Enhancing Low-Resource Question-Answering Performance Through Word Seeding and Customized Refinement

Hariom Pandya, Brijesh Bhatt
Computer Engineering Department, Dharmsinh Desai University
College Road, Nadiad, 387001, Gujarat, India

***Abstract*—The state-of-the-art approaches in Question-Answering (QA) systems necessitate extensive supervised training datasets. In low-resource languages (LRL), the scarcity of data poses a bottleneck, and the manual annotation of labeled data is a rigorous process. Addressing this challenge, some recent efforts have explored cross-lingual or multilingual QA learning by leveraging training data from resource-rich languages (RRL). However, the efficiency of such approaches relies on syntactic compatibility between languages. The paper introduces the innovative method that involves seeding LRL data into RRL to create a bilingual supervised corpus while preserving the syntactical structure of RRL. The method employs the translation and transliteration of selected parts-of-speech (POS) category words. Additionally, the paper also proposes a customized approach to fine-tune the models using bilingual data. Employing the bilingual data and the proposed fine-tuning approach, the most successful model has achieved a 75.62 F1 score on the XQuAD Hindi dataset and a 68.92 F1 score on the MLQA Hindi dataset in a zero-shot architecture. In the experiments conducted using few-shot learning setup, the highest F1 scores of 79.17 on the XQuAD Hindi dataset and 70.42 on the MLQA Hindi dataset have been achieved.**

***Keywords*—Embedding learning; words seeding; bilingual dataset generation; low-resource question-answering**

I. Introduction

In recent years the pre-trained models have shown notable performance on many downstream Natural Language Processing (NLP) tasks such as Question-Answering(QA), summarization, machine translation, sentiment analysis, etc. [1], [2], [3], [4], [5], [6]. To use the pre-train models for the task other than the one on which it has been trained [7], fine-tuning on the task-specific supervised dataset is required. While the fine-tuning datasets are available in many resource-rich languages(RRLs) like English, French, and German[8], there are some languages that suffer from the bottleneck of the unavailability of supervised task-specific data.

In various fields of NLP [9], [10], [11], [12], [13], there have been efforts to tackle the situation of LRL data scarcity by annotating RRL datasets.

This paper introduces a method for integrating Hindi terms into English supervised corpora. It is noted that variations in syntactic structures between languages can detrimentally impact the effectiveness of question answering tasks. For example, English follows SVO (Subject - Verb - Object) word order whereas SOV (Subject - Object - Verb) word order is followed in the Hindi language. The proposed approach not only maintains a syntactic structure but also improves the word overlapping between question and context tokens.

It is observed that through the integration of Hindi noun category terms into English supervised data, a supervised QA dataset for LRL can be produced with minimal manual labeling required. Furthermore, it has been demonstrated that this newly generated LRL dataset can be effectively utilized alongside a tailored transfer-learning approach to attain benchmark performance levels. The methodology of transfer-learning is discussed in IV section.

Our major contributions are as follows:

1) For the LRL, a method is presented to construct a bilingual QA supervised dataset by integrating LRL words into the RRL corpora.
2) The proposed transfer-learning mechanism leverages bilingual supervised QA dataset to enable task-specific learning and language structure learning together.
3) A method is proposed to modify the position of *answer_start* during the generation of bilingual annotated data. This method relies on n-gram matching between the answer and context tokens.
4) An analysis of the translation and transliteration of nouns from the source RRL to the destination LRL is also furnished, along with its repercussions on the QA task.

The remaining paper is organized as follows. The next section describes the existing work in the directions of LM learning and QA task. The noun seeding approaches and challenges of transliteration and translation are given in Section III. The proposed approach to QA learning is mentioned in the Section IV. In Section V, the discussion revolves around the impact on performance and the analysis of the obtained results.

II. Related Work

The development of the state-of-the-art QA models ([14], [15], [16], [17], [18], [19], [20], [21], [22]) is facilitated by numerous supervised large-scale question-answering datasets. Majority of QA datasets are either labelled manually by crowdworkers (e.g., SQuAD [23], HotPotQA [24], NewsQA [25]) or originated from human inputs such as conversations or search query logs (e.g., MS MARCO [26], NaturalQuestions

[27], CoQA [28]). All these datasets are generated in English languages.

There have been recent efforts to develop LRL QA corpora involving cross-lingual and multilingual information transfer from English or other RRLs. Authors [29] approach the cross-lingual transfer learning by pre-fetching the support passages. Authors [30] an approach to retrieving related documents for a specific question first and using them as extra assistance in predicting an answer. To generate the language they looked at fine-tuning for retrieval-assisted generation models by combining pre-trained parametric and non-parametric memory. Authors [31] proposed a cross-lingual training approach that utilizes the generative architecture with resource-rich language. Authors [32] explored the direction of creating a dataset by the utilization of generative pre-trained language models in unsupervised environment followed by model fine-tuning by leveraging the guidance provided by the synthesized dataset.

There have been many efforts [33], [34] to generate supervised QA data in multilingual environments or for low-resource language(s). By translating documents into English or other RRLs before providing the answer, some research converts the cross-lingual study into a monolingual task . These approaches propagate the translation issues to the answer generation stage [35]. The approach of question classification for low-resource language proposed by [36] suggests the deep learning-based architecture can outperform traditional machine learning-based approaches for any higher level tasks. Furthermore, numerous research concentrates on creating weakly aligned data using different translation approaches. Few techniques for cross-lingual learning use the shared-encoder strategy [37], [38], [39], [40], allowing the linguistic patterns learned in one language to be transferred to all other languages without changing the model parameters. Author [41] used weakly supervised model architecture with text matching and relation detection tasks. In the approach authors, leverage the results of text retrieval to construct positive and negative text pairs followed by fine-tuning it on QA dataset.

Authors in [42] have proposed the approach to translating the whole dataset into low-resource language and during annotation analyze the quality of translation. To adopt this framework in QA, the system is allowed to ignore the question if the best probable answer seems to be invalid [43], [44]. Hence, the system will produce a subset of a resource-rich dataset.

Another way of reducing data generation efforts is to replace complete supervision with noisy weak supervision. Authors [45], [46], [47], [48], [49], [50] have explored that direction of dataset generation. In TriviaQA authors [51] uses the noisy distant supervised approach to annotate documents and answer span. The continuous user feedback-based learning approach is proposed by authors [52]. For annotation, the selection of a small subset from the whole dataset based on relevancy score is the approach used in the active learning strategy. The annotation cost could be reduced by prioritizing annotation samples. The techniques like maximizing expected model change [53], data-driven function learning [54] and model uncertainty [55] are frequently adopted for annotation.

