Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054673

ARTICLE

LKMT: Linguistics Knowledge-Driven Multi-Task Neural Machine Translation for Urdu and English

Muhammad Naeem Ul Hassan[1,2], Zhengtao Yu[1,2,*], Jian Wang[1,2], Ying Li[1,2], Shengxiang Gao[1,2], Shuwan Yang[1,2] and Cunli Mao[1,2]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China

*Corresponding Author: Zhengtao Yu. Email: ztyu@hotmail.com

Received: 04 June 2024 Accepted: 26 August 2024 Published: 15 October 2024

ABSTRACT

Thanks to the strong representation capability of pre-trained language models, supervised machine translation models have achieved outstanding performance. However, the performances of these models drop sharply when the scale of the parallel training corpus is limited. Considering the pre-trained language model has a strong ability for monolingual representation, it is the key challenge for machine translation to construct the in-depth relationship between the source and target language by injecting the lexical and syntactic information into pre-trained language models. To alleviate the dependence on the parallel corpus, we propose a Linguistics Knowledge-Driven Multi-Task (LKMT) approach to inject part-of-speech and syntactic knowledge into pre-trained models, thus enhancing the machine translation performance. On the one hand, we integrate part-of-speech and dependency labels into the embedding layer and exploit large-scale monolingual corpus to update all parameters of pre-trained language models, thus ensuring the updated language model contains potential lexical and syntactic information. On the other hand, we leverage an extra self-attention layer to explicitly inject linguistic knowledge into the pre-trained language model-enhanced machine translation model. Experiments on the benchmark dataset show that our proposed LKMT approach improves the Urdu-English translation accuracy by 1.97 points and the English-Urdu translation accuracy by 2.42 points, highlighting the effectiveness of our LKMT framework. Detailed ablation experiments confirm the positive impact of part-of-speech and dependency parsing on machine translation.

KEYWORDS

Urdu NMT (neural machine translation); Urdu natural language processing; Urdu Linguistic features; low resources language; linguistic features pretrain model

1 Introduction

Neural Machine Translation (NMT) has become a predominant approach in developing machine translation systems. NMT, as introduced by [1], represents the cutting-edge methodology for machine translation. It has garnered prominence in both academic circles, as evidenced by the work of [2–4], and the industrial domain, with notable contributions from [5,6]. In the realm of advancements, recent studies by [7–10] have expanded the capabilities of NMT to encompass multilingual translation.

This entails the development of a unified model capable of seamlessly translating between multiple language pairs. Such endeavours mark significant strides in the evolution of NMT, positioning it as a versatile and powerful tool for overcoming language barriers across diverse linguistic landscapes. Language generation, distinct from understanding, focuses on creating natural language sentences based on given inputs. This involves various tasks such as NMT as demonstrated by [1,11,12] text summarization, as explored by [13,14] and conversational response generation, as researched by [15], these tasks often require large amounts of data, with many operating under low-resource or even zero-source conditions regarding training data.

Pre-trained Language Models (PLMs) have emerged as effective tools for text representation by incorporating rich contextual information. Among these, auto-encoding PLMs like BERT [16] and RoBERTa [17] are widely used for Natural Language Understanding (NLU) tasks. These models differ from auto-regressive PLMs, such as GPT [18], which rely on standard language models for training. Instead, auto-encoding PLMs depend on specific pre-training tasks to grasp contextual data. A key task in this domain is the Masked Language Model (MLM), introduced by BERT and widely adopted in others like RoBERTa, ALBERT [19], ERNIE [20], and DeBERTa [21]. MLM focuses on restoring words from masked text, chosen randomly, indicating its linguistic-agnostic nature. While PLMs are recognized for encompassing extensive linguistic knowledge, some researchers suggest enhancing PLMs with external knowledge. Efforts to integrate linguistic knowledge into PLMs include adding structural knowledge and other linguistic tasks. Despite these initiatives, previous works have limitations, mainly focusing on integrating various linguistic features without thoroughly analyzing their individual contributions or the interplay between different tasks. Moreover, these implementations can be complex, as structural knowledge is not readily integrated into PLMs.

Urdu exhibits flexibility in word order, with a Subject-Object-Verb (SOV) structure being common. Table 1 presents the languages used in our experiments and their linguistic features. Traditional machine translation models struggle with capturing such variations, leading to inaccuracies in translated output.

Table 1: Languages and their characteristics. Additionally, English is fusional languages, while Urdu analytic

language	Character	Word order
English	Latin alphabet	SVO
Urdu	Aryan	SOV

Unlike widely spoken languages such as English, there may be a lack of well-established pre-trained language models for Urdu.

Despite progress in machine translation (MT) and pre-trained language models (PLMs), challenges persist, particularly with low-resource languages like Urdu, which lacks extensive, varied datasets. This limitation hinders the development of effective NLP models, affecting performance and generalization. Urdu's intricate morphology and flexible syntactic structures further complicate natural language processing, especially in morphological analysis and understanding [22]. Current models often fail to grasp these complexities, and the integration of specific linguistic knowledge into PLMs is not well-explored.

Our research addresses these gaps by proposing a novel pre-training strategy for Urdu and English, integrating Part-of-Speech (POS) and Dependency (DEP) features into multi-task NMT

models. This approach enhances contextual understanding and translation accuracy by managing Urdu's morphological richness and syntactic variations, advancing the field of NMT for low-resource languages. We explore the effectiveness of this linguistic knowledge integration, demonstrating how our pre-trained model optimizes performance for both language understanding and generation tasks by adapting BERT-like methods to the unique demands of sequence-to-sequence learning frameworks. This targeted strategy addresses the critical issue of corpus insufficiency and paves the way for more robust translation models. The contribution of the paper is as follows:

1. Creation of a unique pre-trained language representation model specifically optimized for Urdu and English, addressing the gap of high-quality pre-trained models for low-resource languages.
2. We proposed method to incorporate linguistic knowledge, specifically POS and DEP features into pretrain model, using a linguistic knowledge pre-training strategy of Urdu and English as a multitask.
3. We develop techniques to better handle the morphological richness and flexibility of Urdu by leveraging morphological features in the translation process.
4. Implementing a multi-task learning framework that allows the model to simultaneously learn translation and linguistic annotations, improving overall translation quality and robustness.

2 Related Work

Pre-trained language models are designed to leverage extensive corpora during pre-training, aiming to capture a broad understanding of language that incorporates contextual nuances. Early approaches to word embeddings relied on static methods, pre-training embeddings using large corpora to capture semantic and syntactic similarities [23,24].

