Reconsidering SMT Over NMT for *Closely Related* Languages: A Case Study of Persian-Hindi Pair

Waisullah Yousofi and **Pushpak Bhattacharyya**
Computation for Indian Langauge Technology (CFILT)
Indian Institute of Technology Bombay, Mumbai, India.
(waisullah, pb)@cse.iitb.ac.in

Abstract

This paper demonstrates that Phrase-Based Statistical Machine Translation (PBSMT) can outperform Transformer-based Neural Machine Translation (NMT) in moderate-resource scenarios, specifically for structurally similar languages, Persian-Hindi pair in our case. Despite the Transformer architecture's typical preference for large parallel corpora, our results show that PBSMT achieves a BLEU score of **66.32**, significantly exceeding the Transformer-NMT score of 53.7 ingesting the same dataset.

Additionally, we explore variations of the SMT architecture, including training on Romanized text and modifying the word order of Persian sentences to match the left-to-right (LTR) structure of Hindi. Our findings highlight the importance of choosing the right architecture based on language pair characteristics, advocating for SMT as a high-performing alternative in such cases, even in contexts commonly dominated by NMT.

1 Introduction

In the current state of NLP affairs, the performance of attention-based (Bahdanau et al., 2014; Vaswani et al., 2017) MT systems reaches BLEU scores of almost one. However, the underlying Neural Network (NN) architectures of such high-performing models, assume that the language pairs have a humongous diverse parallel corpora to achieve such desired performance. Of course, there are certain high-source language pairs such as English and French which benefit from those solutions but the MT system of other natural languages that utilize NN architecture without meeting the architecture's assumptions are on the disadvantage side and will generate translations that are way far being accepted by native speakers.

Beyond the need for large datasets, another great concern when using NNs is their high power consumption and the environmental impact they leave behind—through processes such as training, inference, and experimentation, which all contribute to carbon footprints (Faiz et al., 2023).

To walk through an efficient alternative path, we looked at the big picture of natural languages, focusing on their linguistic families and the factors that group languages. We have observed that the property of linguistic closeness of less divergent languages can be exploited. The key contributions of our paper are:

- The first attempt to build a general domain MT system of Persian-Hindi languages.
- We demonstrate that for structurally close language pairs having a moderate-sized (1M+ sentences) high-quality parallel corpus, SMT outperforms a Transformer-based NMT model.
- Suggesting alternative paths to build computationally and environmentally efficient MT systems.

The upcoming sections are structured as follows: section 2 presents a review of the literature, followed by a detailed description of the parallel corpus used in our experiments and analysis in section 3. Section 4 outlines the experimental setup, while section 5 provides a comprehensive analysis of the results. Finally, section 6 concludes the paper and discusses potential directions for future research.

2 Related Works

Before the revolutionization of the field by the Transformer architecture (Vaswani et al., 2017), the notion of language closeness was being leveraged in various forms for different language pairs. In this section, we will see that our work is not only different from the perspective of studying a new language pair, Persian-Hindi, which to date no formal research has been conducted yet for the pair, but it also varies in terms of past Transformer comparison of the two architectures, NMT and SMT, given that we have access to a moderate amount of

parallel sentences.

Split	Sentences	FA-Tokens	HI-Tokens
Train	1,01M	17.25M	17.75M
Test	3,000	50k	52K
Tune	8,000	1.36M	1.39M
Total	**1M+**	**19.1M**	**19.1M**

Table 1: Corpus statistics after applying LABSE filtration with a threshold of 0.9.

Language	Sentences	Tokens
Persian	13.7M+	190M+
Hindi	13.7M+	207M+

Table 2: Details of normalized but unfiltered monolingual of Persian and Hindi.

Previous works such as,(Toral and Way, 2015), hypothesize that translations between related languages tend to be more literal, with complex phenomena (e.g., metaphors) often transferring directly to the target language. In contrast, these phenomena are more likely to require complex translations between unrelated languages. Other instances such as (Rios and Sharoff, 2015), (Kunchukuttan et al., 2017), and (Kunchukuttan and Bhattacharyya, 2017) utilize lexical similarities. Except (Jauregi Unanue et al., 2018) which shows that for the scenario of low-resource (unlike our scenario which assumes medium resource) languages, SMT performs better than NMT, there is no comparative analysis of the NMT and SMT architectures for structurally similar languages with an assumption concerning the size of the parallel corpus.

