonness Scoring	top25k	Maths	12.9
		random25k	Maths	15.5
	Commonness Scoring + DSIR	top25k	Maths	27.9
		random25k	Maths	15.9
	Commonness Scoring + Discriminative DSIR	top25k	Maths	21.3
		random25k	Maths	15.6

Results for English-Tamil Corpus

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Bible	14.4
		random25k	Bible	10.5
Bible	Cosine Similarity BOW	top25k	Bible	13.8
		random25k	Bible	9.9
	JSD	top25k	Bible	14.2
		random25k	Bible	10.2
	Cosine Similarity XLMR	top25k	Bible	15.5
		random25k	Bible	10.6
	Classifier	top25k	Bible	11.3
		random25k	Bible	9.2
	Weighted Cosine Similarity (Euclidean)	top25k	Bible	14.8
		random25k	Bible	10.9
	Weighted Cosine Similarity (JSD)	top25k	Bible	6.4
		random25k	Bible	8.8

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Bible	14.4
		random25k	Bible	10.5
Bible	DSIR_N_GRAM_2 (Default)	top25k	Bible	13.7
		random25k	Bible	10.5
	DSIR_N_GRAM_3	top25k	Bible	0.9
		random25k	Bible	0.9
	DSIR_N_GRAM_4	top25k	Bible	13.9
		random25k	Bible	10.5
	DSIR_N_GRAM_5	top25k	Bible	13.6
		random25k	Bible	10.7
	Discriminative_DSIR	top25k	Bible	15.4
		random25k	Bible	11.4
	Discriminative_DSIR + Weighted Cosine	top25k	Bible	15.5
		random25k	Bible	9.3
	Commonness Scoring	top25k	Bible	13.4
		random25k	Bible	10.4
	Commonness Scoring + DSIR	top25k	Bible	13.4
		random25k	Bible	9.5
	Commonness Scoring + Discriminative DSIR	top25k	Bible	11.6
		random25k	Bible	9.7

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Govt	8.8
		random25k	Govt	5.2
Govt	Cosine Similarity BOW	top25k	Govt	6.4
		random25k	Govt	6.8
	JSD	top25k	Govt	8.1
		random25k	Govt	5.9
	Cosine Similarity XLMR	top25k	Govt	9.2
		random25k	Govt	5.3
	Classifier	top25k	Govt	8
		random25k	Govt	3.1
	Weighted Cosine Similarity (Euclidean)	top25k	Govt	7.7
		random25k	Govt	6.7
	Weighted Cosine Similarity (JSD)	top25k	Govt	2.7
		random25k	Govt	6.6
	DSIR_N_GRAM_2 (Default)	top25k	Govt	10.6
		random25k	Govt	6.2

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Govt	8.8
		random25k	Govt	5.2
Govt	DSIR_N_GRAM_3	top25k	Govt	0
		random25k	Govt	0.6
	DSIR_N_GRAM_4	top25k	Govt	9.8
		random25k	Govt	6.4
	DSIR_N_GRAM_5	top25k	Govt	10.4
		random25k	Govt	4.8
	Discriminative_DSIR	top25k	Govt	8.5
		random25k	Govt	6.5
	Discriminative_DSIR + Weighted Cosine	top25k	Govt	9.6
		random25k	Govt	6.7
	Commonness Scoring	top25k	Govt	8.9
		random25k	Govt	6.7
	Commonness Scoring + DSIR	top25k	Govt	8.9
		random25k	Govt	6.1
	Commonness Scoring + Discriminative DSIR	top25k	Govt	9.4
		random25k	Govt	7.1

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	OpenSubtitles	21
		random25k	OpenSubtitles	14.1
OpenSubtitles	Cosine Similarity XLMR	top25k	OpenSubttitles	38.4
		random25k	OpenSubttitles	34.3
	Classifier	top25k	OpenSubttitles	32.2
		random25k	OpenSubttitles	21.7
	Weighted Cosine Similarity (Euclidean)	top25k	OpenSubttitles	31.4
		random25k	OpenSubttitles	37.8
	DSIR_N_GRAM_2 (Default)	top25k	OpenSubttitles	38.4
		random25k	OpenSubttitles	32.7

