Towards Building Question Answering Resources for TELUGU

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

VEMULA RAKESH KUMAR
2019701027
rakesh.kumar@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Towards Building Question Answering Resources for TELUGU**" by **VEMULA RAKESH KUMAR**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Manish Shrivastava

Dedicated to my
Family and Friends

Acknowledgments

As I prepare to submit my Master's thesis, I wish to take a moment to extend my heartfelt gratitude to the individuals who have been instrumental in my journey at IIIT-Hyderabad.

First and foremost, I would like to convey my sincerest appreciation to my mentor, Prof. Manish Shrivastava. Without his unwavering guidance and support, the completion of this work would not have been possible. I am profoundly thankful for his invaluable expertise, insightful suggestions, and constructive feedback, all of which have played a pivotal role in shaping my research endeavors. Prof. Shrivastava's provision of resources, encouragement, and motivation has been instrumental in helping me bring my research to fruition.

I also wish to extend my sincere appreciation to Mani Kanta Sai Nuthi and Ala Hema, who have been integral parts of both my life and my journey within the institute. My heartfelt thanks go out to my fellow labmates, Gopichand, Lokesh, Pavan, Priyanka, and Ashok for their unwavering support and for encouraging a positive and conducive learning environment.

Last, but certainly not least, I want to convey my deepest gratitude to my parents, Laxmi and Rajesham, for their unwavering love and support throughout my journey.

Abstract

Natural Language Processing (NLP) is a cutting-edge field of artificial intelligence that empowers computers to understand and work with human language. It plays a crucial role in various practical applications that impact our daily lives. Its applications span a wide spectrum, including machine translation, text summarization, question-answering, and sentiment analysis, among others. These applications have left an indelible mark on society, giving rise to chatbots, voice assistants like Alexa and Siri, and recommendation systems on platforms such as YouTube, Netflix, and Hotstar.

Within the realm of NLP, Question Answering (QA) represents a pivotal field. QA involves the intricate interplay of query formulation, document retrieval, and, at times, document summarization. Recent strides in this domain have given rise to comprehensive end-to-end systems capable of extracting precise answers from extensive text collections. These systems can be trained on expansive datasets, tailored either to a specific domain (referred to as closed domain) or spanning a wide array of subjects (referred to as open domain).

Nonetheless, the efficacy of QA systems heavily hinges upon the curation of meticulous datasets, a process that demands substantial manual labor and resources. This challenge becomes particularly pronounced when considering Indian languages, where access to dependable and substantial datasets remains limited. In such contexts, the exclusive reliance on data-driven neural network approaches proves inadequate. Therefore, the imperative arises to strengthen available data resources and introduce innovative, data-independent techniques. Recent state-of-the-art models and new datasets have advanced many NLP areas, especially, Machine Reading Comprehension (MRC) tasks have improved with the help of datasets like SQuAD (Stanford Question Answering Dataset). But, large-high-quality datasets are still not a reality for low-resource languages like Telugu to record progress in MRC.

This thesis intends to explore a resource-scarce language like Telugu, one of the widely spoken Dravidian languages in India, with native speakers of around 96 million. In this thesis, we present a Telugu Question Answering Dataset - TeQuAD with the size of 82k parallel triples created by translating triples from the SQuAD. We also introduce a few methods to create similar Question Answering datasets for the low-resource languages. Then, we present the performance of our models which outperform baseline models on Monolingual and Cross-Lingual Machine Reading Comprehension (CLMRC) setups.

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

Language stands out as a crucial element that distinguishes humans from other species. Conversely, machines lack the innate ability to grasp words or text in natural language, as they process information using binary code 1s and 0s. However, enabling machines to comprehend natural language holds the promise of simplifying daily human life. By harnessing vast amounts of processed data generated through global communication in diverse languages, it becomes possible to instruct machines to understand and extract meaning from human languages. In this context, the concept of NLP comes into play as it serves a pivotal role in facilitating communication between computers and human beings.

NLP involves the development of algorithms and models that enable machines to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant. NLP combines elements of computer science, linguistics, and cognitive psychology to bridge the gap between human communication and computer understanding. Nevertheless, the construction of these NLP applications demands a substantial volume of processed (or annotated) text data and significant computational resources. Unfortunately, such processed text data is predominantly accessible for only a handful of widely spoken languages like English. This restriction confines the utility of NLP applications to specific linguistic communities. To broaden the reach of NLP applications and cater to diverse user groups, there is a pressing need to extend the development of these applications to a more extensive array of languages. This thesis serves as a catalyst, motivating the creation of NLP systems specifically for the Indian language Telugu.

NLP has lots of applications like Language Generation, Machine Translation, Speech Recognition, Information Retrieval, Summarization, Sentiment Analysis and Question Answering. QA is a pivotal task in NLP, where the model is specifically crafted to generate answers in response to questions posed by users. This process encompasses various components, including information retrieval, natural language understanding, and natural language generation. QA systems have diverse applications, spanning from chatbots and search engines to information retrieval systems. With the ongoing expansion of unstructured data and an increasing demand

for accessing specific, essential information, QA is transforming into a central and challenging task within the realm of NLP.

1.1 The Importance of Addressing QA task in Telugu

Telugu, one of the Dravidian languages predominantly spoken in the Indian states of Andhra Pradesh and Telangana, showcases a rich linguistic heritage. Addressing the task of QA in Telugu is crucial due to the language's unique linguistic characteristics. These distinctive features encompass not only the script and phonetics but also syntactic structures, semantic nuances, and culturally embedded expressions.

Telugu is recognized for its morphological richness, characterized by agglutination and inflection. Words in Telugu can undergo various morphological changes based on tense, mood, gender, and other grammatical factors. Incorporating these morphological complexities is crucial for accurate question understanding and answer generation. The syntactic structures in Telugu, including word order and sentence formation, may differ from those in languages that have been extensively studied in the context of QA systems. Adapting QA models to comprehend and interpret Telugu syntax ensures effective information extraction from questions and documents.