3 Dataset and Preprocessing

In addition to the "Large Scale Colloquial Persian Dataset" (LSCP) (Abdi Khojasteh et al., 2020), the other datasets we utilized are primarily sourced from OPUS (Tiedemann, 2016), a well-known repository for parallel corpora of various domains for a vast number of language pairs.

Table 3 shows the basic statistics related to all the corpora. After downloading those **10.9M+** sentences, we noticed that most of them were of low quality. To filter them we used LABSE (Feng et al., 2020) during which (Batheja and Bhattacharyya, 2022)'s work was of help. Before the LABSE-filtration, the preprocessing steps for each corpus include, the removal of empty lines, punctuations, emojis, and deduplication of repeated parallel sentence pairs, normalization, and tokenization using language-specific libraries, `indic-nlp-library` (Kunchukuttan, 2020) for Hindi and `ParsiNorm` (Oji et al., 2021) for Persian.

An additional time-taking step applied to the LCSP corpus was to pair the sentences first and then pass to the preprocessing phase. Then, we performed LABSE filtration which the dramatic reduction of the original corpus is shown in Table 1. In one of our SMT experiments, as we will see the details in the next section, the parallel sentences need to be Romanized, for which we employed `uroman` library (Hermjakob et al., 2018). For all NMT experiments, the first step after receiving the raw data (Table 1) was to apply Byte Pair Encoding (BPE) (Sennrich et al., 2016) with 32K merge operations.

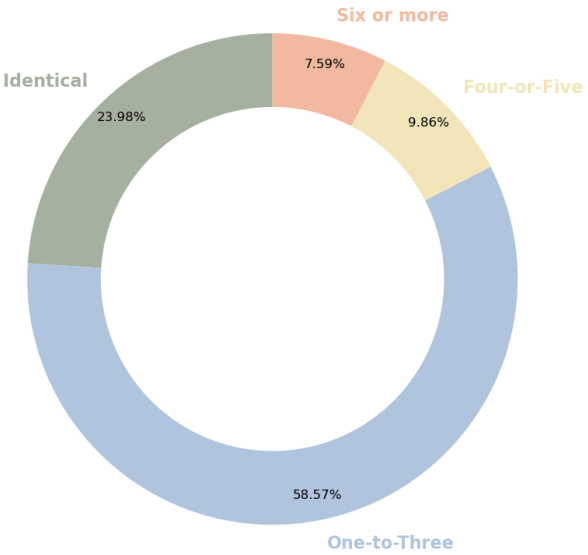

Figure 1: Categories the differences of lengths of parallel sentences length counted in terms of tokens.

It should be mentioned that for the Language Model (LM) component of the SMT model, we used the unfiltered monolingual of the target language, Hindi, and the corresponding numbers are detailed in Table 2.

To evaluate the structural similarity between Persian and Hindi, we analyzed the differences in sentence lengths (measured in token counts) across all parallel sentences. We assumed that if the majority of sentences exhibit a difference of three tokens or less, the alignment achieved through `mgiza` would represent an optimal one-to-one correspondence. As illustrated in the donut chart in Figure 1, more than 72% of the parallel sentences in our dataset,

Corpus	Sentences	FA Tokens	HI Tokens
CCMatrix v1	2,7M+	30M+	32M+
NLLB v1	2M+	30M+	32M+
MultiCCAligned v1.1	1M+	22M+	21M+
XLEnt v1.2	0.4M+	1M+	1M+
Tanzil v1	.01M+	4M+	4M+
KDE4 v2	72K	0.3M+	0.3M+
OpenSubtitles v2018	48K	0.2M+	0.3M+
TED2020 v1	41K	0.7M+	0.7M+
GNOME v1	40K	0.1M+	0.1M+
WikiMatrix v1	20K+	0.3M+	0.3M+
NeuLab-TedTalks v1	16K	0.3M+	0.3M+
QED v2.0a	2k	0.7M	0.6M
ELRC-wikipedia_health v1	1k	1K	1K
GlobalVoices v2018q4	139	1K	1K
TLDR-pages v2023-08-29	58	447	454
Wikimedia v20230407	40	2k	2K
Ubuntu v14.10	6k+	27K	29K
LSCP Corpus	4.6M+	1.3M+	1.1M+
Total	**10.9M+**	**90.9M+**	**93.7M+**