A. Comparison of the Proposed Approach with Existing Work

In comparison to existing literature, our work extends the exploration of bilingual dataset creation for QA task by focusing on the substitution of nouns from RRL with those from LRL. In the proposed approach, the dataset has been annotated by transliterating word-subset from context, question, and answer for the whole SQuAD dataset. While prior studies have examined various aspects of bilingual dataset creation, such as translation and transliteration of text, our research specifically targets the replacement of nouns, which is a crucial component in QA systems. By offering multiple strategies based on the choice of Hindi seeding word, our study provides a nuanced approach to address the challenges inherent in bilingual dataset generation. This comprehensive analysis contributes to the existing literature by offering insights into the effectiveness of different methods in improving the performance of QA systems across languages.

It is essential to note that existing research in similar areas has certain limitations, such as overlooking specific linguistic nuances or failing to adequately address the variability in noun usage across languages. To address these gaps, our study aims to incorporate a comprehensive analysis of noun replacement strategies, considering the limitations identified in previous research. By presenting these insights, we aim to contribute to the existing literature and offer potential solutions to overcome the identified limitations, thereby enhancing the effectiveness of QA systems across languages.

III. Bilingual Dataset Generation

The primary requirement of a machine reading comprehension (MRC) system is to have exact word overlapping between answer and context. Since English and Hindi follow different word ordering, the exact translation negatively impacts word overlapping. For example, as shown in Table I, the word order of all our bilingual seeded datasets is SVO, similar to English whereas in Hindi translated it is updated to SOV. Moreover, the noun phrases in a given passage are the most plausible answer to the asked question [23], [25]. According to the article by Trischler et al. [25], the majority of the answers are common noun phrases.

Aligned with these findings, this paper explores the path of replacing nouns of RRL with LRL. Specifically, Hindi nouns are introduced into the RRL supervised dataset by substituting English nouns. Based on the choice of the Hindi seeding word our approaches are divided into three parts: 1) Replacement of all nouns with Hindi translation, 2) Replacement of all nouns with Hindi transliteration, and 3) Replacement of common nouns with Hindi translation and proper nouns with Hindi transliteration. The remaining of this section gives details of all three approaches. The subsection III-D describes our approach to adjust the position of *answer_start* index after Hindi word seeding.

A. Noun Translation

To generate the LRL supervised data, our first approach is to replace all the English noun words with their Hindi-translated version. The major issue with direct translation is that the translation may replace multiple occurrences of

Table I. Example of Hindi and English Word Ordering with Translation and Transliteration

English Text	Carolina got the ball on their own 24-yard line.
Hindi Text	(Noun Translation) कैरोलिना got the गेंद on their own 24-yard रेखा. (Noun Transliteration) कैरोलिना got the बॉल on their own 24-yard लाइन. (NN Translation & NNP Transliteration) कैरोलिना got the गेंद on their own 24-yard रेखा. (Text Translation) कैरोलिना ने अपने 24 यार्ड लाइन में बॉल प्राप्त किया।

Table II. Example with Multiple Sentence Reasoning and use of Lexical Variation(synonymy). Overlapping Words are Underlined, Synonymy used in Context is shown in Bold Text and the Answer is Highlighted in Blue Color

Question	What is the **क्षेत्रफल** of ग्लेशियर नेशनल **पार्क** ? What is the **area** of Glacier National **Park** ?
Context	ग्लेशियर नेशनल **पार्क** is an American national **उद्यान** located on the कनाडा-संयुक्त राज्य अमेरिका सीमा. The **उद्यान** is located in the उत्तर-पश्चिमी राज्य of मोंटाना in the संयुक्त राज्य and is adjacent to the कनाडा प्रांतों of अल्बर्टा and अंग्रेजों कोलम्बिया. The **उद्यान** covers an **क्षेत्र** of more than one million acres (4,000 km^2) and includes two पर्वत श्रृंखला (उप-श्रेणियाँ of the रॉकी पर्वत), more than 130 named झीलें, over 1,000 different पौधों प्रजातियां , and hundreds of प्रजातियां of वन्यजीवों. Glacier National **Park** is an American national **Park** located on the Canada–United States border. The **park** is located in the northwestern state of Montana in the United States and is adjacent to the Canadian provinces of Alberta and British Columbia. The **park** covers an **area** of more than one million acres (4,000 km^2) and includes two mountain ranges (sub-ranges of the Rocky Mountains), more than 130 named lakes, over 1,000 different plant species, and hundreds of species of wildlife.

the same word with word synonyms. Additionally, context-independent translation of proper nouns may produce a word that diverts the sentence focus from the actual linguistic meaning. Next, some example cases are mentioned to highlight these issues.

a) Replacement of Proper Noun(NNP): The NNP *"British Columbia"* is translated to "अंग्रेजों कोलंबिया" in context-independent translation. The meaning of the word "अंग्रेजो" represents *"the British community"* instead of it's actual meaning i.e. place.

b) Replacement with synonyms: The translation performance is dependent on the third-party translation tool. Situations were observed wherein the translation substitutes various instances of a word with synonyms. Table II shows the example where definite pronoun and word synonyms deflect the overlapping between the question and the answer sentence from the context. To represent the word *"park"* in the Hindi context paragraph, the lexical variations "पार्क" and "उद्यान" are used. The overlapping between context and question emphasizes the word "पार्क" but the answer statement contains a synonym word "उद्यान" as shown in a bold letter in the example. Additionally, word *"area"* is written as "क्षेत्रफल" in the question and it diminished to "क्षेत्र" in the answer sentence. Further, the overlapping noun "ग्लेशियर नेशनल पार्क" is present in a non-answer sentence and it is replaced with its definite pronoun "उद्यान" in the answer statement.

B. Noun Transliteration

To explore the impact of transliteration, in our first experiment, all words of the NNP category were replaced with their Hindi transliterated version. In the next experiment, all noun tokens (NNP, NN, NNPS, NNS) of the question, context, and answer words were transliterated.

Both experiments produce bilingual datasets for QA training. However, before starting the training in the annotated dataset, the following two situations need to be addressed: 1) similar to translation, in transliteration few Hindi word replacements have a negative impact on the quality of the transliteration. 2) after the transliteration, the invalid position of the *answer_start* needs to be updated. The next subsection describes examples that affect the transliteration quality and an approach to handling such erroneous situations. The approach of adjusting the *answer_start* is described in III-D.

a) Replacement of Common Noun(NN): Despite producing the correct transliterated version of common nouns, the seeding does not improve language learning along with task learning. Instead, such transliteration produces words that do not present in the test set that is fully in Hindi. For example, replacing the word *"agriculture"* with "कृषि" is more significant than with "एग्रीकल्चर".

b) Replacement of Proper Noun(NNP): It is observed that there were a few cases where the proper noun transliterations produced misleading Hindi words. For example, *"Main"* is converted to "मैं" in transliteration version. Degree *"MBA"* is translated to "ब", *"SUNDAYS"* is translated to "संदेश" in transliteration version. Given the limited occurrence of such misleading words, a dictionary was compiled to address the problem of incorrectly transliterated words. These words were subsequently replaced with their original English counterparts before commencing the training process

c) Erroneous POS labeling: The instances have been observed where the word *"Which"* from the question is labeled as NNP or Adjective(JJ) instead of Wh-determiner(WDT). To handle such unnecessary transliteration due to erroneous labeling, all WH words are added to the dictionary mentioned in the above step.