2.1 Pre-Trained Language Models for Machine Translation

Recent advancements in pre-trained language models such as BERT and GPT have significantly enhanced NMT by leveraging extensive pre-training on large corpora, capturing a broad understanding of language. Integrating these pre-trained models, such as by replacing the Transformer encoder, has led to improvements in WMT14 En↔De and En↔Fr tasks [25,26]. Combining BERT and GPT-2 with Transformer architecture further enhanced translation quality on the WMT-14 dataset [27]. CeMAT, a conditional masked language model, achieved significant performance gains [28], while leveraging pre-trained checkpoints for sequence generation tasks resulted in state-of-the-art results [25]. To address data scarcity, pre-trained models improved performance on the IWSLT'14 En↔De dataset [29], and mBART enhanced low-resource and unsupervised MT tasks [30]. Additionally, reinforcement learning-based curriculum learning improved model performance [31], and synthetic pre-training mitigated issues such as toxicity and bias [32]. Enhancing NMT for low-resource languages by integrating syntactic and semantic knowledge showed substantial improvements [33]. Notably, developing an English-Urdu NMT system achieved a high score [34].

2.2 Knowledge-Enhanced Machine Translation

Incorporating various types of linguistic knowledge into NMT models has significantly enhanced their performance. Lexical integration, for instance, utilizes fine-tuned vector-based linguistic information from BERT to improve generalization in NMT, resulting in notable improvements [35]. A knowledge-aware NMT approach models additional linguistic features using RNNs, achieving significant BLEU (Bilingual Evaluation Understudy) score improvements in Chinese↔English and

English→German tasks [36]. Morpheme segmentation and linguistic features have been used to enhance translation predictions, particularly improving performance in low-resource conditions [37]. Incorporating these features in BPE-based NMT models for Indian languages has also shown significant performance boosts [38]. Multi-source neural models utilize separate encoders for the source word sequence and linguistic feature sequences, improving translation quality for Turkish-English and Uyghur-Chinese tasks [39]. NMT systems developed for English-Kannada using a pre-trained Cross-Lingual Language model have achieved significant performance improvements [40]. Moreover, low-resource multilingual NMT using linguistic feature-based relevance mechanisms has resulted in significant BLEU score improvements for multiple Asian languages [41].

2.3 Syntactic Integration in Machine Translation Models

Integrating syntactic knowledge into NMT models has consistently shown to enhance translation quality. For instance, a framework was proposed for leveraging pre-trained models through dynamic fusion mechanisms and knowledge distillation, significantly improving translation quality [33]. In another study, NMT was applied to English-Tamil and English-Malayalam, using pre-trained Byte-Pair-Encoded (BPE) embeddings and MultiBPE embeddings to address the out-of-vocabulary (OOV) issue in low-resource languages, which substantially outperformed Google Translate [42]. A multi-task learning approach was employed to incorporate auxiliary tasks such as semantic parsing, syntactic parsing, and named-entity recognition, effectively injecting semantic and syntactic knowledge into the translation model and resulting in enhanced performance for English-French, English-Farsi, and English-Vietnamese translations [43]. Additionally, a method was introduced for learning latent feature representations from input sentences using a latent feature encoder, significantly boosting translation performance in large-scale tasks [44]. A fine-tuning procedure combining embeddings freezing with adversarial loss for domain adaptation in pre-trained multilingual NMT models was introduced, leading to improved performance on specialized data with minimal loss in general domain quality [45]. Furthermore, a penalty mechanism was developed to regulate copying behaviors during pre-training, enhancing translation quality [46]. In another development, a language-generation model was pre-trained using a Masked Sequence-to-Sequence pre-training method for Korean and Japanese, achieving high performance in unsupervised NMT by leveraging shared syntactic structures [47]. Finally, a framework was introduced for integrating BERT into NMT, ensuring the retention of pre-trained knowledge and avoiding catastrophic forgetting, which led to significant BLEU score improvements [48]. Even with the advances in pre-trained language models (PLMs) and machine translation (MT), there are still issues, particularly with low-resource languages like Urdu that lack large datasets. The construction of efficient NLP models is hampered by this scarcity, which has an impact on the models' generalisation and performance. Morphological analysis and syntactic comprehension in natural language processing are made more difficult by Urdu's intricate morphology and flexible syntax. These complexity are frequently beyond the capabilities of current models, and there is still much to learn about how to incorporate language expertise into PLMs.

3 Model

Our proposed model as depicted in Fig. 1 introduces transformative innovations in linguistic analysis and translation tasks. It integrates extensive linguistic knowledge embeddings, including POS, DEP tags, punctuation sensitivity, verb aspect, and noun case embeddings, which enhance syntactic parsing and semantic understanding. The advanced encoder-decoder architecture features specialized multi-head attention mechanisms optimized for linguistic tasks, each head focusing on different linguistic features like syntactic dependencies and semantic roles for improved contextual

awareness. A key innovation is the self-adjusting attention mechanism in the translation module, which dynamically adjusts focus based on text complexity and linguistic features, ensuring idiomatic and accurate translations. Additionally, our novel positional encoding combines fixed and dynamic components, better handling long-range dependencies and varied sentence structures, beneficial for languages with flexible syntax like Urdu. The output stage finely tunes translations by applying processed linguistic features for linguistic accuracy and contextual nuance. The model's training involves a synergistic multi-task learning strategy, enhancing its capabilities across various linguistic tasks by leveraging the interconnectedness of linguistic features, thus boosting overall performance.

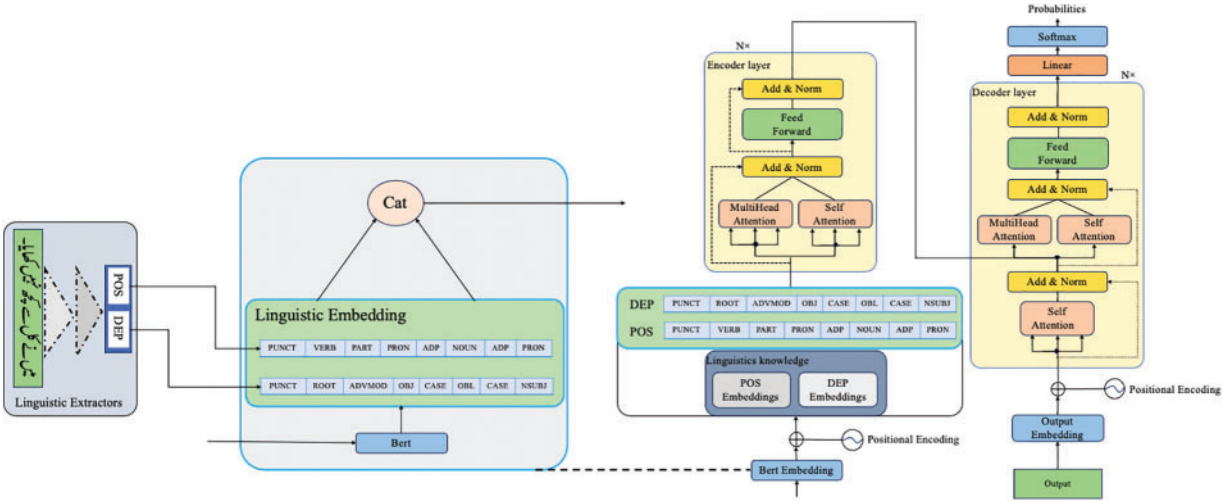


Figure 1: Overview of our model. We use one Urdu and English in the input for simplicity as Multilingual with linguistic feature (POS and DEP), NMT for only two language Urdu and English