Tailoring QA systems to accommodate these characteristics not only enhances the accuracy and effectiveness of the system but also contributes to linguistic inclusivity, ensuring that Telugu speakers can fully benefit from advancements in NLP technology.

1.2 Motivation for Telugu QA

Extensive research has explored into the foundational realms of NLP in Telugu, covering morphological, lexical, and syntactic dimensions. Despite this comprehensive exploration, a conspicuous void emerges in the domain of QA tools and resources for Telugu. The current landscape primarily spotlights endeavors directed at tasks such as POS Tagging, NER tagging, and parsing. However, a distinct lack of emphasis is evident in the dedicated development of tools and resources tailored specifically to unravel the intricate challenges posed by QA in the Telugu language.

The motivation behind undertaking this research endeavor stems from the recognition of the profound impact that NLP technologies can have on linguistic communities, especially those underserved by existing resources. Telugu, with its rich linguistic heritage and a vast community of speakers, presents a compelling case for the development of dedicated QA resources tailored to its unique linguistic characteristics. By embarking on the journey towards building QA resources for Telugu, we aim to address several critical gaps in the existing landscape of NLP research. Firstly, the lack of comprehensive QA datasets in Telugu impedes the development

and evaluation of robust QA systems for this language. Secondly, existing QA models trained on high-resource languages often exhibit suboptimal performance when applied to low-resource languages like Telugu, highlighting the need for language-specific resources and methodologies.

Moreover, the development of QA resources for Telugu holds immense potential to encourage digital inclusion and accessibility for Telugu-speaking communities. By enabling native speakers to interact with technology in their mother tongue, we can empower individuals with limited proficiency in English and promote the preservation and expansion of Telugu language and culture in the digital age. Creating QA datasets for Telugu involves building language-specific resources. This, in turn, supports the development of tools, models, and datasets that are crucial for further research in Telugu NLP. QA research in Telugu can have a positive impact on educational resources and also adds to the global body of knowledge in NLP. It provides insights into language-specific challenges and solutions, contributing to a more comprehensive understanding of linguistic diversity in the context of NLP.

In light of these considerations, this research endeavors to contribute towards the creation of foundational QA resources for Telugu, laying the groundwork for future advancements in Telugu NLP and facilitating more inclusive and impactful human-computer interaction experiences for Telugu speakers.

1.3 Challenges in Creating Telugu Corpus for NLP Applications

Creating datasets for low-resource languages presents a set of unique challenges that can impact the quality and effectiveness of NLP models. Here are some common challenges faced while creating datasets for low-resource languages:

1. **Limited Availability of Data**

   Low-resource languages often suffer from a scarcity of available data. This limitation can hinder the creation of diverse and comprehensive datasets, affecting the performance of NLP models that rely on extensive training data.

2. **Lack of standard Guidelines**

   Low-resource languages may lack standardized linguistic resources and guidelines, making it challenging to create consistent and well-annotated datasets. This can lead to variations in data quality and hinder the development of robust models.

3. **Lack of language Models**

   Existing language models, such as pre-trained embeddings or word vectors, may not be readily available for low-resource languages. This scarcity can impact the initialization and performance of models, especially those relying on pre-trained representations.

4. **Morphological Complexity**

   Languages with complex morphological structures, common in low-resource languages, pose challenges in tokenization, stemming, and lemmatization. Handling these complexities requires linguistic expertise and specialized preprocessing techniques.

5. **Annotation Challenges**

   Annotating data for low-resource languages can be labor-intensive and costly. The lack of established annotation standards and guidelines may result in inconsistencies and difficulties in achieving inter-annotator agreement.

In addition to the challenges mentioned earlier, specific obstacles related to the QA task will be explored in the upcoming chapters.

1.4 Our Contributions

We have introduced TeQuAD, a comprehensive Telugu Question Answering Dataset, which consists of 82,000 parallel triples. These triples were meticulously generated by translating corresponding triples from the widely acclaimed SQuAD dataset, ensuring linguistic consistency and relevance in the Telugu context.

Furthermore, our endeavor extends beyond dataset creation. We have developed and presented guidelines and methodologies that establish a structured framework for the creation of Question Answering datasets specifically tailored for low-resource languages like Telugu. These guidelines aim to streamline the dataset creation process, ensuring high-quality outputs that meet the unique linguistic needs and challenges of low-resource languages.

In our commitment to accessibility and collaboration within the research community, we have made the generated datasets publicly available. By offering open access to these resources, we aim to facilitate widespread exploration, experimentation, and advancement in the field of question-answering, particularly in the context of low-resource languages.

Our dataset is available here[1].

[1] https://github.com/ltrc/TeQuAD

1.5 Organization of Thesis

This thesis is organized into five chapters. The current chapter introduces the QA tasks in Telugu, explains the rationale behind choosing these challenges, and describes the obstacles encountered in constructing data resources for Telugu. Chapter 2 offers an extensive overview of previous research conducted on question-answering systems. Chapter 3 explains corpus collection methodologies. The experiments and results of our models for Telugu language QA are clarified in Chapter 4. Finally, we present conclusions and a few ideas for future works in Chapter 5.

Chapter 2

Literature Review

In this chapter, we will explore various types of QA tasks in Section 2.1, will provide an overview of different word representation techniques in Section 2.2, followed by a discussion on the modeling approaches utilized in this thesis in Section 2.3. We will explain the evaluation metrics employed for QA in Section 2.4. Finally, related work that happened in the field of QA will be explained in Section 2.5.

2.1 Introduction to QA

Question Answering is one of the hot research topics in NLP. A QA system can be thought of as a computer program that can extract answers from a knowledge base or natural language documents, where the questions asked by the QA system are also posed in natural language. Here are some common types of QA systems.