C. Combining Translation and Transliteration

By considering the above-discussed challenges of translation and transliteration, in our third approach to bilingual dataset generation, translation and transliteration were combined. Specifically, the approach replaces the English proper nouns with Hindi transliterated words and common nouns with translated words.

The next subsection describes our approach to adjusting the position of *answer_start* in bilingual data generation.

D. Position of Answer_start

The incorrect position of *answer_start* degrades the performance when words from the question, answer, and context are replaced with their appropriate Hindi transliteration or translation. To tackle the situation of adjusting the correct position of *answer_start* and to produce the context-aligned answer, n-gram similarity between context and answer statement as shown in Algorithm 1 is used. Here, *n* value of n-gram is equivalent to the answer length. NG represents *ngrams()* function from *nltk* library and SM is *SequenceMatcher()* from *difflib*. First, in the list *grams* all possible n-grams of context paragraph were stored. Next for each n-gram value of *grams*, the matching sequence with *answer* text was computed and all the computed results were stored in the list *score*. Maximum score from the list *score* is the most probable candidate for *context_answer*. To computer *answer_start* the find() function was used and the index of *context_answer* was calculated accordingly.

Algorithm 1 ngram similarity for adjusting *answer_start* and *context_answer*

Input: ***answer, context***
Output:*answer_start,context_answer*

$len \leftarrow length(answer)$
$grams \leftarrow NG(context.split(), len)$
$ngrams \leftarrow []$
$score \leftarrow []$
$index \leftarrow 0$
$max\_index \leftarrow 0$
$max\_score \leftarrow 0$
while $grams \neq empty$ **do**
  $score[index] \leftarrow SM(answer, grams[i])$
  $index \leftarrow index + 1$
end while
$max\_score \leftarrow max(score)$
$max\_index \leftarrow score.index(max\_score)$
$context\_answer \leftarrow ngrams[max\_index]$
$answer\_start \leftarrow find(context\_answer)$
return $answer\_start$,$context\_answer$

IV. Proposed Model Training Approach

To assess the significance of POS categories in QA, the NLTK library is employed to determine the category of every token within the questions, answers, and context passages of the SQuAD dataset. For question, answer, and context tokens Table IV indicates the count of words belonging to the 8 most frequent POS categories from answer tokens. Fig. 1 indicates the percentage-wise distribution of individual top POS categories (more than 3% of total tokens) for the question, answer, and context tokens.

Fig. 1 shows around 21.97% answer words are labeled with NNP category. Moreover, Table IV reveals that noun with their subcategories (NN, NNP, NNS) occupies 48.52% (almost half tokens) of total answer tokens. The same distribution is 30.64% in context and 30.24% in question tokens.

As shown in Fig. 2, the proposed method consists of the following steps:

1) Fine-tune the model on the Question-Answering task using the English SQuAD dataset(part-A of Fig. 2). At this stage, there is no update to the embedding weights.
2) The embedding layer of the pre-trained transformer model is trained on Hindi unlabelled text corpora[1] with MLM objective (as shown in part-B of Fig. 2). During the MLM training, all layers except embedding are kept frozen. During this step, the model is trained to learn the language structure of the Hindi language.
3) In a transfer-learning step (part-C of Fig. 2), the embedding layer of the above setup is updated with the embedding layer learned in Step 1.
4) Fine-tune the model on downstream task using bilingual labeled data of English and Hindi. The data is annotated as mentioned in Section III.
5) For a few-shot setup, further fine-tune the model on the downstream task using Hindi QA data (part-D of Fig. 2).
6) Evaluate the model performance on the Hindi QA test dataset.

As shown in Fig. 2, the embedding of a pre-trained transformer model was trained with an MLM objective. During this step, the unsupervised Hindi data was supplied with a 15% masking probability. Except for the embedding layer, the weights of layers were kept unchanged to enable language learning. To learn the QA task, our models were fine-tuned using SQuAD English dataset. The learned QA head was added with the Hindi embedding layer to form the transformer model that knows the Hindi embedding and QA task. Next, to see the impact of noun transliteration, the models were trained on the bilingual annotated dataset.

In a few shot setup, the QA learning is further fine-tuned using MLQA or XQuAD Hindi dataset, depending on the model. This step is omitted for the zero-shot learning setup. Finally, all the trained models are evaluated on the Hindi test dataset of MLQA or XQuAD evaluation set.

A. Models

The mBERT model is pre-trained in 104 languages and XLM-R is pre-trained in 100 languages. The training set of both includes the Hindi language as a subset. Based on the annotation approach mentioned in III, the following models of XLM-R$_{Large}$ and mBERT have been trained using the approach mentioned above.

[1]The experiments are conducted on the pre-trained models from the huggingface: https://huggingface.co/

Table III. Example Context Paragraph from the Article *Armenia* of SQuAD Train Set. The Example is Taken from Bilingual Dataset that has been Generated using the Annotation Method Mentioned in the Section III