3.1 Linguistic Features

Our goal is to employ linguistic features in a straightforward manner. To achieve this, the linguistic features we generate must possess two essential traits: high accuracy and uniqueness. High accuracy implies that these features should reliably represent the text's structure. While existing language analysis tools like Stanford Stanza [49] can analyze linguistic features, not all of them exhibit high accuracy. Uniqueness dictates that each input token should be associated with precisely one target tag for a specific linguistic feature, as illustrated in Fig. 2a for Urdu and Fig. 2b for English. In this study, we utilize Stanford Stanza tools, following the approach outlined by [49], to annotate the input text with two primary types of linguistic features: POS and DEP.

PUNCT	VERB	PART	PRON	ADP	NOUN	ADP	PRON
۔	کھایا	نہیں	کچھ	سے	کل	نے	میں

(a)

PRON	VERB	ADV	VERB	NOUN	ADP	NOUN	PUNCT
I	have	not	eaten	anything	since	yesterday	.

(b)

Figure 2: (a) POS features of Urdu (SOV) right to left based language; (b) POS features of English (SVO) right to left based language

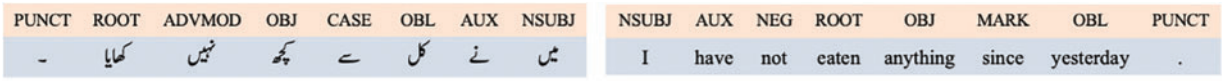
These features play a pivotal role in our analysis. To facilitate comprehension, we present a comprehensive list of linguistic tags used in our approach, we provide specific insights into the scope and categories of linguistic annotations employed in our study. For POS tagging, each input token

receives a unique POS label, resulting in a total of 20 distinct POS tag types for both Urdu and English languages showing in Table 2.

Table 2: POS tag on the input sequence use in our approach

No.	Pos tags	Descriptions
1.	ADJ	Adjective
2.	ADV	Adverb
3.	INTJ	Interjection
4.	NOUN	Noun
5.	PROPN	Proper noun
6.	VERB	Verb
7.	ADP	Ad position
8.	AUX	Auxiliary verb
9.	CONJ	Coordinating conjunction
10.	DET	Determiner
11.	NUM	Numeral
12.	PART	Particle
13.	PRON	Pronoun
14.	SCONJ	Subordinating conjunction
15.	PUNCT	Punctuation
16.	SYM	Symbol
17.	X	Other
18.	CCONJ	Coordinating conjunction
19.	PART	Particle
20.	.	Punctuation

Regarding DEP, syntactic dependency parsing is conducted on the input sequence. It's important to note that we assign the relation label to its dependent, guaranteeing that each token receives a unique label. This approach leads to a total of 38 different types of dependency relations and the different type of DEP in showing Fig. 3a,b.



(a) (b)

Figure 3: (a) DEP features of Urdu (SOV); (b) DEP features of English (SVO)

We present a comprehensive list of linguistic tags used in our approach, detailed in Table 3 for DEP.

Table 3: Dependency parsing on the input sequence use in our approach

No.	DEP	Descriptions	No.	DEP	Descriptions
1.	ACL	Clausal modifier of noun	2.	FIXED	Fixed multiword expression
3.	ADVCL	Adverbial clause modifier	4.	FLAT	Flat multiword expression
5.	ADVMOD	Adverbial modifier	6.	GOESWITH	Goes with
7.	AMOD	Adjectival modifier	8.	IOBJ	Indirect object
9.	APPOS	Appositional modifier	10.	LIST	List
11.	AUX	Auxiliary	12.	MARK	Marker
13.	CASE	Case marking	14.	NMOD	Nominal modifier
15.	CC	Coordinating conjunction	16.	NSUBJ	Nominal subject
17.	CCOMP	Clausal complement	18.	NUMMOD	Numeric modifier
19.	CLF	Classifier	20.	OBJ	Object
21.	COMPOUND	Compound	22.	OBL	Oblique nominal
23.	CONJ	Conjunct	24.	ORPHAN	Orphan
25.	COP	Copula	26.	PARATAXIS	Parataxis
27.	CSUBJ	Clausal subject	28.	PUNCT	Punctuation
29.	DEP	Unclassified dependent	30.	REPARANDUM	Overridden disfluency
31.	DET	Determiner	32.	ROOT	Root
33.	DISCOURSE	Discourse element	34.	VOCATIVE	Vocative
35.	DISLOCATED	Dislocated elements	36.	XCOMP	Open clausal complement
37.	EXPL	Expletive	38.	NEG	Negation modifier

3.2 Linguistics Knowledge-Driven

For every linguistic task, we approach it as a classification endeavor. Each input token undergoes projection to its respective linguistic features (POS, DEP), annotated via the methodology outlined in the preceding section. Upon scrutiny of these linguistic attributes, it becomes evident they possess varying degrees of significance. We hypothesize that POS stands as the foundational linguistic attribute, trailed by DEP. Considering their interdependencies, we allocate distinct learning rates to each linguistic feature, thereby facilitating quicker acquisition of POS compared to DEP. This mirrors the natural learning process observed in humans, where fundamental concepts are typically grasped before delving into more complex, dependent knowledge domains. To achieve this, we implement a fully-connected layer for mapping input tokens to their corresponding linguistic labels across each task.

In the model, each token from the input sequence undergoes a sophisticated transformation process. This transformation integrates the token's original embedding with the embeddings generated for POS and DEP tags. this is represented by the equation as:

$$e_i^{'} = ReLU(W_e.e_i + W_p.e_{pi} + W_d.e_{di} + b_e) \tag{1}$$

In this equation, e_i is the original embedding of the token, e_{pi} and e_{di} are the embeddings for POS and DEP tags, and W_e, W_p, W_d, and b_e are the weights and bias parameters learned during training. The

ReLU (Rectified Linear Unit) function is applied to introduce non-linearity, enhancing the model's ability to learn complex patterns.