Table 3: Basic Statistics of Individual and Merged Corpora.

Table 1, have a length difference of less than three tokens. Moreover, through the ***language divergence*** *(the phenomenon of languages expressing meaning in divergent ways)* setting proposed by (Dorr, 1993), we studied the fact that the Persian-Hindi pair is almost isomorphic.

From the perspective of structure and syntax of German, Spanish, and English, Dorr proposes a set of seven types of divergences (Bhattacharyya, 2015). For our pair, we examine some types of syntactic divergence through examples provided in the Appendix A:

4 Experimental Setup

4.1 Moses SMT

SMT which gave rise to NMT has been around for quite a long time. The basic idea of SMT is to learn the word alignment first and then expand it to phrases to build a phrase table that will be used for predictions. All three SMT-based experiments that we performed generally follow the same pipeline which is illustrated in Figure 2. We utilized the open-source toolkit, Moses (Koehn et al., 2007) to train a PBSMT model. First, a word alignment model between the Persian and Hindi languages was trained on the training data using MGIZA++ toolkit (Och and Ney, 2000). Next, a 5-gram Language Model (LM) employing Kneser-Ney smoothing and interpolation was built using *SRILM* toolkit developed by (Kneser and Ney, 1995). Since these two languages do not require transforming their scripts into lower-case, true-case, etc, we neither applied those transformations nor used Moses's default tokenizer- we used language-specific libraries for better results. Finally, Moses decoder was used to translate sentences based on these components.

4.2 OpenNMT

We fed the same data splits, Table 1, that were used for SMT, to an NMT Transformer (Vaswani et al., 2017) model with the help of open-source OpenNMT (Klein et al., 2017) library. The NMT model consists of 8 layers of encoder and decoder each with 8 attention heads, the embedding layer of size 512 with positional encoding enabled. Coming to hyperparameters, the batch size was set to 4096 tokens iterating over 300K steps with an initial learning rate of 2 along with Adam optimizer setting β_1 and β_2 to 0.9 and 0.998 respectively. Additionally, we utilized 8000 warmup steps.

In terms of GPU usage, two NVIDIA GeForce RTX 2080 GPUs were occupied during training.

4.3 Evaluation

Throughout the experiments, we used the BLEU metric (Papineni et al., 2002). Although, the way Moses [1] calculates this score is correct but a refined version that better handles BLEU's hyperparameters and computes it is SACREBLEU (Post, 2018).

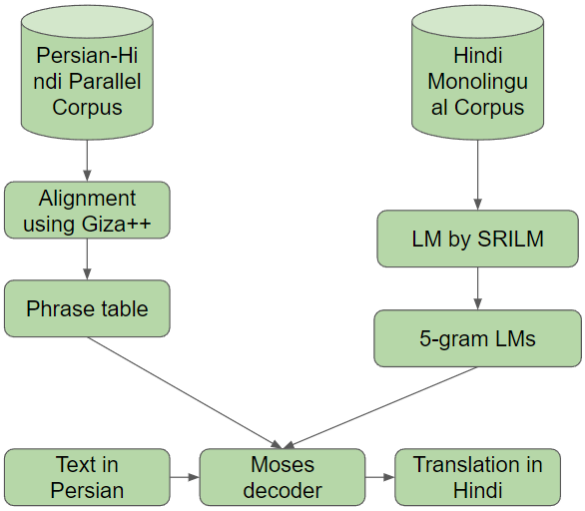

Figure 2: Architecture of Persian to Hindi SMT Model.