Translation of common nouns and transliteration of proper nouns
कृषि accounted for less than 20 % of both net सामग्री उत्पाद and total रोज़गार before the विघटन of the सोवियत यूनियन in 1991. After आजादी, the महत्त्व of कृषि in the अर्थव्यवस्था increased markedly, its शेयर करना at the समाप्त of the 1990s rising to more than 30 % of जीडीपी and more than 40 % of total रोज़गार. This बढ़ोतरी in the महत्त्व of कृषि was attributable to भोजन सुरक्षा ज़रूरत of the आबादी in the चेहरा of अनिश्चितता during the first चरणों of संक्रमण and the गिर जाना of the non-agricultural सेक्टरों of the अर्थव्यवस्था in the early 1990s. As the economic परिस्थिति stabilized and वृद्धि resumed, the शेयर करना of कृषि in जीडीपी dropped to slightly over 20 % (2006 जानकारी), although the शेयर करना of कृषि in रोज़गार remained more than 40 %.
Hindi Translation (manual)
1991 में सोवियत संघ के विघटन से पहले कुल भौतिक उत्पाद और कुल रोजगार दोनों में कृषि का हिस्सा 20% से भी कम था। स्वतंत्रता के बाद, अर्थव्यवस्था में कृषि का महत्व स्पष्ट रूप से बढ़ गया, 1990 के दशक के अंत में इसका हिस्सा बढ़कर जीडीपी का 30% और कुल रोजगार का 40% से अधिक हो गया।। कृषि के महत्व में यह वृद्धि जनसंख्या की खाद्य सुरक्षा आवश्यकताओं के कारण संक्रमण के पहले चरणों के दौरान अनिश्चितता और 1990 के दशक की शुरुआत में अर्थव्यवस्था के गैर-कृषि क्षेत्रों के पतन के कारण हुई थी। जैसे-जैसे आर्थिक स्थिति स्थिर हुई और विकास फिर से शुरू हुआ, जीडीपी में कृषि का हिस्सा घटकर 20% (2006 डेटा) से थोड़ा अधिक हो गया, हालांकि रोजगार में कृषि का हिस्सा 40% से अधिक रहा।
Transliteration of all nouns
एग्रीकल्चर accounted for less than 20 % of both net मटेरियल प्रोडक्ट and total एम्प्लॉयमेंट before the डिसॉलूशन of the सोवियत यूनियन in 1991. After इंडिपेंडेंस, the इम्पोर्टेंस of एग्रीकल्चर in the इकॉनमी increased markedly, its शेयर at the एन्ड of the 1990s rising to more than 30 % of जीडीपी and more than 40 % of total एम्प्लॉयमेंट. This इनक्रीस in the इम्पोर्टेंस of एग्रीकल्चर was attributable to फ़ूड सिक्योरिटी नीड्स of the पापुलेशन in the फेस of अनसर्टेनिटी during the first फेसेस of ट्रांजीशन and the कलपसे of the non-agricultural सेक्टर्स of the इकॉनमी in the early 1990s. As the economic सिचुएशन stabilized and ग्रोथ resumed, the शेयर of एग्रीकल्चर in जीडीपी dropped to slightly over 20 % (2006 डाटा), although the शेयर of एग्रीकल्चर in एम्प्लॉयमेंट remained more than 40 %.
Hindi Translation (manual)
1991 में सोवियत संघ के विघटन से पहले कुल भौतिक उत्पाद और कुल रोजगार दोनों में कृषि का हिस्सा 20% से भी कम था। स्वतंत्रता के बाद, अर्थव्यवस्था में कृषि का महत्व स्पष्ट रूप से बढ़ गया, 1990 के दशक के अंत में इसका हिस्सा बढ़कर जीडीपी का 30% और कुल रोजगार का 40% से अधिक हो गया।। कृषि के महत्व में यह वृद्धि जनसंख्या की खाद्य सुरक्षा आवश्यकताओं के कारण संक्रमण के पहले चरणों के दौरान अनिश्चितता और 1990 के दशक की शुरुआत में अर्थव्यवस्था के गैर-कृषि क्षेत्रों के पतन के कारण हुई थी। जैसे-जैसे आर्थिक स्थिति स्थिर हुई और विकास फिर से शुरू हुआ, जीडीपी में कृषि का हिस्सा घटकर 20% (2006 डेटा) से थोड़ा अधिक हो गया, हालांकि रोजगार में कृषि का हिस्सा 40% से अधिक रहा।

Table IV. Distribution of Tokens as Per POS Categories. Table is Sorted in Non-ascending Order of Answer Token Counts per POS Category

Training Data	NNP	NN	JJ	CD	NNS	IN	DT	CC	Other	Total
Question	106518	145906	62734	13711	46571	108387	83005	8820	413057	988709
Answer	64966	56265	26242	24103	22215	21446	20970	9843	49600	295650
Context	294436	335086	198461	67784	147719	312017	247682	85681	847126	2535992

- MODEL-NNP: The transformer models[2] trained with Hindi MLM objective are further trained using annotated bilingual QA dataset as mentioned in the proposed approach. Here, in the annotation process, the tokens that fall in the NNP POS category are only transliterated in Hindi and other tokens are kept in English.
- MODEL-Nouns-Transliterate: The Hindi MLM-trained models are further trained using an annotated bilingual QA dataset as mentioned in the proposed approach. Here, in the annotation process, all noun tokens are transliterated in Hindi and other tokens are kept in English.
- MODEL-Nouns-Translate: The Hindi MLM-trained models are trained using our annotated bilingual QA dataset. Here, in the annotation process, all noun tokens are translated into Hindi. The other tokens are kept in English.
- MODEL-Nouns-Combined: The Hindi MLM-trained models are trained using our annotated bilingual QA dataset. Here, in the annotation process, all proper noun tokens are transliterated and common noun tokens are translated into Hindi. The other tokens are kept in English.
- MODEL-SQuAD: The Hindi MLM-trained models are trained for QA learning on the SQuAD dataset.
- MODEL-SQuAD-NNP: The MODEL-SQuAD model is further trained using annotated bilingual QA dataset as mentioned in the proposed approach. Here, in the annotation process, the tokens that fall in the NNP POS category are only transliterated in Hindi and other tokens are kept in English.
- MODEL-SQuAD-Nouns: The MODEL-SQuAD model is trained using an annotated bilingual QA dataset as mentioned in the proposed approach. Here, in the annotation process, all proper noun tokens are transliterated and common noun tokens are translated into Hindi and other tokens are kept in English.

[2]Here, MODEL is either XLM-R$_{Large}$ or mBERT

V. Experimental Setup and Result Analysis

A. Model Parameters

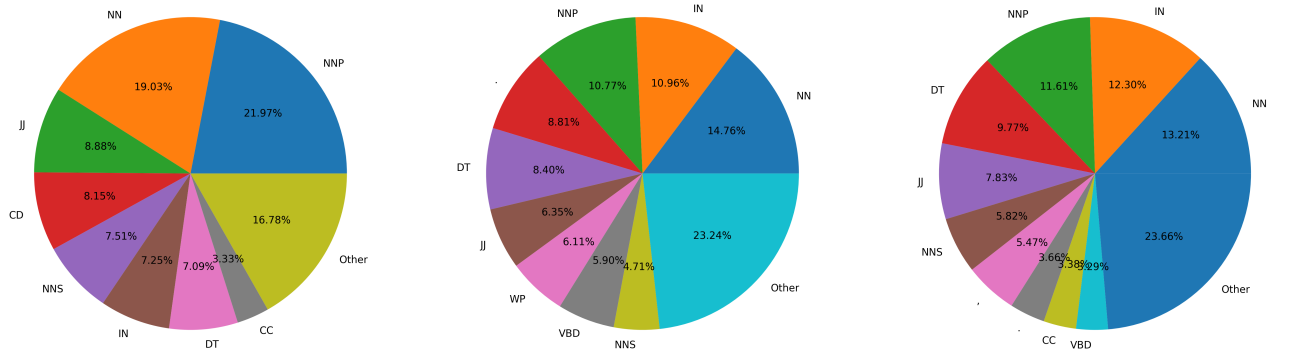
For model training, 128 doc_stride and 2e-5 learning rate were used. The Adam optimizer was used for all the experiments. It adjusts the learning rate for individual parameters by utilizing estimates of the gradients' first and second moments. By keeping track of moving averages of gradients, Adam achieves faster and more dependable convergence compared to conventional optimizers with static learning rates. It includes bias correction to counteract initialization bias and updates parameters with scaled gradients, leading to efficient updates. For QA training batch size is kept to 4 and models are trained for 2 epochs. All other hyper-parameters of our training are similar to [56]. The NVIDIA Quadro GP100 GPU was used for fine-tuning all the transformer models. Fig. 3 indicates the

Table V. F1 Score and EM of models on MLQA Hindi dataset in zero-shot setup and after few-shot Hindi XQuAD training.