For POS tagging, each token is projected onto a POS space using the weight matrix W_{pos} and bias b_{pos}. The projected feature is then classified using a softmax layer to determine the probability distribution over all possible POS tags. This is expressed as:

$$POS(t) = W_{pos}.e_i^{'} + b_{pos} \tag{2}$$

$$P_{pos}(t) = \mathrm{s}oftmax(POS(t)) \tag{3}$$

For DEP tagging, a similar approach is used, where each token is projected onto a DEP space using its respective weight matrix W_{dep} and bias b_{dep}. The classification through a softmax layer then assigns probabilities to each DEP label:

$$DEP(t) = W_{dep}.e_i^{'} + b_{dep} \tag{4}$$

$$P_{dep}(t) = \mathrm{s}oftmax(DEP(t)) \tag{5}$$

The learning speed is adjusted to prioritize the learning of POS tags over DEP tags. This is done by modifying the learning rate α with a factor γ for POS tagging, such that $\alpha_{pos} = \alpha \times \gamma$.

And $\alpha_{dep} = \alpha$. This adjustment ensures that the model learns the fundamental linguistic features at a faster rate before progressing to the more complex dependencies inherent in DEP tagging. This model will be used to understand the language beyond the surface level tokens, enabling it to delve into the grammatical structure and relationships between words, which is particularly beneficial for languages with rich morphological features like Urdu.

3.3 MLM Task

The pre-training of the model follows the methodology introduced in the original BERT paper by [16], incorporating the masked language model task. This task entails training the model to predict tokens that have been masked at random within a sequence. Typically, 15% of the tokens in a given sequence are randomly chosen for masking. In this process, each selected token s_i in the sequence has an 80% probability of being replaced by a [MASK] token, a 10% chance of being substituted with a random token, and a 10% probability of remaining unaltered, as illustrated in the masking procedure below:

$$S = [s_1, s_2, s_3, s_4, \ldots., s_n] \tag{6}$$

The MLM model then predicts a probability distribution over the entire vocabulary v for each masked token, calculated as:

$$P(v|e_i) = softmax(W.e_i + b) \tag{7}$$

Here, e_i is the embedding of the masked token, and $W$and b are trainable parameters of a linear layer applied to the embeddings. The model's effectiveness in languages like Urdu and English hinges on its ability to capture and predict contextually relevant tokens, considering the linguistic intricacies and structural differences between these languages.

میں نے کل سے [MASK] نہیں کھایا۔

3.4 Decoder Layers

In the initial embedding layer, the model utilizes a Transformer encoder to perform additional processing on these embeddings. Each encoder layer consists of two essential sub-layers: a multi-head self-attention mechanism and a position-wise feed-forward network. Multi-head self-attention mechanisms are employed to handle sequences of words, a design renowned for its capacity to effectively capture long-range dependencies in text. Within this architecture, every token in the input sequence undergoes transformation into query Q, key K, and value V vectors via learned linear transformations. The self-attention score for each token is computed as follows:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{8}$$

Here, d_k is the dimensionality of the key vector. This score determines how much focus to put on other parts of the input sequence when encoding a specific part. The attention mechanism computes a weighted sum of the values, where the weight assigned to each value is determined by a compatibility function of the query with the corresponding key.

In the Transformer decoder, the self-attention mechanism is modified to incorporate masking. This is critical in a generation task to ensure that the prediction for a particular token can only depend on previously generated tokens. The masked self-attention mechanism can be mathematically represented as follows:

$$MaskedAttention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{9}$$

Here, Q, K, and V represent the queries, keys, and values, respectively, similar to the standard attention mechanism. M is a mask matrix applied to prevent future tokens from influencing the generation of the current token. The mask effectively eliminates information about future tokens, aligning with the autoregressive nature of language generation.

Besides the masked self-attention, the decoder incorporates an additional layer of attention that focuses on the output of the encoder. This encoder-decoder attention mechanism is crucial for integrating information from the source text into the target translation process.

The encoder-decoder attention operates as follows:

$$EncDecAttention\,(Q_{dec}, K_{enc}, V_{enc}) = softmax\left(\frac{Q_{dec}K_{enc}^T}{\sqrt{d_k}}\right)V_{enc} \tag{10}$$

In this equation, Q_{dec} represents the queries from the decoder, while K_{enc} and V_{enc} are the keys and values from the encoder output. This attention layer effectively allows the decoder to 'attend' to different parts of the input sentence as it generates each word of the translation.

Each decoder layer is the feed-forward network (FFN). This network consists of two linear transformations with a ReLU activation in the middle, further processing the information from the attention layers. The FFN can be expressed as:

$$FFN\,(x) = \max\,(0, xW_1 + b_1)\,W_2 + b_2 \tag{11}$$

The FFN enhances the decoder's ability to capture complex language features, facilitating the generation of grammatically and contextually coherent translations. The input to the FFN, denoted as x is first passed through a linear layer. This linear layer is characterized by a weight matrix W_1 and

a bias term b_1. The linear transformation can be mathematically represented as $xW_1 + b_1$. While W_2 and b_2 project the input data into a higher-dimensional space, allowing the model to represent more complex patterns and relationships. After the ReLU activation, the resulting features will be further refined through a second linear transformation with weight matrix W_2 and bias term b_2, enabling the model to capture intricate linguistic features necessary for producing accurate translations. The Transformer's non-sequential processing and attention mechanisms are particularly advantageous in translating between Urdu and English, which differ significantly in syntax and structure. It efficiently captures contextual nuances and long-range dependencies that are crucial for accurate translation between these languages.

4 Experiments

In this section, we conduct a detailed examination of our newly developed linguistic feature, starting with an assessment of its classification accuracy scores. Subsequently, we incorporate these extracted linguistic features into the NMT framework, as outlined in Section 3.3. This integration allows us to evaluate the impact of these features on the performance of various models.

4.1 Data Sources

We utilized a substantial corpus consisting of 5,464,575 sentences, which encompass approximately 95.4 million tokens, as documented by [50] show in Table 4. This dataset not only served as the primary training set but also facilitated model validation, with 0.2% of the data reserved specifically for validation purposes. we selected a news.2010.en.shuffled.deduped dataset of English.