2.1.1 Open-Domain QA

Open-domain QA [13], [10] refers to the task of automatically answering questions posed by users using information from a wide range of sources, without restricting the domain or topic of the questions. Open-domain QA aims to provide comprehensive answers by drawing upon vast repositories of knowledge, such as the entire internet, encyclopedias, books, articles, and other textual sources. The system must possess a broad understanding of natural language and the ability to interpret and analyze diverse types of text to generate accurate responses to a wide variety of questions.

Open-domain QA systems typically employ large-scale knowledge bases, search engines, and sophisticated algorithms to retrieve and analyze relevant information in response to user queries. These systems may utilize machine learning and deep learning approaches to improve performance and adapt to user needs over time. The goal of open-domain QA is to provide users with accurate, informative, and contextually relevant answers to their questions, regardless of the topic or domain. This task has numerous applications, including virtual assistants, search

engines, question-answering chatbots, and information retrieval systems, and continues to be an active area of research and development in the field of NLP.

2.1.2 Closed-Domain QA

Closed-domain QA refers to the task of automatically answering questions within a specific domain or topic, where the information is limited to a predefined set of documents or knowledge sources. Closed-domain QA focuses on providing accurate responses within a constrained domain or subject area. In closed-domain QA, the system is typically trained and tailored to a specific domain, such as medicine, finance, or law. The knowledge sources used by the system are restricted to a predefined corpus of documents or databases relevant to the chosen domain. This focused approach allows for more precise and targeted answers to questions within the domain.

Systems often leverage domain-specific ontologies, knowledge graphs, or structured databases to extract and organize information relevant to the domain. They may also incorporate domain-specific language models and specialized algorithms to understand and process text within the domain's context. The primary goal of closed-domain QA is to provide accurate and relevant answers to questions within the specified domain, enabling users to quickly access information and make informed decisions within their area of interest or expertise. This task has applications in various domains, including customer support systems, expert systems, and domain-specific information retrieval tools. While closed-domain QA systems are limited to specific domains, they often achieve higher accuracy and efficiency compared to open-domain QA systems due to their focused nature and specialized knowledge sources. However, they may require manual curation and updating of domain-specific data and knowledge sources to ensure relevance and accuracy over time.

2.1.3 Factoid QA

Factoid QA [2] systems are designed to provide precise, factual answers to questions based on verifiable information. These systems excel at addressing queries that have straightforward answers, typically in the form of short phrases or single words. They are expert at answering questions related to names, dates, locations, measurements, and other factual details.

- *Q1: Where was Virat Kohli born ?*
- A1: Delhi
- *Q2: What jersey number does Virat Kohli wear?*
- A2: 18

2.1.4 Non-Factoid QA

Non-factoid QA [8] systems are designed to address questions that require more complex or nuanced responses beyond simple factual information. These systems deal with questions that may have subjective or opinion-based answers. These questions often involve interpretation, evaluation, or speculation rather than verifiable facts. They may require deeper understanding, inference, or reasoning to generate meaningful responses.

- *Q1: What leadership qualities does Virat Kohli exemplify in his role as captain of the Indian cricket team ?*

- A1: Virat Kohli demonstrates strong leadership qualities through his assertiveness, tactics, ability to inspire teammates, and relentless pursuit of excellence, shaping the team's performance on and off the field.

- *Q2: How has Virat Kohli's aggressive playing style influenced the culture of Indian cricket?*

- A2: Virat Kohli's aggressive playing style has instilled a sense of confidence, competitiveness, and resilience in the Indian cricket team, redefining the team's approach to win matches.

This work is related to a Closed-Domain QA, in which, like reading comprehension, the question-answering system is expected to understand a given paragraph (context) with the associated question and then extract the answer to the question. The answer (as a segment or span of text) text should be located and extracted from the provided paragraph itself. Stanford Question Answering Dataset (SQuAD) [37] is one of the popular closed-domain QA Dataset in English. SQuAD contains triples. A triple is a collection of a paragraph, a question, and an answer. There is a paragraph (or para) and a question related to the para (context), such as the answer to that question lies in the para(context). The figure below shows an example of a triple from SQuAD. The SQuAD dataset was created via crowdsourcing and is exclusively available in English. To compile this dataset, high-quality Wikipedia articles spanning various topics were gathered, and crowd workers were tasked with generating multiple questions based on these articles. In total, SQuAD comprises 100,000 parallel triples, each consisting of paragraphs, questions, and corresponding answers. In many cases, several question-answer pairs were produced for a single paragraph. Notably, the answer to each question is contained within its corresponding paragraph, and it can range from a single word to a lengthy-phrase comprising multiple words. To accurately pinpoint the answers, they are represented using span indices, indicating their positions within the paragraphs. This approach helps resolve ambiguity when multiple instances of potential answer phrases exist within the paragraphs. See 2.1 for a SQuAD instance example.

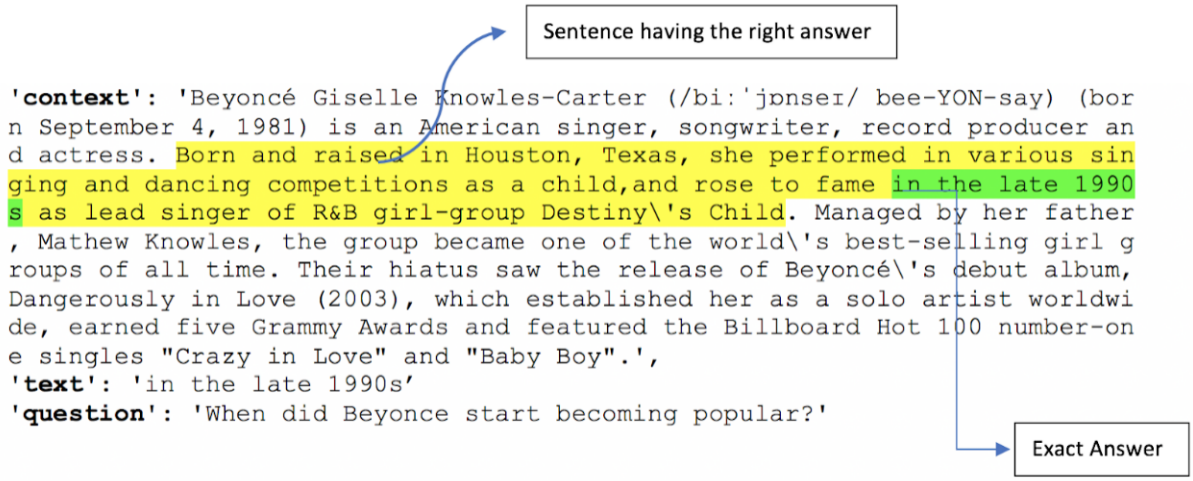

Figure 2.1: Example for Span Extractive Question Answering.