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	OpenSubtitles	21
		random25k	OpenSubtitles	14.1
OpenSubtitles	DSIR_N_GRAM_3	top25k	OpenSubttitles	27.9
		random25k	OpenSubttitles	36.1
	DSIR_N_GRAM_4	top25k	OpenSubttitles	29.7
		random25k	OpenSubttitles	33.5
	DSIR_N_GRAM_5	top25k	OpenSubttitles	29.2
		random25k	OpenSubttitles	33.4
	Discriminative_DSIR	top25k	OpenSubttitles	36.3
		random25k	OpenSubttitles	35.1
	Discriminative_DSIR + Weighted Cosine	top25k	OpenSubttitles	27.1
		random25k	OpenSubttitles	36.3
	Commonness Scoring	top25k	OpenSubttitles	31.7
		random25k	OpenSubttitles	33.6
	Commonness Scoring + DSIR	top25k	OpenSubttitles	38.3
		random25k	OpenSubttitles	35
	Commonness Scoring + Discriminative DSIR	top25k	OpenSubttitles	32
		random25k	OpenSubttitles	32.2

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Maths	30.9
		random25k	Maths	22.9
OpenSubtitles	Cosine Similarity XLMR	top25k	Maths	34
		random25k	Maths	26
	Classifier	top25k	Maths	26
		random25k	Maths	15
	Weighted Cosine Similarity (Euclidean)	top25k	Maths	21
		random25k	Maths	18
	DSIR_N_GRAM_2 (Default)	top25k	Maths	24
		random25k	Maths	25

Score				Chrf++
Dataset			Evaluation Set	
Baseline	LASER 3 [17]	top25k	Maths	30.9
		random25k	Maths	22.9
OpenSubtitles	DSIR_N_GRAM_3	top25k	Maths	27
		random25k	Maths	26
	DSIR_N_GRAM_4	top25k	Maths	29
		random25k	Maths	23
	DSIR_N_GRAM_5	top25k	Maths	27
		random25k	Maths	25
	Discriminative_DSIR	top25k	Maths	28
		random25k	Maths	23
	Discriminative_DSIR + Weighted Cosine	top25k	Maths	17
		random25k	Maths	25
	Commonness Scoring	top25k	Maths	16
		random25k	Maths	23
	Commonness Scoring + DSIR	top25k	Maths	24
		random25k	Maths	23
	Commonness Scoring + Discriminative DSIR	top25k	Maths	28
		random25k	Maths	26

Method 2 & 3: Translating Monolingual Domain Specific Corpus & Combining Both Translating & Extracting Methods



Dataset	chrf++
Extracted 25k(Method 1)	18.8
Original Bible (47546) (Method 2)	27.5
Original Bible + Extracted 25k	27.5
Original Bible + Extracted 50k	28.7
Original Bible + Extracted 75k	28.1
Original Bible + Extracted 100k	24.3



Dataset	chrF++
Extracted 25k	23.8
Original Govt (53319)	30.8
Original Govt + Extracted 25k	30.8
Original Govt + Extracted 50k	35.3
Original Govt + Extracted 75k	35.1
Original Govt + Extracted 100k	34.3



Dataset	chrF++
Extracted 25k	34
Original Opensubtitles 50k	46
Original Opensubtitles 50k + Extracted 25k	51.2
Original OpenSubtitles 50k + Extracted 50k	54.6
Original Opensubtitles 50k + Extracted 75k	52.7
Original Opensubtitles 50k + Extracted 100k	54.5



Dataset	chrF++
Extracted 25k	34
Original Opensubtitles 100k	53.1
Original Opensubtitles 100k + Extracted 25k	54.8
Original OpenSubtitles 100k + Extracted 50k	56.2
Original Opensubtitles 100k + Extracted 75k	54.8
Original Opensubtitles 100k + Extracted 100k	55.7



Dataset	chrF++
Extracted 25k	33
Original Maths 50k	48.9
Original Maths 50k + Extracted 25k	52.7
Original Maths 50k + Extracted 50k	54.7
Original Maths 50k + Extracted 75k	56.2
Original Maths 50k + Extracted 100k	55



Dataset	chrF++
Extracted 25k	33
Original Maths 100k	54.5
Original Maths 100k + Extracted 25k	54
Original Maths 100k + Extracted 50k	54.8
Original Maths 100k + Extracted 75k	55.1
Original Maths 100k + Extracted 100k	56.6

Progress Since Last Review

New Approaches Implemented

- Commonness Score
- Commonness Score + DSIR/Discriminative DSIR
- Discriminative_DSIR + Weighted Cosine Similarity

Solving the issue in OpenSubtitles Domain and Re-Evaluation for EnSi & EnTa

Optimizing codes for large datasets (eg: Opensubtitles, Maths)

Evaluation in Maths Domain for EnSi & EnTa

Extending the Experiments into the Translation Method and Combined Method



Future Work

- Experimenting with CCAAlign Dataset
- Expanding the dataset size to 50k for further experiments

Reference

- [1] 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>.
- [2] 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
- [3] 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>
- [4] 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>

Reference [Ctd..]