Table 4: Urdu dataset statistics

Corpus	Sentences	Tokens	Vocabulary
Urd Planet	4,793,736	78,045,722	536,789
Urd BBC	423,828	11,974,394	96,008
Urd library	96,240	1,692,948	44,812
Urd books	83,282	2,458,402	39,955
Urd iFastnet	24,639	427,324	28,103
Urd Awaz	22,031	388,498	20,591
Urd noman diary	18,664	375,531	19,770
Urd faisaliat	2155	49,008	5542
Total	5,464,575	95,411,827	582,795

Text Processing Approach: We extract POS and DEP features using Sandford library stanza[1] for Urdu and English. We adopted the Bert tokenizer. For linguistic processing tasks, we utilized the stanza tool [49]. The vocabulary set, consistent with the BERT-Multilingual-case[2], contains 21,128 entries. For tasks specific to machine translation, the pipeline extends to handle parallel data. sentences in both the source and target languages are tokenized using BERT same as for linguistics features, considering the linguistic characteristics of English and Urdu.

[1] https://stanfordnlp.github.io/stanza/usage.html (accessed on 25 August 2024)
[2] https://github.com/google-research/bert/blob/master/multilingual.md (accessed on 25 August 2024)

Model Configuration: For all task both the encoder and the decoder incorporating 6 layers each for both the encoder and decoder. we set $d_{model} = 786$ and $N_{head} = 8$. The dropout ratio is set to 0.3.

For processing input data, the model accommodates a maximum sequence length of 512 tokens. The more detailed show in Table 5. For model optimization, we employed the ADAM algorithm [51] with a weight decay rate of 0.1. We started with an initial learning rate of 5e-5. Each model underwent training for 2 million steps, including a 10,000-step linear warm-up phase for the learning rate. All models were trained from the ground up. We set the maximum sequence length in our models to 512 and established an overall masking ratio of 20% for the training data.

Table 5: Parameters of our approach

Parameters	Value
Optimizer	Adam
Learning rate	5e-5
Training regime	Epoch-based (32 epochs)
weight decay rate	0.1

4.2 Main Results

LKMT, which utilizes a 6-layer transformer, stands out with its significant increases in BLEU scores, achieving a +1.97 point jump for Urdu to English and a +2.42 point rise for English to Urdu translations on monolingual corpus.

Our results are illustrated in Fig. 4. Compared with others previous methods [52] and XLM [53], MASS [54] mBART [30] for language generation tasks showing in Table 6, our method not only encapsulates the cumulative knowledge gleaned from prior models but also suggests the integration of more effective linguistic processing techniques. Such an increase is indicative of a model that is adept at navigating the linguistic complexities inherent in translation tasks, setting a new standard in the field with its notable performance. The BLEU score improvements achieved by **LKMT** underscore the potential for NMT to continue its path of rapid development, promising even greater levels of language understanding and translation fidelity in the future.

Low-Resource Neural Machine Translation: In the low-resource NMT scenario, we select the Tanzil corpus, comprising 0.7 million paired sentences from the Urdu-English parallel corpus [60], to assess the efficacy of our method across various low-resource contexts. Additionally, we incorporate the religious corpus [61], which contains 13,000 sentences, for further evaluation. Employing the BPE (Byte Pair Encoding) codes acquired during the pre-training phase, we tokenize the training sentence pairs. Subsequently, we fine-tune the pre-trained model on the paired data using the Adam optimizer, with the learning rate appropriately configured. The selection of the optimal model is based on its performance accuracy on the development set. Initially, we outline the experiments focusing on low-resource NMT for Urdu and English. We continue to employ the same learned BPE codes from the pre-training phase for tokenization of the training sentence pairs, coupled with POS and DEP labels utilizing the stanza library. Furthermore, we adjust the hyperparameters for the NMT fine-tuning task, setting it to 1e-4. The results, along with the ablation study, are presented in Table 8.

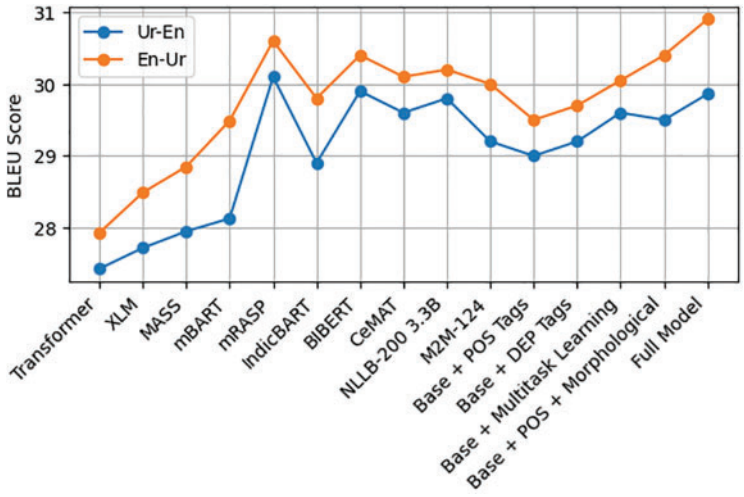


Figure 4: The comparisons between LKMT and the baseline with different NMT model on translation task. The results are reported in BLUE score

Table 6: We report the BLEU score for LKMT with 6 layer transformer, NMT, and their combinations on only two directed language pairs. Results are obtained on Tanzil corpus for ur – en and en – ur

Method	Setting	ur – en	en – ur
Transformer [52]	–	27.43	27.93
XLM [53]	4-layer Transformer	27.72	28.49
MASS [54]	6-layer Transformer	27.95	28.85
mBART [30]	4-layer Transformer	28.13	29.49
mRASP [55]	–	30.10	30.60
IndicBART [56]	–	28.90	29.80
BIBERT [57]	–	29.90	30.40
CeMAT [28]	–	29.60	30.10
NLLB-200 3.3B [58]	–	29.80	30.20
M2M-124 [59]	–	29.20	30.00
Base + POS Tags	6-layer Transformer	29.00	29.50
Base + DEP Tags		29.20	29.70
Base + Multitask Learning		29.60	30.05
Base + POS + Morphological		29.50	30.40
Full model		29.86	30.90

4.3 Comparative Experiment of POS and DEP

We mainly compare our results with pre-trained language models that use a similar amount of training data for POS and DEP. Experimental results are shown in Table 7.

