

Evaluating Effect of Training Corpus Size on Performance of Statistical Translation Systems

Chetan Agarwal*, Kamlesh Dutta\$, Pardeep Singh**

chetanswadeshi@gmail.com, \$kd@nith.ac.in, #pardeep@nith.ac.in
Computer Science and Engineering Department,
National Institute of Technology, Hamirpur (H.P.), INDIA

Abstract— While the performance of Statistical Machine Translation (SMT) systems can be influenced by numerous factors, the size of the training corpus remains a foundational determinant in shaping the learning behavior and eventual translation quality of such systems. This paper reaffirms the significance of training data volume in the context of Phrase-Based Statistical Machine Translation (PB-SMT) and investigates the impact of varying training corpus sizes on translation performance using MOSES, a well-established SMT framework. Through this empirical lens, this research work explicitly contributes to: (i) building a structured experimental framework to evaluate corpus size influence, (ii) evaluate quantitative performance trends across multiple metrics. Outcomes presented not only validate existing beliefs about the importance of data scale but also serve as a diagnostic foundation for scaling SMT systems in under-resourced language contexts. Observations suggested a logarithmic growth in translation quality with increased training data size.

Index Terms— Natural Language Processing, Translation Model, Statistical Machine Translation, MOSES.

I. INTRODUCTION

MACHINE Translation (MT), which aims to translate a text from one language (source language) to another language (target language) with or without human assistance has emerged as one of the very thriving and important subfields of Natural language Processing (NLP) in last two decades. According to computational methodology employed in translation process, all approaches are classified in two major classes: Rule Based Machine Translation (RBMT), and Corpus Based Machine Translation (CBMT). Due to static nature, requirement of enormous input and assistance by human experts, RBMT are almost obsolete. Corpus-Based Machine Translation is an alternative approach or methodology of MT, which automatically extracts knowledge by analyzing translation examples from a parallel corpus built by human experts. There are two types of CBMT: “Statistical Machine Translation” (SMT) and Example-Based Machine Translation (EBMT). SMT focuses on generating many thousands of hypothetical translations for the input string and then working out which one of those is most likely [1]. Statistical SMT has long been a dominant approach to automated translation.

MOSES [2] is one such very prominent open-source toolkit that works like a multi-lingual translation platform built upon SMT approach. MOSES consists of a very large size parallel corpus for a variety of source and target language pairs and systematically implements up-to-date solutions in the field of SMT and allows us to build multiple SMT models. Once a translation model is trained for an object language pair, it can produce the translated text for given input text. Translation quality of this output can be evaluated by using any of the prevalent automatic evaluation metrics like BLEU, TER, METEOR etc.

This paper is aimed to (i) build translation system for English to Hindi translation using MOSES platform with varying size of training sets, and (ii) analyze the influence of size of training sets on translation quality. Being in line with the above-mentioned objectives, the rest of the paper has been organized as follows.

Section 2 insides the details about datasets along-with experimental setup and Section 3

presents results analysis. Section 4 finally concludes the paper.