2.2 Word Embeddings

Word embeddings in NLP refer to the representation of words as vectors in a continuous vector space. This technique is designed to capture the semantic relationships and contextual meanings of words based on their usage patterns in large corpora of text. Instead of representing words as discrete symbols, word embeddings enable machines to understand the semantic similarity between words, offering a more nuanced understanding of language. Word representations can be broadly classified into Non-Contextual and Contextual word representations.

2.2.1 Non Contextual Embeddings

Non-contextual embeddings represent words as fixed vectors, irrespective of their context within a sentence or document. Each word is assigned a single vector representation, which remains constant across different contexts. These embeddings are typically trained on large corpora using unsupervised learning, capturing word semantics based on global co-occurrence patterns. Non-contextual embeddings are suitable for tasks where word meaning remains consistent regardless of context. Common applications include sentiment analysis, named entity recognition, and document classification. Some non-contextual embeddings are explained below.

1. **Word2Vec**

   Word2Vec [25], [31] represents words as dense vectors in a continuous vector space. Each word is assigned a vector with real-numbered components, capturing semantic relationships between words. Word2Vec [39] offers two training methods: Skip-gram and Continuous Bag of Words (CBOW). Skip-gram method, which predicts context words given

a target word. CBOW method, which predicts a target word given its context words. Word2Vec employs a shallow neural network with a single hidden layer. The network is trained to predict context words or target words, depending on the chosen method Skip-gram or CBOW.

2. **GloVe**

   GloVe [34], which stands for Global Vectors for Word Representation, aims to capture global statistical information about the co-occurrence patterns of words in a large corpus of text. The key idea behind GloVe is to generate word vectors based on the global context of word co-occurrences, emphasizing the semantic relationships between words.

3. **Fasttext**

   FastText extends the principles of Word2Vec by incorporating subword information, making it particularly effective for handling morphologically rich languages and out-of-vocabulary (OOV)words. FastText addresses the limitation of Word2Vec when dealing with OOV and rare words. Unlike Word2Vec, which struggles to effectively represent such words, FastText employs a unique approach. It learns representations for words by considering their constituent n-grams. For example, the word 'word2vec' is broken down into tri-grams like 'wor', 'ord', 'rd2', 'd2c', '2ve' and 'vec'. The embedding for 'word2vec' is then derived by summing up the vectors of these n-grams. This methodology enables FastText to effectively capture representations for OOV and rare words, enhancing its ability to handle a broader range of vocabulary.

2.2.2 Contextual Embeddings

Contextual embeddings [4] consider the surrounding context of a word when generating its vector representation. The vector for a word varies depending on its context within a sentence, allowing the model to capture different meanings of the word in different contexts. These models are pre-trained on large datasets with tasks like language modeling or masked language modeling, enabling them to learn contextual information. Contextual embeddings excel in tasks that require an understanding of word meaning in different contexts. Applications include machine translation, question answering, and tasks where word sense disambiguation is crucial.

- **BERT Embeddings**

  BERT, short for Bidirectional Encoder Representations from Transformers, generates contextual embeddings by considering the entire context of a word within a sentence. It is based on the transformer architecture, which allows it to efficiently process and understand relationships within long-range dependencies in text. BERT's [35], [46] attention mechanism allows it to assign different weights to different parts of the input sequence, focusing more on relevant words and capturing intricate relationships between words. In

contrast to uni-directional models, which analyze text linearly, either from left to right or right to left, BERT processes the entire word sequence simultaneously, embodying a bidirectional approach.

2.3 Modeling Strategies

The input text is converted into vector representation then it is processed with a model that predicts the output. The model can be machine learning or neural network-based architecture. In this thesis, for the question-answering task, we have explored the BERT and mBERT language models.

2.3.1 BERT Language Model

The BERT model considers both the left and right contexts of each word in a sentence, allowing it to understand the meaning of words concerning their surrounding context. This bidirectional approach enhances the model's ability to comprehend nuances and ambiguities in language. BERT is built upon the TRANSFORMER architecture, which facilitates parallel processing of input sequences through self-attention mechanisms. This architecture enables BERT to capture long-range dependencies in text and effectively model relationships between words. As mentioned in [18], [30] BERT is pre-trained using two primary tasks: masked language modeling (MLM) and next sentence prediction (NSP). BERT is trained on both MLM (50%) and NSP (50%) at the same time. The BERT language model has significantly advanced the field of NLP by providing powerful contextual representations of text.

2.3.1.1 Transformers

The Transformer is a deep learning architecture explained in [45], [23]. It has become a foundational building block for various NLP tasks, including machine translation, text summarization, and question-answering. The Transformer architecture replaces recurrent neural networks (RNNs) and convolutional neural networks (CNNs) with a self-attention mechanism, enabling it to capture long-range dependencies in sequential data more effectively. The Transformer consists of two main components: an encoder and a decoder. The encoder processes the input sequence, while the decoder generates the output sequence. Each component contains multiple layers of self-attention mechanisms and feed-forward neural networks. The core innovation of the Transformer is the self-attention mechanism, which allows the model to weigh the importance of each word in the input sequence when generating representations. Self-attention enables the model to capture dependencies between words regardless of their distance in the sequence

2.3.1.2 Masked Language Modelling

The goal of MLM is to train a language model to predict masked or missing words within a sentence or sequence of text. The model is then trained to predict the original masked tokens given the context of the surrounding words. In other words, it learns to estimate the probability distribution of each word in the vocabulary given its context. MLM is an effective pre-training objective because it encourages the model to learn contextual representations of words and phrases.