Models	Zero-shot results		Few-shot results	
	F1	EM	F1	EM
mBERT†	43.8	29.87	54.3	41.03
mBERT-NNP	45.49	30.23	59.84	44.74
mBERT-Nouns-Transliterate	47.85	32.33	60.83	45.91
mBERT-Nouns-Translate	44.91	30.09	59.12	44.43
mBERT-Nouns-Combined	48.86	33.88	60.89	46.07
mBERT-SQuAD	46.54	31.39	57.64	42.36
mBERT-NNP-SQuAD	46.03	30.45	59.75	45.12
mBERT-Nouns-SQuAD	49.45	34.55	61.74	47.76
XLM-R$_{Large}$†	64.37	45.23	66.38	50.27
XLM-R$_{Large}$-NNP	65.93	46.95	69.14	53.78
XLM-R$_{Large}$-Nouns-Transliterate	66.79	48.40	70.02	53.92
XLM-R$_{Large}$-Nouns-Translate	64.98	46.19	68.83	52.81
XLM-R$_{Large}$-Nouns-Combined	67.56	48.97	70.31	54.27
XLM-R$_{Large}$-SQuAD	66.44	48.53	69.52	54.21
XLM-R$_{Large}$-NNP-SQuAD	67.10	48.89	70.19	54.34
XLM-R$_{Large}$-Nouns-SQuAD	68.92	52.24	70.42	54.51

Table VI. F1 Score and EM of models on XQuAD Hindi dataset in zero-shot setup and after few-shot Hindi MLQA training for 2 epochs.

Models	Zero-shot results		Few-shot results	
	F1	EM	F1	EM
mBERT†	48.93	34.02	70.02	55.52
mBERT-NNP	49.63	34.37	70.98	56.01
mBERT-Nouns-Transliterate	52.98	37.39	71.42	55.21
mBERT-Nouns-Translate	49.03	32.74	69.24	54 .48
mBERT-Nouns-Combined	55.47	39.91	71.50	55.39
mBERT-SQuAD	51.38	35.94	68.83	54.27
mBERT-NNP-SQuAD	50.78	34.96	71.70	56.39
mBERT-Nouns-SQuAD	56.04	40.50	71.52	55.46
XLM-R$_{Large}$†	71.79	51.53	77.38	60.36
XLM-R$_{Large}$-NNP	71.56	51.77	79.12	62.11
XLM-R$_{Large}$-Nouns-Transliterate	73.13	54.79	79.02	61.89
XLM-R$_{Large}$-Nouns-Translate	70.12	51.37	78.83	61.01
XLM-R$_{Large}$-Nouns-Combined	74.22	57.93	79.12	62.09
XLM-R$_{Large}$-SQuAD	72.05	52.54	77.36	60.53
XLM-R$_{Large}$-NNP-SQuAD	73.36	54.87	79.06	62.14
XLM-R$_{Large}$-Nouns-SQuAD	75.62	58.65	79.17	62.18



(a) Answer tokens distribution as per POS category

(b) Question tokens distribution as per POS category

(c) Context tokens distribution as per POS category

Fig. 1. Distribution of tokens as per POS categories with count of categorical tokens > 3% of total tokens. Tokens with count < 3% are labelled as Other.

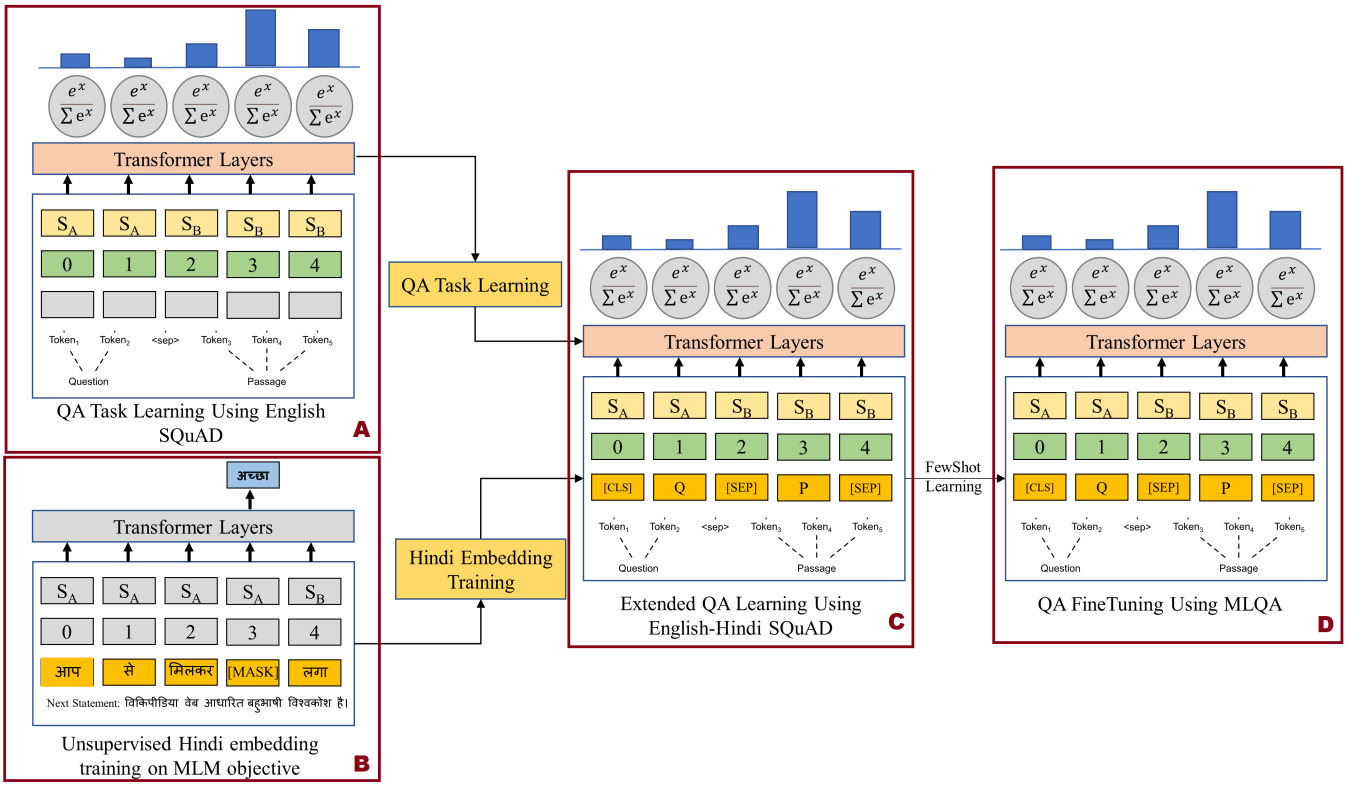


Fig. 2. Proposed approach of low-Resource hindi question-answering learning.

training loss at each phase of the model training process. The graphs are generated using 0.925 smoothing rate.