Our model in the table represents the culmination of our proposed approach showing in Fig. 5. It exhibits superior performance across all measured metrics. Notably, in POS tagging, our model

achieved an F1 score of 94.07 for Urdu and 96.19 for English, which represents a marginal but consistent improvement over the strong baseline established by RoBERTa, which scored 93.63 and 95.57, respectively, on our dataset. In the more complex task of dependency parsing, our model's enhancements become more pronounced. It achieved a Labelled Attachment Score (LAS) of 85.71 for Urdu and 86.99 for English and an Unlabelled Attachment Score (UAS) of 92.14 for Urdu and 94.67 for English. These results show a marked improvement over the previously leading jPTDP-v2 model by [62], which scored 80.44 and 84.71 in LAS and 86.74 and 87.55 in UAS for Urdu and English, respectively. When considering the average performance across languages, our model maintains a lead with an average POS F1 score of 95.13 and an average DEP score (LAS and UAS) of 86.35 and 93.40, denoting the robustness of our model across both languages and tasks.

Table 7: Experimental results on our linguistic features for Urdu and English F1 score of POS and LAS/UAS score of our annotated

Model	POS		DEP						
	F1		LAS		UAS				
	ur	en	ur	en	ur	en	AVG POS	AVG LAS	AVG UAS-DEP
jPTDP-v2 [62]	93.35	95.48	80.44	84.71	86.74	87.55	–	–	–
XLM-Rbase	93.54	95.12	81.24	84.94	88.56	90.95	–	–	–
RoBERTa	93.63	95.57	82.10	85.67	89.87	91.99	–	–	–
Our	**94.07**	**96.19**	**85.71**	**86.99**	**92.14**	**94.67**	**95.13**	**86.35**	**93.40**

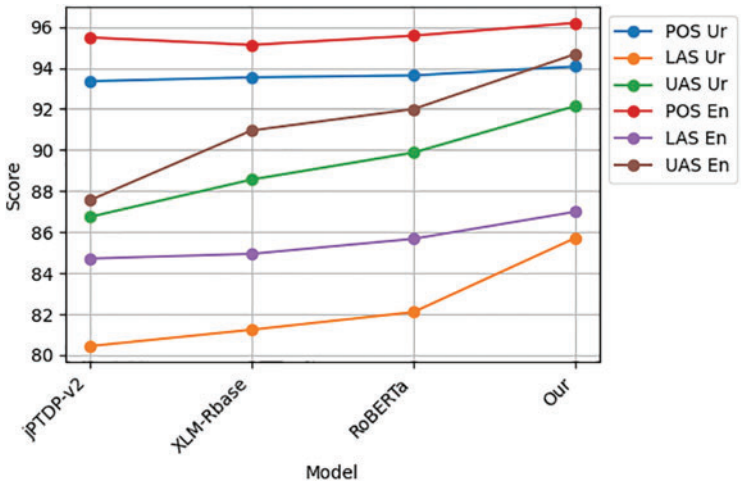


Figure 5: The comparisons between LKMT and different model on POS,DEP task. The results are reported F1 score for POS and LAS,UAS score for DEP

5 Ablation Study

In our paper, we conduct two ablation studies: one to evaluate the impact of each linguistic task in Section 5.1, and the other focusing on the NMT task. The purpose of these studies is to ascertain the individual effectiveness of each linguistic task in our model.

Our ablation study, devised to analyze the impact of different components in our Urdu-English NMT model trained on the Tanzil corpus, aimed to assess the performance of our method across diverse low-resource scenarios, as detailed in Section 4.2. With sentences from each corpus and 1.2k for validation, the study yielded valuable insights. Each configuration examined in the study contributed valuable data regarding the influence of various linguistic features on translation quality, as depicted in Table 8.

Table 8: Ablation results on using different linguistic features of NMT

Configuration	BLEU (ur – en)	BLEU (en – ur)	Average BLEU	Δ
Without POS and DEP	26.33	27.62	28.47	–
With POS only	26.90	28.03	28.54	+0.06
With DEP only	27.50	28.85	29.10	+0.62
Full model	28.40	29.70	30.15	+1.67

Without POS and DEP serves as the baseline. It uses a fundamental NMT system without any syntactic enrichment from POS or DEP embeddings. The Average BLEU score is 28.475, setting a standard for comparison with other enriched models. Incorporating POS information alone results in a minor increase in translation quality with an Average BLEU score of 28.54. The Δ Average BLEU of +0.065 suggests that POS features have a slight positive impact, likely by providing the model with better syntactic context, aiding in more accurate word choice and sentence structure. Adding DEP alone leads to a more significant improvement in the Average BLEU score, reaching 29.1. The Δ Average BLEU of +0.625 indicates that DEP features contribute more substantially to the model's performance. This suggests that understanding the grammatical relationships between words is highly beneficial for generating coherent translations. The full model, presumably integrating POS and DEP along, achieves the highest Average BLEU score of 30.15. The Δ Average BLEU of +1.675 reflects the aggregate effect of all model components and possibly additional advanced features or training techniques. This indicates that while POS and DEP embeddings are valuable, the greatest translation performance is achieved when they are part of a comprehensive NMT system. The ablation study effectively demonstrates the varying impacts of syntactic features on an NMT system's performance. POS and DEP embeddings each independently enhance the system, with their combined use leading to significant improvements. However, the full model's superior performance underscores the multiplicative benefit of a holistic approach that leverages all available linguistic and contextual information to achieve the highest translation accuracy. This analysis not only validates the importance of POS and DEP features but also encourages further research into other model enhancements that can contribute to the effectiveness of NMT systems.

5.1 Ablation Study for Linguistics Features Task

The ablation study involves adding each linguistic task to the masked language model baseline, to assess their individual contributions. The results of this study are presented in Table 8. Our findings indicate that all three types of linguistic features POS, DEP positively contribute to the overall performance. Notably, the DEP features stand out as particularly crucial, especially for downstream POS tasks. In addition, when we combine all three linguistic features in our model, we observe a further enhancement in its final performance. This comprehensive approach leads to consistent improvements across all downstream tasks, underscoring the value of integrating multiple linguistic features in the

model. The study was carried out using the Universal Dependencies (UD) dataset alongside our annotated dataset, ensuring a broad coverage of linguistic features and syntactic structures. The model architecture is based on a robust transformer framework with positional encoding, which is critical for maintaining the order of tokens in sequences.

Our ablation study meticulously deconstructed this complex relationship, shedding light on the individual and collective impacts of POS Tagging and DEP on model performance. The results, while substantial, align with the expected incremental progress characteristic of high-dimensional NLP models. Without the integration of POS and DEP features, the baseline model's performance served as a control point, registering F1 scores of 91.50 and 93.70 for Urdu and English, respectively. The Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) similarly reflected foundational translation capabilities without the benefit of syntactic or structural guidance. Incorporating POS Tagging, the model witnessed an appreciable enhancement, with average improvements of +1.25 across F1, LAS, and UAS metrics shown in Table 9. This uptick underscores the criticality of syntactic information in translation precision, as POS Tagging provides granular cues about the functional roles of words, thereby refining the model's linguistic comprehension.