2.3.1.3 Next Sentence Prediction

The goal of NSP is to train a language model to understand the relationship between two consecutive sentences in a text. Pairs of consecutive sentences are sampled from a large corpus of text. These pairs consist of a "context" sentence and a "target" sentence, where the target sentence is the sentence that immediately follows the context sentence in the original text. The model is then trained to predict whether the target sentence follows the context sentence in the original text. Specifically, it learns to classify whether the target sentence is the actual next sentence in the text or a randomly chosen sentence from the corpus

2.3.2 mBERT Language Model

mBERT, short for Multilingual BERT, is a variant of the BERT language model that has been pre-trained on text from multiple languages. In this thesis, we have used the Telugu-English mBERT language model. mBERT is trained on a large corpus of text from diverse languages, covering a broad spectrum of linguistic diversity. This multilingual training data enables mBERT to capture language-agnostic patterns and generalize across different languages. It leverages cross-lingual transfer learning, where knowledge learned from one language can be transferred to improve performance on tasks in other languages. This transfer learning approach enables mBERT to achieve competitive performance across a wide range of languages, even for languages with limited training data.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Figure 2.2: F1-Score

$$EM\ Score = \frac{\text{Number of questions with exact match}}{\text{Total number of questions}} \times 100\%$$

Figure 2.3: EM Score

2.4 Evaluation Metrics

The QA system's performance is assessed using metrics such as the F1 score and Exact Match (EM) score. The following sections provide a concise explanation of these evaluation metrics.

1. **F1 Score**

   The F1 score, also known as the F-score or F-measure, is a metric used to evaluate the accuracy of a classification model. It combines the precision and recall of the model into a single value, providing a balanced assessment of the model's performance. The formula for calculating the F1 score shown in 2.2 is where Precision measures the proportion of true positive predictions among all positive predictions made by the model. Recall measures the proportion of true positive predictions among all actual positive instances in the dataset. The F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating poor performance. In the context of a QA system, the F1 score is used to assess how accurately the system identifies correct answers to questions. A higher F1 score indicates better performance in retrieving relevant information and answering questions accurately.

2. **EM Score**

   The EM score is a metric used to evaluate the performance of question-answering systems. It measures the percentage of questions for which the model provides an exact and correct answer, without any errors or deviations as shown in 2.3. In essence, the EM score is calculated by comparing the predicted answer from the model to the ground truth answer for each question in the evaluation dataset. If the predicted answer exactly matches the ground truth answer, the question is considered correctly answered, and the EM score for that question is 1. If there is any discrepancy between the predicted and ground truth answers, the EM score for that question is 0. The overall EM score for the QA system is then calculated as the percentage of questions for which the model provides an exact match to the ground truth answers. A higher EM score indicates better performance of the QA system in providing precise and accurate answers to questions.

2.5 Related Work on QA

Several datasets such as SQuAD[37], NewsQA dataset [44] and CNN/Dailymail [12], etc fulfill the necessity of resources in English for QA tasks. Although these datasets helped in attaining enormous progress for this specific language in NLP, other languages are still unexplored in this area due to the scarcity of high-quality annotated datasets in corresponding languages. While the generation of reading comprehension corpora in other languages is costly and time-consuming, few works such as Lim et al. [27], Efimov et al. [20], Cui et al. [15], d'Hoffschmidt et al. [19] developed RC datasets natively. Clark et al. [14] presented a question-answering dataset covering 11 typologically diverse languages including Telugu. Question answering system corpora [28], [21] is available for Bengali, Telugu, and Tamil. Statistical and ML-based QA systems are explored for a few Indian languages [32], [6], [40], [17], [7]. Many automatic QA systems [48], [42] have been developed over the past few decades. In the Indian context, QA systems for several Indian languages are constructed using methods such as keyword identification, pattern matching, and sentence ranking to extract answers [22], [3], [41].

Few others propose methods to boost the functioning of the model in low-resource settings. Hsu et al. [24] explored zero-shot cross-lingual transfer learning on reading comprehension tasks and suggested that translation from source to target languages is not necessary. Bornea et al. [9] presents a translation-based data augmentation mechanism to improve multilingual transfer learning.

Liu et al. [29] and Cui et al. [16] talk about leveraging translated information from high-resource languages to perform well in low-resource languages. Cui et al. [16] presented several back translational approaches for cross-lingual experiments. They have also discussed techniques to align the answer phrases in the target language. Stating the disadvantages of such approaches and the necessity to overcome them, they introduced a novel model called 'Dual BERT', which can learn semantic information from bilingual QA pairs and utilize the learned knowledge to improve MRC in low-resource languages.

Yuan et al. [49] introduced phrase boundary supervision tasks to improve the answer boundary detection capability in the low-resource MRC models which are trained with training data from high-resource languages to exploit cross-lingual transfer learning. Post-correction methods to improve the span of the extracted answer are addressed in Reddy et al. [38]. They added additional layers on top of a pre-trained transformer-based language model to re-examine and modify the predicted answers.

Chapter *3*

TeQuAD: Telugu Question Answering Dataset

3.1 Introduction

Cutting-edge models and high-caliber datasets have significantly enhanced various domains within NLP, particularly MRC tasks, which have flourished with the emergence of resources like SQuAD. However, despite these advancements, the availability of top-tier MRC datasets remains scarce for numerous low-resource languages such as Telugu. To tackle this challenge, our strategy involves leveraging existing high-quality datasets from well-resourced languages to achieve respectable MRC performance in low-resource languages. We utilized the SQuAD dataset as a foundation and translated it to create QA resources in Telugu, our target language. However, due to linguistic disparities, the positioning of answers within the target paragraphs changes post-translation. Consequently, employing simplistic matching methods, like identifying translated answer phrases within the translated paragraphs, proves ineffective for answer span extraction. Given the critical role of the span extraction process in the QA data creation pipeline, we delved into and deliberated on various heuristic matching techniques. By implementing such techniques, we successfully constructed a parallel Telugu-English QA dataset comprising 82k triples. The primary objective behind creating this parallel QA dataset is to harness the benefits of cross-lingual reasoning.