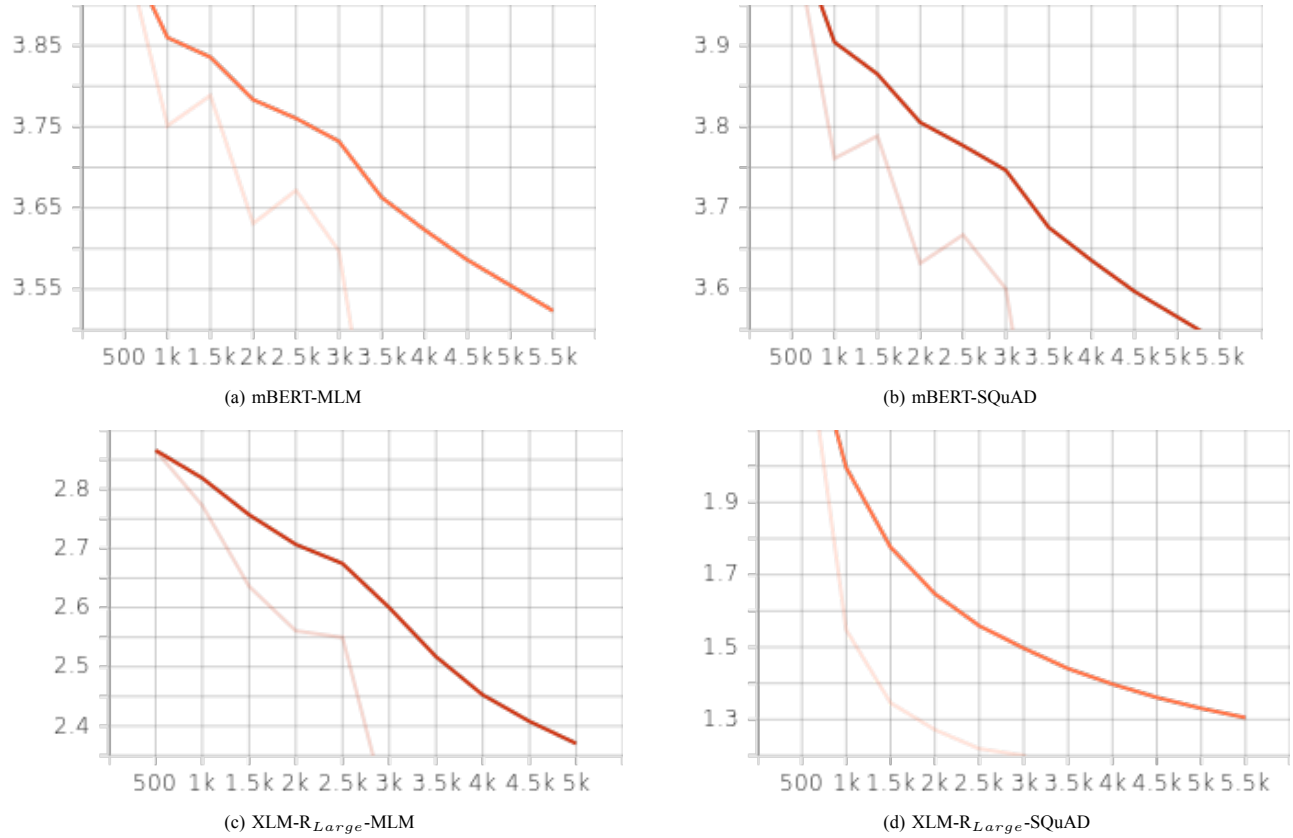
B. Datasets

The unsupervised Hindi text data for Hindi embedding training and annotated QA dataset for task learning were used. The details of the dataset that has been used are as follows:

a) Unsupervised data for Embedding training: For embedding training, 63.1M sentences from IndicCorp ([57]), 2.3M sentences from Wikipedia dump and 8.56M Hindi sentences from *Samanantar* Indic corpora collection ([58]) were combined.

To pre-process the Wikipedia dump and to clean the data, the Wikipedia Extractor tool[3] is used. It involves parsing through the XML dump of Wikipedia articles and removing the markup, templates, and other non-textual elements, leaving behind only the plain text content. This extraction process cleans the Wikipedia text, making it aligned with IndicCorp and *Samanantar* and hence, making it compatible to fine-tuned

[3]https://github.com/attardi/wikiextractor



(a) mBERT-MLM

(b) mBERT-SQuAD

(c) XLM-R$_{Large}$-MLM

(d) XLM-R$_{Large}$-SQuAD

Fig. 3. Training loss of the zero-shot learning steps on mBERT and XLM-R$_{Large}$ models.

the models for the embedding training.

b) Supervised data for Question-Answering training: On the SQuAD 1.1 ([23]) English dataset, the models were trained for single epoch. To train the model further on bilingual QA data, the model were trained on task-specific bilingual corpora. The Hindi subset of MLQA dataset ([59]) and XQuAD ([56]) were used to train the models on Hindi QA task in few-shot setup. The few-shot training was executed for two epochs on XQuAD or MLQA dataset depending on the model. Our models, trained on MLQA, are evaluated on the XQuAD Hindi dataset and visa-versa.

C. Result Analysis

Table III shows an example context paragraph from SQuAD training set. The table indicates the approach of translation of common nouns and transliteration of a proper noun has more word overlapping with the Hindi translation version as compared to the transliteration of all nouns (overlapping is highlighted in blue color). However, there are few cases where the translation-transliteration approach leads to incorrect translation as the Hindi translation of a word is independent of the statement structure and neighborhood words (highlighted in red color). For example, the translation tool has converted the word *end* to समाप्त which is the correct translation. However, for the current context, it should be समाप्ति. Table III also depicts that the synonyms are also playing vital role in the translation as mentioned in III-A0b. Some examples of synonym pairs from the table are (यूनियन-संघ), (आजादी-स्वतंत्रता), and (बढ़ोतरी-वृद्धि).