5 Experiments and Results

The first SMT experiment used the normalized filtered data of Table 1, applying it conventionally to SMT-Moses with the configurations detailed in Section 4. While Moses was processing the data, we simultaneously set up our encoder-decoder Transformer model. The best NMT model achieved *53.7*, whereas the initial SMT model had a BLEU score **64.9**. To verify the high BLEU score of SMT, we conducted a 4-fold cross-validation the BLEU scores of which are ***67.32***, *66.32, 64.90,* and *66.74*, respectively. Therefore, our best SMT model's BLEU has been marked **66.32**- the average of 4-fold's BLEU. Also, by looking at the algorithmic nature of each architecture, it makes sense for the SMT to perform better than NMT in the existence of moderate data size. Because, SMT uses Expectation Maximization (EM) algorithm for alignment, and since the source and target languages are almost always one-to-one mapping, we need less data-size than that of NN.

Since Persian and Hindi share many common words, in our second experiment, parallel sentences were first Romanized to increase text similarity (Hermjakob et al., 2018). However, the BLEU score dropped to *51.21* from *66.7*. One highly probable reason is due to the diacritics (mark that is placed above, below, or through a letter to indicate how it should be pronounced) that some Persian words have in order to determine the phoneme of a word and hence the associated meaning. For example, `gul (flower)` and `gel (mud)` are two words the sound and meaning of which can be determined from the context (without diacritics) or using diacritics. Romanization often results in the loss of these nuances, leading to ambiguities in alignment and translation.

Model	BLEU Score
Initial SMT Model	64.91
Best SMT Model	**66.32**
FOLD 1	67.32
FOLD 2	66.32
FOLD 3	64.90
FOLD 4	66.74
NMT Transformer Model	53.7
Romanized-SMT Model	51.21
Inverted-SMT Model	48.74

Table 4: BLEU scores of various SMT and NMT models.

In the final experiment, we reversed the Persian scripts from right-to-left (RTL) to left-to-right (LTR) to align the writing direction of Hindi, expecting improved alignment quality. Unfortunately, this resulted in a further decrease in the BLEU score, falling to *48.74*. This decline can be attributed to the fact that reversing a sentence alters its meaning. Although both Persian and Hindi are classified as free-word-order languages, we observed that the phenomenon where "only constructs that follow each other can be moved to any other position in the sentence while still preserving meaning" is compromised when inversion occurs, leading to a change in interpretation. See Appendix A for examples. Table 4 summarizes the BLEU scores for the different experiments we performed.

6 Conclusion and Future Works

This study presents a comparative analysis of SMT and NMT for the Persian-to-Hindi language pair. Our findings demonstrate that SMT yields superior results in closely related languages, attributable to their shared linguistic structures. Additionally, we

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

observed that reversing the order of Persian sentences from RTL to LTR negatively impacted the SMT model's performance, resulting in a loss of meaning. In contrast, romanizing the input text showed a beneficial effect compared to the inversion experiment.

Future work will focus on deepening our understanding of these languages and exploring alternative approaches, such as translation through common space word embedding, transfer learning, and pivot-based NMT with English as a bridging language.

References

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. LSCP: Enhanced large scale colloquial Persian language understanding. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6323–6327, Marseille, France. European Language Resources Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2015. *Machine translation*. CRC Press.

Bonnie Jean Dorr. 1993. *Machine translation: a view from the lexicon*. MIT press.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Parteek Sharma, Fan Chen, and Lei Jiang. 2023. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. *arXiv preprint arXiv:2309.14393*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Inigo Jauregi Unanue, Lierni Garmendia Arratibel, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. English-Basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE.

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 association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The indic nlp library. https://github.com/anoopkunchukuttan/indic_nlp_library.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2017. Learning variable length units for SMT between related languages via byte pair encoding. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 14–24, Copenhagen, Denmark. Association for Computational Linguistics.

Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. Utilizing lexical similarity between related, low-resource languages for pivot-based SMT. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 283–289, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Romina Oji, Seyedeh Fatemeh Razavi, Sajjad Abdi Dehsorkh, Alireza Hariri, Hadi Asheri, and Reshad Hosseini. 2021. Parsinorm: A persian toolkit for speech processing normalization. In *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE.