Table 9: Ablation results on using different linguistic features, Urdu and English F1 score of POS and LAS/UAS score of UD (deep universal dependencies 2.8) dataset

Model	POS		DEP				
	F1		LAS		UAS		
	ur	en	ur	en	ur	en	Average
Without POS and DEP	91.50	93.70	79.00	81.50	87.00	88.50	–
With POS Only	92.80	94.80	80.50	82.50	88.50	90.00	+1.25
With DEP Only	92.87	94.90	81.50	83.00	89.00	90.50	+1.50
Full POS and DEP	93.50	95.60	82.21	83.49	89.64	91.17	+2.00

Integrating DEP features alone, the model demonstrated a further performance boost, with a +1.50 average increase in LAS and UAS. These metrics particularly benefited from the structural insights offered by DEP features, as they inform the model about the grammatical relationships between words, facilitating a more coherent translation output. The comprehensive model, equipped with both POS and DEP, emerged as the most adept variant, exhibiting an average delta of +2.00. This configuration epitomized the synergetic potential of fusing multiple linguistic features, leading to superior translation fidelity. Each feature's contribution to capturing the intricacies of language was accentuated when combined, suggesting a multiplicative rather than merely additive effect on translation quality. The ablation study's findings delineate a clear trajectory for future enhancements. The measured yet significant improvements observed reaffirm the essential roles of POS and DEP features in NMT. Furthermore, they invite exploration into the integration of additional linguistic dimensions, such as semantic role labeling and discourse analysis, to potentially catalyze further advancements in the field.

6 Conclusion

In our research, we introduce a pioneering pre-trained language model that integrates two key linguistic features POS and DEP through a linguistics enhancement approach. This model, which

includes a masked language model, undergoes multi-task tasks including linguistics features and translation task. It uniquely predicts not only the original word but also its linguistic tags for masked tokens. Our Linguistics Knowledge-Driven Multi-Task is designed to first grasp basic language elements before progressing to complex ones, a method empirically proven effective. This model significantly surpasses existing pre-trained language models, particularly in low-resource languages, highlighting the advantage of incorporating linguistic knowledge. We have conducted extensive experiments on natural language understanding tasks in Urdu and English. Crucially, our model requires just one pre-training phase and can then be fine-tuned for language generation tasks, including NMT. It has achieved BLEU scores in both Urdu-English and English-Urdu NMT, surpassing previous records by over +1.97 point for Urdu to English and a +2.42 point rise for English to Urdu translations. Looking forward, we plan to enhance our model with more linguistic features, like semantic dependency parsing, for both Urdu and English and transfer learning. Additionally, the success of our task warmup strategy prompts us to explore its application in other multi-task learning settings and different languages for NMT.

Acknowledgement: The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grant (61732005, 61972186), Yunnan Provincial Major Science and Technology Special Plan Projects (Nos. 202103AA080015, 202203AA080004).

Author Contributions: Muhammad Naeem Ul Hassan and Ying Li collected information and prepared the manuscript. Muhammad Naeem Ul Hassan, Ying Li, Shengxiang Gao, Shuwan Yang and Jian Wang drafted the article and revised it critically for important intellectual content. Cunli Mao and Zhengtao Yu finalized the manuscript. Zhengtao Yu participated in the acquisition of funding. Zhengtao Yu approved the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*.

[2] O. Bojar *et al.*, "Findings of the 2016 conference on machine translation (WMT16)," in *In First Conf. Mach. Trans., Assoc. Comput. Linguist.*, 2016, pp. 131–198.

[3] O. Bojar *et al.*, "Findings of the 2017 conference on machine translation (WMT17)," in *Proc. Second Conf. Mach. Trans.*, 2017, pp. 169–214. doi: 10.18653/v1/w17-4717.

[4] T. Kocmi and O. Bojar, "Trivial transfer learning for low-resource neural machine translation," in *Proc. Third Conf. Mach. Trans.: Res. Pap.*, 2018, pp. 244–252. doi: 10.18653/v1/w18-6325.

[5] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," Sep. 2016, *arXiv:1609.08144*.
[6] H. Hassan *et al.*, "Achieving human parity on automatic chinese to english news translation," Mar. 2018, *arXiv:1803.05567*.
[7] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Joint Conf. Natural Lang. Process.*, 2015, vol. 1, pp. 1723–1732. doi: 10.3115/v1/P15-1.
[8] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Yarman Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," in *Proc. 2016 Conf. Empir. Methods Natural Lang. Process*, 2016, pp. 268–277. doi: 10.18653/v1/d16-1026.
[9] T. -L. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," Nov. 2016, *arXiv:1611.04798*.
[10] M. Johnson *et al.*, "Google's multilingual neural machine translation system: enabling zero-shot translation," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 339–351, 2017. doi: 10.1162/tacl_a_00065.
[11] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches," Sep. 2014, *arXiv:1409.1259*.
[12] A. Vaswani *et al.*, "Attention is all you need," Jun. 2017, *arXiv:1706.03762*.
[13] J. Suzuki and M. Nagata, "Cutting-off redundant repeating generations for neural abstractive summarization," in *Proc. 15th Conf. Europ. Chapter Assoc. Comput. Linguist.*, 2017, vol. 2, pp. 291–297. doi: 10.18653/v1/E17-2.
[14] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning, Convolutional sequence to sequence learning," in *Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.
[15] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever and G. Hinton, "Grammar as a foreign language," *Adv. Neural inf. Process. Syst.*, vol. 28, pp. 2773–2781, 2015.
[16] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.
[17] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019, *arXiv:1907.11692*.
[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. Accessed: May 10, 2024. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.
[19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," Sep. 2019, *arXiv:1909.11942*.
[20] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun and Q. Liu, "ERNIE: enhanced language representation with informative entities," May 2019, *arXiv:1905.07129*.
[21] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," Jun. 2020, *arXiv:2006.03654*.
[22] Z. Mao, C. Chu, and S. Kurohashi, "Linguistically driven multi-task pre-training for low-resource neural machine translation," *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–29, 2022. doi: 10.1145/3491065.
[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Oct. 2013, *arXiv:1310.4546*.
[24] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[25] R. Weng, H. Yu, W. Luo, and M. Zhang, "Deep fusing pre-trained models into neural machine translation," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 11468–11476, 2022. doi: 10.1609/aaai.v36i10.21399.
[26] L. Han, G. Erofeev, I. Sorokina, S. Gladkoff, and G. Nenadic, "Examining large pre-trained language models for machine translation: What you don't know about it," 2022, *arXiv:2209.07417*.
[27] Z. Sun, M. Wang, and L. Li, "Multilingual translation via grafting pre-trained language models," 2021, *arXiv:2109.05256*.