Table V indicates zero-shot and few-shot learning results on the MLQA Hindi dataset. The baseline results obtained for mBERT and XLM-R$_{Large}$ models are highlighted with † sign in the table. The models trained after all noun replacement are producing the best results. In the zero-shot configuration, XLM-R$_{Large}$ model has achieved the best (68.92/52.24) (F1/EM) scores and the best score of the mBERT model is (49.45/34.55). In the few-shot configuration when the same models are trained on XQuAD, the XLM-R$_{Large}$ model has achieved (70.42/54.51) (F1/EM) scores. The best few-shot F1 score is 1.5% better than zero-shot. Additionally, for the MLQA dataset, the best performance difference between zero-shot and few-shot setup for the mBERT is 11.29% which is just 1.5% in XLM-R$_{Large}$ model. This shows for the mBERT models, the few-shot XQuAD training helps in boosting the overall performance.

Table VI shows zero-shot and few-shot learning results on the XQuAD Hindi dataset. The baseline results obtained for mBERT and XLM-R$_{Large}$ models are highlighted with † sign in the table. In the zero-shot setup, the best performance on the XQuAD Hindi dataset has been observed by the setup of the models trained on all nouns seeding dataset, followed by SQuAD training. Specifically, XLM-R$_{Large}$ model has achieved (75.62/58.65) (F1/EM) and (56.04/40.50) (F1/EM) is the score of the mBERT for the same configuration. When

the same models were trained on MLQA to report a few-shot learning outcome, the same XLM-R$_{Large}$ model has achieved (79.17/62.18) (F1/EM) scores and (71.52/55.46) (F1/EM) is the mBERT result. The best few-shot F1 score is 3.55% better than zero-shot.

Results obtained in both tables suggest that common noun translation and proper noun transliteration have improved the performance of XLM-R and mBERT models for both MLQA and XQuAD datasets as it involves the replacement of 31.93% English tokens by its aligned Hindi version.

VI. Conclusion and Future work

In this paper, a novel method is introduced aimed at seeding low-resource words to establish a bilingual supervised QA dataset while ensuring the syntactic structure of the RRL is maintained. The proposed approach leverages the RRL and incorporates transliteration or translation techniques for nouns into the LRL. This method facilitates the creation of a robust bilingual dataset for question-answering tasks, addressing the challenge of limited resources in certain languages while preserving syntactic coherence and linguistic structure across languages. By utilizing this approach, the availability and quality of datasets for training and evaluating QA systems in bilingual settings has been enhanced, contributing to advancements in NLP and QA research. Moreover, the issue of aligning *answer_start* following the LRL word seeding process, has been addressed. Performance analysis of our approach and bilingual corpora on MLQA and XQuAD Hindi datasets has been conducted utilizing the mBERT and XLM$_{Large}$ architectures. In the zero-shot setup, our best-performing models have shown (75.62 / 58.65) (F1/EM) on the XQuAD Hindi dataset and (68.92/52.24) (F1/EM) scores on the MLQA Hindi dataset. In the few-shot setup, our best-performing models have shown (79.17/62.18) (F1/EM) on the XQuAD Hindi dataset and (70.42/54.51) (F1/EM) scores on the MLQA Hindi testset.

The proposed work opens avenues for future research in several areas. An intriguing direction is the analysis of POS category-based Hindi translation or transliteration and text annotation using all possible translated synonyms. However, it is important to acknowledge that in translation, synonyms might alter the sentence focus, even though they refer to the same concept, thus potentially introducing ambiguity. Another area worth exploring is the identification of the most suitable word replacement by translation or transliteration based on POS category, coupled with an in-depth analysis of the impact of all word replacements. This comprehensive approach would help address the limitations inherent in the current method and provide insights for improving accuracy and effectiveness. Additionally, examining the impact of word replacement by synonyms could be a promising avenue for further investigation, shedding light on potential limitations and challenges. Furthermore, regarding the mBERT model, while it demonstrates a notable improvement in few-shot learning compared to XLM-R$_{Large}$, further investigation into the underlying reasons for this disparity is warranted to gain a deeper understanding of model performance. By addressing these limitations and delving into these research directions, future studies can enhance the current work of multilingual QA systems.

References

[1] J. Liu, Y. Chen, and J. Xu, "Document-level event argument linking as machine reading comprehension," *Neurocomputing*, vol. 488, pp. 414–423, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222002867

[2] B. Ofoghi, M. Mahdiloo, and J. Yearwood, "Data envelopment analysis of linguistic features and passage relevance for open-domain question answering," *Knowledge-Based Systems*, vol. 244, p. 108574, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122002568

[3] D. Suleiman and A. Awajan, "Multilayer encoder and single-layer decoder for abstractive arabic text summarization," *Knowledge-Based Systems*, vol. 237, p. 107791, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121010005

[4] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Question-aware transformer models for consumer health question summarization," *Journal of Biomedical Informatics*, vol. 128, p. 104040, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422000569

[5] S. M. Jain, *Fine-Tuning Pretrained Models*. Berkeley, CA: Apress, 2022, pp. 137–151. [Online]. Available: https://doi.org/10.1007/978-1-4842-8844-3_6

[6] S. Tarek, H. M. Noaman, and M. Kayed, "Enhancing question pairs identification with ensemble learning: Integrating machine learning and deep learning models," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01411100

[7] C. Zhang, Y. Lai, Y. Feng, and D. Zhao, "A review of deep learning in question answering over knowledge bases," *AI Open*, vol. 2, pp. 205–215, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000292

[8] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021, https://www.sciencedirect.com/science/article/pii/S1574013720304366.

[9] M. Marchal, M. Scholman, and V. Demberg, "Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin," in *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, Nov. 2021, pp. 84–94. [Online]. Available: https://aclanthology.org/2021.codi-main.8

[10] M. A. Hedderich, L. Lange, and D. Klakow, "ANEA: distant supervision for low-resource named entity recognition," *CoRR*, vol. abs/2102.13129, 2021. [Online]. Available: https://arxiv.org/abs/2102.13129

[11] W. Ali, N. Ali, Y. Dai, J. Kumar, S. Tumrani, and Z. Xu, "Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: Association for Computational Linguistics, Apr. 2021, pp. 188–194. [Online]. Available: https://aclanthology.org/2021.wassa-1.20

[12] G. Singh, Z. Sabet, J. Shawe-Taylor, and J. Thomas, *Constructing Artificial Data for Fine-Tuning for Low-Resource Biomedical Text Tagging with Applications in PICO Annotation*. Cham: Springer International Publishing, 2021, pp. 131–145. [Online]. Available: https://doi.org/10.1007/978-3-030-53352-6_12

[13] X. Li, D. R. Mortensen, F. Metze, and A. W. Black, "Multilingual phonetic dataset for low resource speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6958–6962.