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.

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. Association for Computational Linguistics.

Miguel Rios and Serge Sharoff. 2015. Obtaining SMT dictionaries for related languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 68–73, Beijing, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Opus-parallel corpora for everyone. *Baltic Journal of Modern Computing*, 4(2).

Antonio Toral and Andy Way. 2015. Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver, Colorado, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A

References

Pushpak Bhattacharyya. 2015. *Machine translation*. CRC Press.

A Examples Appendix

A.1 Paralle Sentences of Different Length Categories

A.1.1 Identical Lengths

1.1.1.Fa: `tamam mahsulat hamel shodeh 100% bazorsi mi shvand.`
1.1.1.Hi: `bheje gae sabhee utpaad 100 nireekshan kie jaate hain.`
1.1.1.En: `All shipped products will be 100% inspected.`

1.1.2.Fa: `baraye etlaat bishtar lotfa ba ma tamas begirid`
1.1.2.Hi: `adhik jaanakaaree ke lie krpaya hamase sampark karen`
1.1.2.En: `For more information please contact us`

A.1.2 One-to-Three Token Difference

1.2.1.Fa: `akharin ghimet sakeh ve tala dar bazar. [7-tokens]`
1.2.1.Hi: `baajaar par naveenatam sikka aur sone kee keematen. [8-tokens]`
1.2.1.En: `The latest price of coins and gold in the market.`

1.2.2.Fa: `Har zemestān bahāri dar pay dārad. [6-tokens]`
1.2.2.Hi: `Har sardī ke baad vasant ṛtu hotī hai. [8-tokens]`
1.2.2.En: `After every winter there's spring.`

A.1.3 Four-or-Five Token Difference

1.3.1.Fa: `pish bini ab ve npava dar litvania.[7-tokens]`
1.3.1.Hi: `Havāmāna andāja lithu'āniyā. [3-tokens]`
1.3.1.En: `Weather forecast in Lithuania.`

1.3.2.Fa: `besiar sadeh baraye estefadeh.[4-tokens]`
1.3.2.Hi: `ka upayog karane ke lie bahut hee saral. [8-tokens]`
1.3.2.En: `Very simple to use.`

A.2 Inversion Example

2.1.original-Fa (read from right-to-left):
`.daram arezo azizan baraye zibayi`
2.1.En (of original-Fa):
`I wish beauty for the loved ones.`
2.1.inverted-Fa inverted (read left-to-right):
`daram arezo azizan baraye zibayi.`
2.1.En (of inverted-Fa):
`I wish the loved ones for beauty.`

As we can see, the inverted sentence wishes *the loved ones* FOR *the beauty*, which to some extent does not make sense at all. Hence, the conclusion we made here is that the inversion of the sentence disrupts the intended meaning and perhaps alignment, which consequently affects the overall performance negatively.

A.3 Syntactic Divergence

Through the examples taken from (Bhattacharyya, 2015), we are going to observe some cases where Persian and Hindi sentences do not diverge, which implies syntactic closeness.

A.3.1 Constituent Order Divergence

It is related to the divergence of word order between a pair. For instance, below we can see that both follow the same SOV order.

3.1.1.En: `Jim (S) is playing (V) tennis (O)`

3.1.1.Fa: `jim (S) tenis (O) bazi karde rahi ast(V). [Jim tenis play work being is]`

3.1.1.Hi: `jeem (S) tenis (O) khel rahaa hai (V) [Jim tennis playing is]`

A.3.2 Null Subject Divergence

Null Subject Divergence refers to the phenomenon languages, such as Persian and Hindi, omit the subject pronoun (like "there" in English), because the subject is implied or understood from the verb form or context.

3.2.1.En: `Long ago, there was a king`

3.2.1.Fa: `Khili vaqt pish, yek padshah bud.[Long ago one king was]`

3.2.1.Hi: `bahut pahale ek raajaa thaa [Long ago one king was]`

Similar practices can be performed to observe that both languages diverge only rarely- conflational divergence.