[28] P. Li, L. Li, M. Zhang, M. Wu, and Q. Liu, "Universal conditional masked language pre-training for neural machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2022, pp. 6379–6391. doi: 10.18653/v1/2022.acl-long.442.

[29] J. Hwang and C. -S. Jeong, "Integrating pre-trained language model into neural machine translation," Oct. 2023, *arXiv:2310.19680*.

[30] Y. Liu *et al.*, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 726–742, 2020. doi: 10.1162/tacl_a_00343.

[31] M. Zhao, H. Wu, D. Niu, and X. Wang, "Reinforced curriculum learning on pre-trained neural machine translation models," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 9652–9659, 2020. doi: 10.1609/aaai.v34i05.6513.

[32] Z. He, G. Blackwood, R. Panda, J. McAuley, and R. Feris, "Synthetic pre-training tasks for neural machine translation," 2022, *arXiv:2212.09864*.

[33] R. Weng, H. Yu, S. Huang, S. Cheng, and W. Luo, "Acquiring knowledge from pre-trained model to neural machine translation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 9266–9273, 2020. doi: 10.1609/aaai.v34i05.6465.

[34] S. A. B. Andrabi and A. Wahid, "Machine translation system using deep learning for english to Urdu," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, 2022, Art. no. 7873012. doi: 10.1155/2022/7873012.

[35] H. S. Shavarani and A. Sarkar, "Better neural machine translation by extracting linguistic information from BERT," 2021, *arXiv:2104.02831*.

[36] Q. Li *et al.*, "Linguistic knowledge-aware neural machine translation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 12, pp. 2341–2354, 2018. doi: 10.1109/TASLP.2018.2864648.

[37] S. Chimalamarri and D. Sitaram, "Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages," *Int. J. Speech Technol.*, vol. 24, no. 4, pp. 1047–1053, 2021. doi: 10.1007/s10772-021-09865-5.

[38] S. Dewangan, S. Alva, N. Joshi, and P. Bhattacharyya, "Experience of neural machine translation between Indian languages," *Mach. Transl.*, vol. 35, no. 1, pp. 71–99, 2021. doi: 10.1007/s10590-021-09263-3.

[39] Y. Pan, X. Li, Y. Yang, and R. Dong, "Multi-source neural model for machine translation of agglutinative language," *Future Internet*, vol. 12, no. 6, 2020, Art. no. 96. doi: 10.3390/fi12060096.

[40] S. K. Sheshadri, B. Sai Bharath, A. H. N. S. Chandana Sarvani, P. R. V. Bharathi Reddy, and D. Gupta, "Unsupervised neural machine translation for english to kannada using pre-trained language model," in *13th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, 2022, pp. 1–5.

[41] A. Chakrabarty, R. Dabre, C. Ding, M. Utiyama, and E. Sumita, "Low-resource multilingual neural translation using linguistic feature-based relevance mechanisms," *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, vol. 22, no. 7, pp. 1–36, 2023. doi: 10.1145/3594631.

[42] H. Choudhary, S. Rao, and R. Rohilla, "Neural machine translation for low-resourced Indian languages," 2020, *arXiv:2004.13819*.

[43] P. Zaremoodi and G. Haffari, "Learning to multi-task learn for better neural machine translation," Jan. 2020, *arXiv:2001.03294*.

[44] Y. Li, J. Li, and M. Zhang, "Improving neural machine translation with latent features feedback," *Neurocomputing*, vol. 463, no. 3, pp. 368–378, 2021. doi: 10.1016/j.neucom.2021.08.019.

[45] M. Grosso, P. Ratnamogan, A. Mathey, W. Vanhuffel, and M. F. Fotso, "Robust domain adaptation for pre-trained multilingual neural machine translation models," in *Proc. Mass. Multiling. Nat. Lang. Understanding Workshop (MMNLU-22)*, Abu Dhabi, The United Arab Emirates, 2022, pp. 1–11.

[46] X. Liu *et al.*, "On the copying behaviors of pre-training for neural machine translation," 2021, *arXiv:2107.08212*.

[47] Y. S. Choi, Y. H. Park, S. Yun, S. H. Kim, and K. J. Lee, "Factors behind the effectiveness of an unsupervised neural machine translation system between Korean and Japanese," *Appl. Sci.*, vol. 11, no. 16, 2021, Art. no. 7662. doi: 10.3390/app11167662.

[48] J. Yang *et al.*, "Towards making the most of BERT in neural machine translation," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 9378–9385, 2020. doi: 10.1609/aaai.v34i05.6479.

[49] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," Mar. 2020, *arXiv:2003.07082*.
[50] B. Jawaid, A. Kamran, and O. Bojar, "A tagged corpus and a tagger for Urdu," *LREC*, vol. 2, pp. 2938–2943, 2014.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*.
[52] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," Apr. 2018, *arXiv:1804.07755*.
[53] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 7059–7069, 2019.
[54] K. Song, X. Tan, T. Qin, J. Lu, and T. -Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," May 2019, *arXiv:1905.02450*.
[55] Z. Lin *et al.*, "Pre-training multilingual neural machine translation by leveraging alignment information," in *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2649–2663.
[56] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. M. Khapra and P. Kumar, "IndicBART: A pre-trained model for indic natural language generation," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, Dublin , 2022, pp. 1849–1863. doi: 10.18653/v1/2022.findings-acl.145.
[57] H. Qin *et al.*, "BIBERT: Accurate fully binarized BERT," 2022, *arXiv:2203.06390*.
[58] M. R. Costa-jussà *et al.*, "No language left behind: Scaling human-centered machine translation," 2022, *arXiv:2207.04672*.
[59] N. Goyal *et al.*, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 522–538, 2021.
[60] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," *LREC*, vol. 2012, pp. 2214–2218, 2012.
[61] B. Jawaid and D. Zeman, "Word-order issues in english-to-urdu statistical machine translation," *The Prague Bull. Math. Linguist.*, vol. 95, no. 1, pp. 87–106, 2011. doi: 10.2478/v10108-011-0007-0.
[62] D. Q. Nguyen and K. Verspoor, "An improved neural network model for joint POS tagging and dependency parsing," in *Proc. CoNLL, 2018 Shared Task: Multiling. Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, 2018, pp. 81–91. doi: 10.18653/v1/K18-2008.