[14] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191–211, 03 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00542

[15] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. [Online]. Available: https://aclanthology.org/P17-1171

[16] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang, "R^3: Reinforced reader-ranker for open-domain question answering," *CoRR*, vol. abs/1709.00023, 2017. [Online]. Available: http://arxiv.org/abs/1709.00023

[17] S. Min, V. Zhong, R. Socher, and C. Xiong, "Efficient and robust question answering from minimal context over documents," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1725–1735. [Online]. Available: https://aclanthology.org/P18-1160

[18] J. Lee, S. Yun, H. Kim, M. Ko, and J. Kang, "Ranking paragraphs for improving answer recall in open-domain question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 565–569. [Online]. Available: https://aclanthology.org/D18-1053

[19] B. Kratzwald and S. Feuerriegel, "Adaptive document retrieval for deep question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 576–581. [Online]. Available: https://aclanthology.org/D18-1055

[20] B. Kratzwald, S. Feuerriegel, and H. Sun, "Learning a Cost-Effective Annotation Policy for Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3051–3062. [Online]. Available: https://aclanthology.org/2020.emnlp-main.246

[21] Y. Xie, W. Yang, L. Tan, K. Xiong, N. J. Yuan, B. Huai, M. Li, and J. Lin, *Distant Supervision for Multi-Stage Fine-Tuning in Retrieval-Based Question Answering*. New York, NY, USA: Association for Computing Machinery, 2020, p. 2934–2940. [Online]. Available: https://doi.org/10.1145/3366423.3380060

[22] H. A. Pandya and B. S. Bhatt, "Question answering survey: Directions, challenges, datasets, evaluation matrices," *CoRR*, vol. abs/2112.03572, 2021. [Online]. Available: https://arxiv.org/abs/2112.03572

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[24] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2369–2380. [Online]. Available: https://aclanthology.org/D18-1259

[25] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "NewsQA: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 191–200. [Online]. Available: https://aclanthology.org/W17-2623

[26] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *CoRR*, vol. abs/1611.09268, 2016. [Online]. Available: http://arxiv.org/abs/1611.09268

[27] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 08 2019. [Online]. Available: https://doi.org/10.1162/tacl_a_00276

[28] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. [Online]. Available: https://aclanthology.org/Q19-1016

[29] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 874–880. [Online]. Available: https://aclanthology.org/2021.eacl-main.74

[30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[31] X. Cao, Y. Zhao, and B. Shen, "Improving and evaluating complex question answering over knowledge bases by constructing strongly supervised data," *Neural Computing and Applications*, vol. 35, no. 7, pp. 5513–5533, 2023.

[32] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "Zerogen: Efficient zero-shot learning via dataset generation," 2022.

[33] F. Ture and E. Boschee, "Learning to translate for multilingual question answering," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 573–584. [Online]. Available: https://aclanthology.org/D16-1055

[34] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 547–564. [Online]. Available: https://aclanthology.org/2021.naacl-main.46

[35] J. Zhu, Q. Wang, Y. Wang, Y. Zhou, J. Zhang, S. Wang, and C. Zong, "NCLS: Neural cross-lingual summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3054–3064. [Online]. Available: https://aclanthology.org/D19-1302

[36] E. Trandafili, N. Kote, and G. Plepi, "Question classification in albanian through deep learning approaches," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140385

[37] Y.-H. Lin, C.-Y. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig, "Choosing transfer languages for cross-lingual learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3125–3135. [Online]. Available: https://aclanthology.org/P19-1301

[38] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-shot cross-lingual transfer with meta learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4547–4562. [Online]. Available: https://aclanthology.org/2020.emnlp-main.368

[39] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 87–94. [Online]. Available: https://aclanthology.org/2020.acl-demos.12

[40] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. [Online]. Available: https://aclanthology.org/2022.acl-long.62

[41] J. Ma, Q. Chai, J. Huang, J. Liu, Y. You, and Q. Zheng, "Weakly supervised learning for textbook question answering," *IEEE Tr on Image Processing*, vol. 31, pp. 7378–7388, 2022.

[42] A. F. T. Martins, M. Junczys-Dowmunt, F. N. Kepler, R. Astudillo, C. Hokamp, and R. Grundkiewicz, "Pushing the Limits of Translation Quality Estimation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 205–218, 07 2017. [Online]. Available: https://doi.org/10.1162/tacl_a_00056

[43] J. Zhao, Y. Su, Z. Guan, and H. Sun, "An end-to-end deep framework for answer triggering with a novel group-level objective," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1276–1282. [Online]. Available: https://aclanthology.org/D17-1131

[44] A. Kamath, R. Jia, and P. Liang, "Selective question answering under domain shift," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5684–5696. [Online]. Available: https://aclanthology.org/2020.acl-main.503

[45] P. Petrushkov, S. Khadivi, and E. Matusov, "Learning from chunk-based feedback in neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 326–331. [Online]. Available: https://aclanthology.org/P18-2052

[46] K. Clark, M.-T. Luong, C. D. Manning, and Q. Le, "Semi-supervised sequence modeling with cross-view training," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1914–1925. [Online]. Available: https://aclanthology.org/D18-1217

[47] C. Liang, J. Berant, Q. Le, K. D. Forbus, and N. Lao, "Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 23–33. [Online]. Available: https://aclanthology.org/P17-1003

[48] S. Iyer, I. Konstas, A. Cheung, J. Krishnamurthy, and L. Zettlemoyer, "Learning a neural semantic parser from user feedback," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 963–973. [Online]. Available: https://aclanthology.org/P17-1089

[49] I. Gur, S. Yavuz, Y. Su, and X. Yan, "DialSQL: Dialogue based structured query generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1339–1349. [Online]. Available: https://aclanthology.org/P18-1124

[50] Z. Yao, Y. Su, H. Sun, and W.-t. Yih, "Model-based interactive semantic parsing: A unified framework and a text-to-SQL case study," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5447–5458. [Online]. Available: https://aclanthology.org/D19-1547

[51] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: https://aclanthology.org/P17-1147

[52] B. Kratzwald and S. Feuerriegel, "Learning from on-line user feedback in neural question answering on the web," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 906–916. [Online]. Available: https://doi.org/10.1145/3308558.3313661

[53] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 51–60.

[54] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 595–605. [Online]. Available: https://aclanthology.org/D17-1063

[55] A. Siddhant and Z. C. Lipton, "Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2904–2909. [Online]. Available: https://aclanthology.org/D18-1318

[56] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4623–4637. [Online]. Available: https://aclanthology.org/2020.acl-main.421

[57] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4948–4961. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.445

[58] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra, "Samanantar: The largest publicly available parallel corpora collection for 11 indic languages," 2021.

[59] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7315–7330. [Online]. Available: https://aclanthology.org/2020.acl-main.653