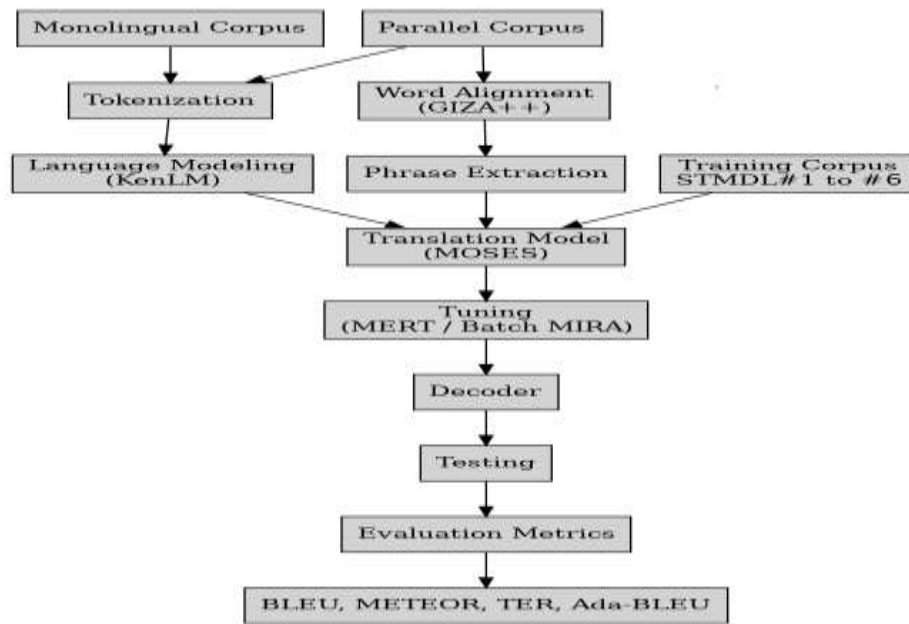
II. EXPERIMENTAL SETUP

According to the objective of this experiment, “study and analyze the impact of training corpus size on translation quality of an SMT”, translation system was trained with different and varying size of training sets and then were evaluated with different automatic evaluation metric. This section provides complete details about the experimental set up upon which system was trained, and experiments were conducted. The overall approach is shown in figure 1.

A. Corpus Preparation:

Since statistical approach-based translation systems are founded on the use of corpora, large quantity and good quality data is essential and the key for the success of SMT systems. MOSES also need the creation of a large quantity of parallel as well as monolingual data to train the systems. MOSES needs both types of corporuses, parallel as well as monolingual at different stages of translation process.

Fig. 1. Working flowchart of MOSES



Parallel and Monolingual corpus: Parallel corpus is needed at training, tuning, and testing stage while monolingual corpus is required at language modelling stage. Since the systems were trained for English to Hindi translation, datasets also contain parallel corpora of English and Hindi and monolingual corpora of Hindi and have been extracted in individual files for training, tuning, testing and language modelling.

The size of parallel corpus for training was 1 lac (99327 after cleaning), 2 lac (198864 after cleaning), 3 lac (297484 after cleaning), 4 lacs (397104 after cleaning), 5 lacs (494685 after cleaning) and 6 lacs (593426 after cleaning) respectively for STMDL#1 - STMDL#6.

Training corpus: Training corpus for STMDL# 1 was comprised of 50 k sentences from one of the repositories of git hub and 50 k sentences from SAMANANTAR [3] corpus. Training corpus for each of successive STMDL# was comprised of training data of previous STMDL and by adding 1 lac more parallel sentences from SAMANANTAR

corpus, i.e., training corpus of STMDL# 2 was constituted by adding 1 lac more parallel sentences with training corpus of STMDL# 1, similarly training corpus of STMDL# 3, STMDL# 4, STMDL#5 and STMDL#6.

Tuning and Testing corpus: To train the LM for Hindi, Hindi text of corresponding training corpus was merged with another data set of approximately 50K sentences, which was separately collected from different sources like newspaper, social media etc. For tuning stage, 6 separate parallel data sets of size 10k, 20k, 30k, 40k, 50k and 60 k sentences respectively for STMDL#1 to STMDL#6 were formed from SAMANANTAR corpus. All models were tested on the same testing data set of 6K parallel sentences. 3K sentences of this testing corpus were taken from SAMANANTAR corpus while remaining 3K sentences were collected by author their own from different sources like newspaper, social media, etc. From the test set we randomly select 100 sentences for a separate human evaluation. From the test set we randomly select 100 sentences for a separate human evaluation. Table 1 summarizes the data collection statistics and other details of data sets.

B. Language modelling, training, tuning, and testing set-up:

All corpus sentences were encoded in utf-8 format and wx2 notation was used for easy tokenization of Hindi sentences. In corpus preparation step, Hindi corpus was tokenized by using IndicNLP tokenizer and English corpus was tokenized by MOSES tokenizer. The language modelling was accomplished by building n-gram ($n = 3$) language model with modified Kneser-Ney smoothing [4] using Ken-LM [5]. After creating LM, system training was completed by using Word alignment tool GIZA++. Learning word alignments from the parallel training corpus was accomplished by using GIZA++ [6] with grow-diag-final-and heuristics for extracting phrases and lexicalized reordering. In word alignment model the GIZA++ will check the word alignment between source and target sentences.

For building phrase-based systems, we use msd-bidirectional-fe as reordering model, set distortion limit to 6. For other parameters of Moses, default values were used. Finally, the trained system was tuned with Minimum Error Rate Training (MERT) [7] to learn the weights of different parameters of the model and Batch MIRA [8] for tuning (default parameters). After training the system, sample sentences were tested by using moses.ini. From reality check and fair analysis point of view, separate data sets from training corpus were used for testing and tuning the system.

Table 2 summarizes the various settings and tools used at different stages in the entire process along with major tasks performed in every step.