MRC is one of the key tasks in NLP, where we test the ability of machines to understand and answer the questions using provided textual knowledge. In common Machine Reading Comprehension tasks, for a given query, the machine needs to extract the answer from the context (paragraph) in the form of span indices. we have seen the rise of Reading Comprehension datasets [33], [26], [43], [47]. A popular large-scale annotated reading comprehension dataset - SQuAD Rajpurkar et al. [37], revolutionized the research interest in this area of English. And though decent research work has been done in MRC for a few Indian languages, languages like Telugu, which is a Dravidian language, still need similar resources for such a Natural Language Understanding task.

Creating an RC dataset of good quantity & quality is difficult, requires manpower, and is time-consuming. For a few languages, the dataset is created by translating SQuAD and using a

few matching techniques to extract the span indices of answers in the target language (Carrino et al. [11], Abadani et al. [1], Artetxe et al. [5]). For others, the dataset is created by using the methodology followed in the creation of SQuAD (Lim et al. [27], Efimov et al. [20], Cui et al. [15], d'Hoffschmidt et al. [19]).

Our idea is to introduce a few heuristics-based approaches to create the datasets for a low-resource language (Telugu) using the resources from a high-resource language via translation. An obvious challenge is to extract the span of the answers in the translated Contexts. With translation, due to language divergences, the position and structure of the answer in the context will vary in the translated language, making it difficult to use straight-forward approaches, like translation candidate matching, to find the position of the answer in the context. We focused on the span extraction process, which is crucial for such a dataset creation after translation. We applied these methods to SQuAD v1.1 and created TeQuAD, an MRC dataset for Telugu consisting of 82k parallel Telugu-English triples (Paragraphs, Questions, and Span indices of Answers). The intention to create a parallel dataset is to exploit the advantage of Cross-lingual reasoning. We also introduce a supervised approach to extract the span of the most probable answer from the target paragraph. This span extractor can later aid in data augmentation for MRC in low-resource languages. In cases where the heuristics do not work, our supervised method performs better than the matching techniques due to its ability to consider contextual semantic information using pre-trained language models.

Both monolingual and cross-lingual setups of multilingual Bidirectional Encoder Representations from Transformers (BERT) were trained on TeQuAD and evaluated on TiDyQA [14] and on two other test datasets, which we created manually by correcting a few samples from translated SQuAD and by using Wikipedia articles respectively.

3.2 Corpus Creation

3.2.1 Data Source

We considered SQuAD as the data source to create the QA dataset in the Telugu language. Because of its high quality and quantity, and adaptability to recent implementations of deep learning models, SQuAD has been chosen as the source to create QA datasets in different languages [11], [1], [5]. SQuAD was created through crowd-sourcing. Crowd workers posed questions on paragraphs extracted from English Wikipedia articles. 1,00,000+ triples were constructed, each triple consisting of a paragraph, question, and span indices of answer. Span indices of answer indicate the position of the answer phrase in the paragraph.

3.2.2 Data Creation

A simple and cost-efficient technique to create a dataset for an NLP task is to translate a well-annotated existing dataset. When it comes to MRC tasks, SQuAD is a favorite for

its quality and adaptability to recent implementations of deep learning models. This span extractive QA data is created from English Wikipedia articles by crowd workers. More than 100000 triples were generated in SQuAD1.1. A triple consists of a Question, an Answer to the question in the form of span indices, and a Context where the answer can be found.

We translated the English SQuAD triples to the Telugu language using online Google translator[1], obtaining translated triples consisting of translated Telugu paragraphs, questions, and answers.

After translation, a well-known issue is the difficulty in the extraction of the span indices of the translated answers. Considering the different possibilities of the translated Telugu answer phrase's presence in the translated Telugu context, we followed multiple techniques to extract the span for answer phrases. The purpose of following different techniques is to create as much synthetic data as possible for Telugu MRC. We also present a supervised span extraction technique to handle the cases where rule-based methods fail.

3.2.2.1 Matching

We used matching algorithms like cosine similarity [36] and fuzzy search with a threshold value of greater than 0.7. A window sliding through the translated Telugu context computes the matching score between the phrase inside the window and the translated Telugu answer phrase. Samples are considered if such a matching phrase (matching score greater than the threshold) is found in the translated Telugu context, and else ignored. There might be a possibility of the presence of multiple answer phrases in the context. For such samples, we considered the index of the actual English answer phrase among its repetitions present in the English context and selected the corresponding index as the answer from repetitions of the Telugu answer in the translated Telugu context. For example, if the word 'apple' is present 3 times in the English context, and if the answer is the second repeated instance, then we consider the second repetition of the Telugu word ఆపిల్ (Āpil) as the answer in the translated Telugu context.

3.2.2.2 Explicit Position Indicator

If phrases with matching scores greater than the threshold value are not found, span extraction for such samples is skipped when matching techniques are applied. These skipped samples are considered to extract the answer span using the explicit position indicator technique. Using span indices of the answers, English answer phrases in the English paragraphs are identified and marked with a special symbol ('|'). After translating English paragraphs to Telugu, everything gets translated, but the symbol remains unchanged. Locating the symbol, we can identify the translated answer phrases in translated Telugu paragraphs. See table 3.1 for example.