- [5] N. He et al., “SoftDedup: an Efficient Data Reweighting Method for Speeding Up Language Model Pre-training.” Accessed: Sep. 12, 2024. [Online]. Available: <https://arxiv.org/pdf/2407.06654>
- [6] 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>
- [7] 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
- [8] Sohangir, S., Wang, D. Improved sqrt-cosine similarity measurement. J Big Data 4, 25 (2017). <https://doi.org/10.1186/s40537-017-0083-6>
- [9] Lin, J. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 1991, 37, 145–151
- [10] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale.” Available: <https://arxiv.org/pdf/1911.02116>

Reference [Ctd..]

[11] 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>

[12] 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.

[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] P. Lison and J. Tiedemann, “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles,” 2016. Available: http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf

[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.

Reference [Ctd..]

[16] M. Ott et al., "FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling," 2019. Available: <https://aclanthology.org/N19-4009.pdf>

[17] 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/>

[18] 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.

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

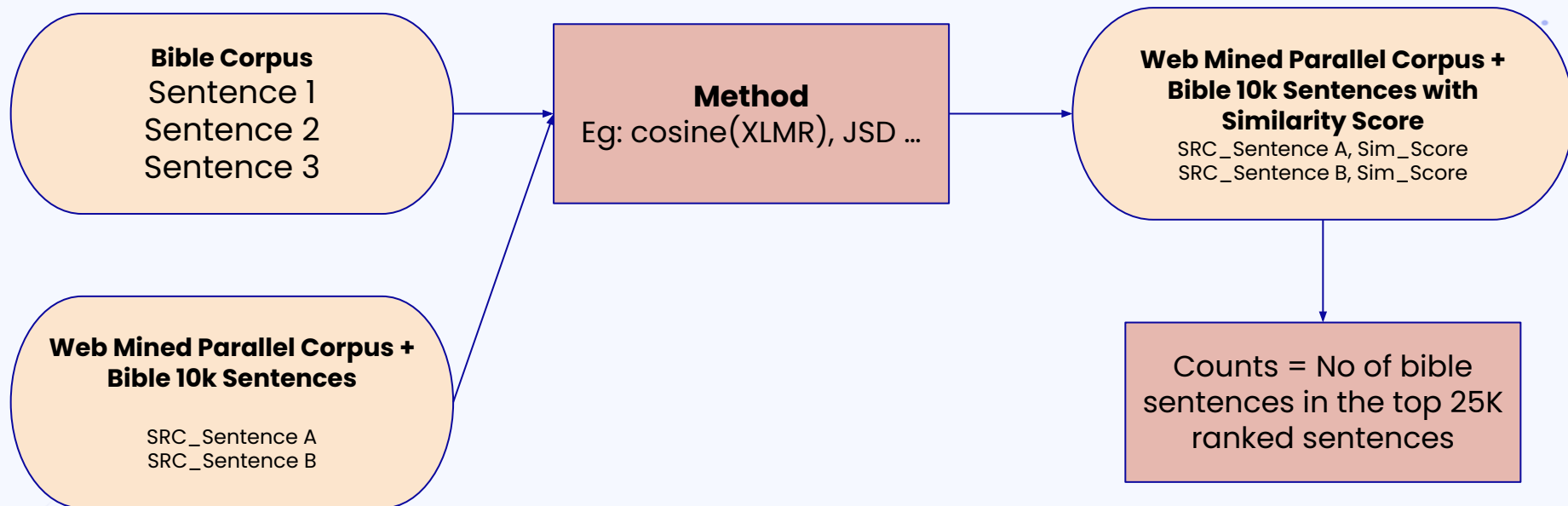


Thanks !

Classifiers

train_bs, valid_bs, lr, epochs				
Classifier - train 50k	5749		Classifier - train 90k	7046
classifier_32_32_5e-05_2	2962		classifier_train90k_32_32_5e-05_2	7765
classifier_32_32_5e-05_3	2962		classifier_train90k_32_32_5e-05_3	7754
classifier_32_32_5e-05_4	7436		classifier_train90k_32_32_5e-05_4	6752
classifier_32_32_3e-05_2	7626		classifier_train90k_32_32_3e-05_2	7413
classifier_32_32_3e-05_3	7778		classifier_train90k_32_32_3e-05_3	5550
classifier_32_32_3e-05_4	6486		classifier_train90k_32_32_3e-05_4	8222
classifier_32_32_3e-05_2	7644		classifier_train90k_32_32_3e-05_2	4760

Code Validation



sqrt cosine	19
SemDedup	67