III. RESULTS

All six phrase-based SMT models (STMDL#1 through STMDL#6) built using the MOSES toolkit were evaluated on a consistent test set of 6,000 English-Hindi sentence pairs using five widely accepted automatic evaluation metrics. Automatic evaluation metrics used for translation quality measurement in this experiment were multi-BLEU [9], NLTK-BLEU [10], TER [11], METEOR [12] and Ada-BLEU [13] metric. All evaluation scores are graphically represented in Fig. 2 and Fig. 3. A detailed analysis reveals a generally positive correlation between the size of the training corpus and the translation quality, as reflected by increasing scores in most metrics. Notably, Multi-BLEU and Ada-BLEU show relatively steady logarithmic growth, suggesting consistent improvement in n-gram overlap and word choice quality as the corpus size scales up. TER, which measures edit distance, also decreases in efficiency at higher scales but still follows a growth curve, indicating reduced post-editing effort with larger corpora. The NLTK-BLEU metric exhibits a slight drop at STMDL#2, before recovering strongly by

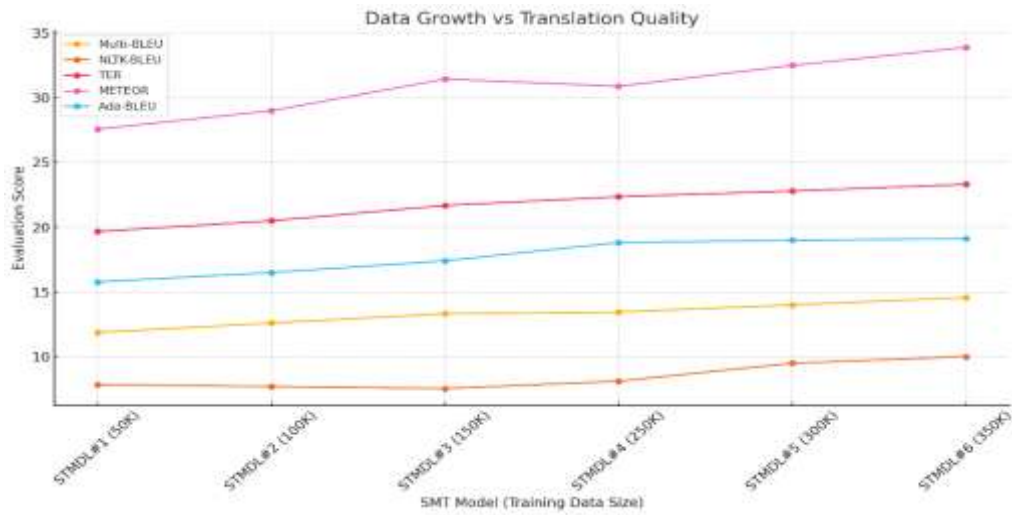
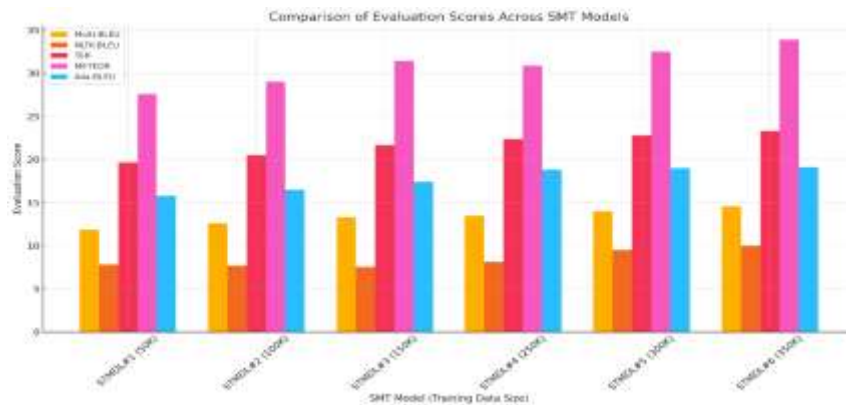
STMDL#4. This non-monotonic behavior could reflect sensitivity to specific phrase patterns or tokenization inconsistencies in intermediate data sizes. A similar dip-and-recovery pattern is observed in the METEOR score at STMDL#3. Such anomalies warrant further scrutiny and suggest that evaluation metrics may respond differently to certain corpus characteristics like domain overlap, vocabulary redundancy, or syntactic variability.