Using the above-discussed methods, we obtained 82,605 English-Telugu parallel triples, creating a Telugu Question Answering dataset - TeQuAD. See table 3.2 for a parallel English-

[1] https://translate.google.co.in/

English text	Translated Telugu text (ISO 15919)
China Unicom 's service in Wenchuan and four nearby counties was cut off , with more than 700 | **towers** | suspended.	Veñcuvān mariyu samīpanlōni nālugu kauṇṭīlalō cainā yunikām sēva nilipivēyabaḍindi, 700ki paigā | **ṭavarlu** | nilipivēyabaḍḍāyi.

Table 3.1: Using Special Symbols to Indicate Answer Span for Extraction.

	English	**Telugu (ISO 15919)**
Context	China Mobile had more than "**2,300**" base stations suspended due to power disruption or severe telecommunication "**traffic congestion**". Half of the wireless communications were lost in the Sichuan province . China Unicom 's service in Wenchuan and four nearby counties was cut off , with more than 700 towers suspended.	Vidyuttu antarāya lēdā tīvramaina ṭelikamyūnikēṣan "**ṭrāphik raddī**" kāraṇaṅgā cainā mobail "**2,300**" ki paigā bēs'sṭēṣanlanu nilipivēsindi. Sicuvān prānslō saga vairles kamyūnikēṣanlu pōyāyi. Veñcuvān mariyu samīpanlōni nālugu kauṇṭīlalō cainā yunikām sēva nilipivēyabaḍindi, 700ki paigā ṭavarlu nilipivēyabaḍḍāyi.
Question 1	Besides power disruption , what caused telecommunications to be suspended ?	Vidyuttu antarāyantō pāṭu, ṭelikamyūnikēṣanlanu nilipivēyaḍāniki kāraṇamēmiṭi?
Span	16 - 17	5 - 6
Answer	traffic congestion	ṭrāphik raddī
Question 2	How many base stations are suspended?	Enni bēs sṭēṣanlu saspeṇḍ cēyabaḍḍāyi?
Span	5 - 5	10 - 10
Answer	2,300	2,300

Table 3.2: Representation of QA pairs in parallel corpora

Telugu instance in TeQuAD. For evaluation and experiments, we have created two different test datasets. We later use these test datasets to analyze the performance of Telugu MRC models and present the results.

1. **Translated & Corrected dataset**

   2000 English triples from the dev set of SQuAD1.1 are translated to Telugu and corrected manually. A set of guidelines explained in 3.2.4 to correct the translated Telugu context, questions, and answers.

2. **Wiki dataset**

   Similar to SQuAD, we created this data from Wikipedia articles. Randomly selected Wikipedia articles are split into paragraphs. From 125 Telugu Wikipedia paragraphs, 947

QA pairs are created manually by framing questions with answer types such as Person, Location, Date/Time, Quantities, Clauses, Verb phrases, Adjective phrases, and others. A minimum of five and a maximum of ten questions were created for each paragraph/context. To make it challenging, for fair evaluation, multiple types of queries were posed while creating the dataset.

3.2.3 Span Extractor

While creating the reading comprehension dataset for Telugu, TeQuAD, we used the discussed techniques to retrieve span indices for the translated Telugu answers. These techniques might fail in cases where

1. The translated answer might not exist in the translated paragraph. After the translation, there is a possibility that information about the answer phrase might have been lost or transformed into a different word form. See 3.1 for example.

2. Multiple instances of the translated answer might be in the translated paragraph. See 3.2 for example

3. A random phrase in the Telugu paragraph returns a greater matching score than the actual answer phrase when compared with the translated Telugu answer phrase. See 3.3 for example

To handle the span extraction of answers in such cases, we introduced a supervised span extraction approach. We employed the Dual BERT method proposed in [16] for Chinese MRC. Along with parallel QA pairs, we additionally added answer phrases as inputs to the model, and the span indices of the answers are predicted as output. See 3.4. Dual BERT considers the contextual semantic information from both Telugu-English parallel triples and can identify the answer phrases even if they exist in different forms in the translated Telugu paragraphs. As this model relies on contextual information to identify the answer phrase, it can find the correct instance of the answer phrase, even if there are multiple instances of the answer phrase present in the Telugu paragraph. Along with the translated Telugu answer, information about English answers will help predict span indices of the exact Telugu answer phrases, avoiding the retrieval of partial answer phrases.

To handle such cases, we introduce a supervised method to extract span indices for the translated answers. We use the Dual BERT approach proposed in Cui et al. [16], but along with parallel QA pairs, we also pass their parallel answers as input to the model and the span indices of the answers are predicted (See Figure 3.4). Due to its ability to exploit semantic information from both Telugu-English parallel triples, it can identify a modified variant of the answer phrase in the translated context, even if the translated answer phrase is not present in the translated context completely.

English Context:
. . . Trade liberalization may shift economic inequality from a global to a domestic scale . . .
English Question:
What scale does trade liberalization shift economic inequality from ?
English Answer:
Global
Translated Telugu Context:
. . . వాణిజ్య సరళీకరణ ఆర్థిక అసమానతలను ప్రపంచ స్థాయి నుంచి దేశీయ స్థాయికి మార్చవచ్చు . . .
(. . . Vāṇijya saraḷīkaraṇa ārthika asamānatalanu prapan̄ca sthāyi nun̄ci dēśīya sthāyiki mārcavaccu . . .)
Translated Telugu Question:
వాణిజ్య సరళీకరణ ఏ స్థాయి నుండి ఆర్థిక అసమానతను మారుస్తుంది ? (Vāṇijya saraḷīkaraṇa ē sthāyi nuṇḍi ārthika asamānatanu mārustundi)
Plausible Telugu Answer: ప్రపంచ స్థాయి (prapan̄ca sthāyi)
Translated Telugu Answer: గ్లోబల్ (Glōbal)

Figure 3.1: Example for the absence of translated Answer in the translated Context. Both *'prapan̄ca sthāyi'* and *'glōbal'* share the similar meaning.