Table 1. Corpus Statistics

Purpose	Data Set	STMDL #	Size	Source of Collection	Language
Training	Sentence level	STMDL # 1	0.5 Lac	Downloaded from website [SAMA]	English-Hindi
		STMDL # 2	1.5 Lac		
		STMDL #3	2.5 Lac		
		STMDL # 4	3.5 Lac		
		STMDL # 5	4.5 Lac		
		STMDL # 6	5.5 Lac		
	Discourse level	STMDL # 1	50 K	Newspapers, Literature Social-Media	English-Hindi
		STMDL # 2	50 K		
		STMDL #3	50 K		
		STMDL # 4	50 K		
		STMDL # 5	50 K		
		STMDL # 6	50 K		
Language Modelling	Sentence level	STMDL # 1	0.5 Lac	Newspapers, Literature Social-Media	Hindi
		STMDL # 2	1 Lac		
		STMDL #3	1.5 Lac		
		STMDL # 4	2.0 Lac		
		STMDL # 5	2.5 Lac		
		STMDL # 6	3.0 Lac		
	Discourse level	STMDL # 1	50 K	Downloaded from website [SAMA]	Hindi
		STMDL # 2			
		STMDL #3			
		STMDL # 4			
		STMDL # 5			
		STMDL # 6			
Tuning	Sentence level	STMDL # 1	10K	Downloaded from website [SAMA]	English-Hindi
		STMDL # 2	20 K		
		STMDL #3	30 K		
		STMDL # 4	40 K		
		STMDL # 5	50 K		
		STMDL # 6	60 K		
Testing	Sentence level	STMDL # 1	3 k	Newspapers, Literature Social-Media	English-Hindi
		STMDL # 2			
		STMDL #3			
		STMDL # 4			
		STMDL # 5			
		STMDL # 6			
	Discourse level	STMDL # 1	3 K	Downloaded from website [SAMA]	English-Hindi
		STMDL # 2			
		STMDL #3			
		STMDL # 4			
		STMDL # 5			
		STMDL # 6			

Table 2. Various settings and tools used at different stages

Step	Task	Tools
1. [Corpus Preparation]	Tokenization	IndicNLP for Hindi MOSES tokenizer for English
2. [Language Modelling]	LM	KenLM
3. [Training]	Word alignment Smoothing N-Gram	GIZA++ KneserNey 3-gram
4. [Tuning]	Weight Reducing	MERT
5. [Binarization]	Reordering	Msd-bidirectional-fe
6. [Testing]	Translation decoding Searching	Moses.ini BEAM
7. [Evaluation]	Translation score	BLEU, TER, METEOR, and Ada-BLEU

Fig. 2. Data Growth vs Translation Quality**FIG 3: COMPARISON OF EVALUATION SCORES ACROSS SMT MODLES**

Overall, these results suggest that while data size positively influences translation quality, the rate of gain is not uniform across all metrics, and qualitative improvements plateau after a certain training volume.

IV. CONCLUSION

This study sets out to evaluate the influence of training corpus size on the performance of phrase-based statistical machine translation (PB-SMT) systems using the MOSES framework and English-Hindi language pairs. Through a sequence of systematically constructed translation models — ranging from 50,000 to 350,000 training pairs — and a consistent experimental setup across language modeling, tuning, and evaluation, the paper has demonstrated the tangible impact of scaling training data on translation quality.

Key findings highlight that translation performance improves with training corpus size, but the rate of improvement diminishes over time. While metrics like Multi-BLEU and Ada-BLEU reflect a stable positive trend, other metrics such as NLTK-BLEU and METEOR exhibit nonlinear behavior, suggesting that metric selection is critical for robust evaluation of SMT models.

REFERENCES

- [1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.
- [2] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open-source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.
- [3] Ramesh, G., Doddapaneni, S., Bheemaraj, A., et al. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. (2022).
- [4] Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.
- [5] Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.
- [6] Franz Josef Och and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, 29(1):19-52. Computational Linguistics.
- [7] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.
- [8] David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of LREC*, pages 697–702.
- [9] 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*, pages 311–318. Association for Computational Linguistics.

- [10]G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceeding of the 2nd International Conference on Human Language Technology Resources, San Diego, CA, 2002, pp. 138–145.
- [11]B. Satanjeev, and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, 2005, pp. 65–72.
- [12]G. Ankush, S. Venkatapathy, and R. Sangal, "METEOR Hindi: Automatic MT evaluation metric for Hindi as a target," in Proceedings of ICON-2010: 8th International Conference on Natural Language Processing. India: Macmillan Publishers, 2010.
- [13]Matthew Snover and Bonnie Dorr, "A Study of Translation Edit Rate with Targeted Human Annotation".