Translated Telugu Context:
. . . పూర్తిగా పెట్టుబడిదారీ ఉత్పత్తి పద్ధతిలో కార్మికుల వేతనాలు ఈ సంస్థలు లేదా యజమాని ద్వారా నియంత్రించబడవు, కానీ మార్కెట్ ద్వారా, వేతనాలు ఏ ఇతర మంచి కోసం ధరల మాదిరిగానే పనిచేస్తాయి. అందువలన, వేతనాలు నైపుణ్యం యొక్క మార్కెట్ ధర యొక్క విధిగా పరిగణించబడతాయి . . .
(. . . Pūrtigā peṭṭubaḍidārī utpatti pad'dhatilō kārmikula vētanālu ī sansthalu lēdā yajamāni dvārā niyantrin̄cabaḍavu, kānī mārket dvārā, vētanālu ē itara man̄ci kōsaṁ dharala mādirigānē panicēstāyi. Anduvalana, vētanālu naipuṇyaṁ yokka mārkeṭ dhara yokka vidhigā parigaṇin̄cabaḍatāyi . . .)
Translated Telugu Question:
పూర్తిగా పెట్టుబడిదారీ ఉత్పత్తి పద్ధతిలో వేతనాలను ఏది నియంత్రిస్తుంది ? (Pūrtigā peṭṭubaḍidārī utpatti vidhānanlō vētanālanu ēdi niyantristundi ?)
Translated Telugu Answer:
మార్కెట్ (mārkeṭ)

Figure 3.2: Example for multiple instances of Answer in the Context

Translated Telugu Context:
. . . కొంతమంది చట్టపరమైన అవిధేయతలు సామాజిక ఒప్పందం యొక్క చెల్లుబాటుపై వారి విశ్వాసం కారణంగా శిక్షను అంగీకరించడం తమ బాధ్యత అని భావిస్తారు . . .
(. . . Kontamandi caṭṭaparamaina avidhēyatalu sāmājika oppandaṁ yokka cellubāṭupai vāri viśvāsaṁ kāraṇaṅgā śikṣanu aṅgīkarin̄caḍaṁ tama bādhyata ani bhāvistāru . . .)
Translated Telugu Question:
చట్టపరమైన అవిధేయతలు దేనిపై విశ్వాసం కారణంగా శిక్షను అంగీకరిస్తారు ?
(Caṭṭaparamaina avidhēyatalu dēnipai viśvāsaṁ kāraṇaṅgā śikṣanu aṅgīkaristāru ?)
Plausible Telugu Answer: సామాజిక ఒప్పందం యొక్క చెల్లుబాటుపై (Sāmājika oppandaṁ yokka cellubāṭupai)
Translated Telugu Answer: సామాజిక ఒప్పందం యొక్క ప్రామాణికత (Sāmājika oppandaṁ yokka prāmāṇikata)

Figure 3.3: Example for partial matching answer scenario.

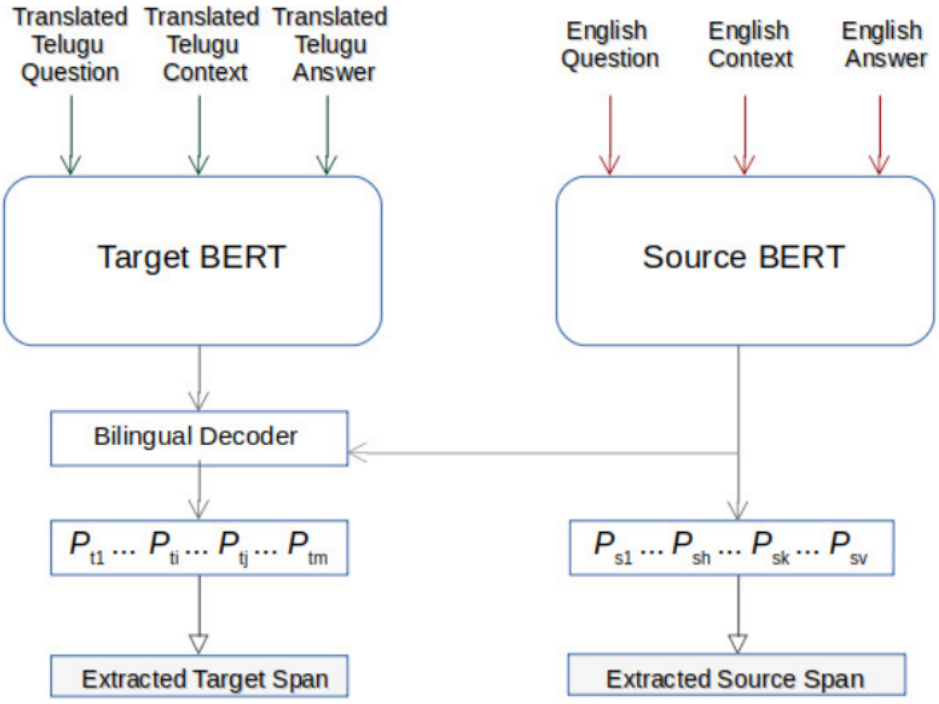

Figure 3.4: Architecture of Span Extractor

Unlike the above-mentioned matching techniques, this model can identify the correct instance of the answer in the translated context, even if there are multiple instances present. In addition to the translated Telugu answers, information from English Answers will help the model to retrieve span indices of the complete Telugu answer phrases in the translated contexts.

As this is a supervised method that needed training data, we considered 82k parallel triples from TeQuAD consisting of span indices obtained by using matching techniques, for training. For evaluation, we use the Translated & Corrected test dataset where span indices of the Telugu answers are manually corrected. We pass the translated Telugu answers as input to the model and evaluate the predictions with corrected Telugu answers. The experimental setup is similar to the Cross-lingual section. Results attained show the performance of **88%** F1 Score and **73%** EM Score. Besides apparent advantages, such supervised methods need sufficient resources to perform well and predict a span even if the answer information is not present in the context (e.g. might have been lost in machine translation).

3.2.4 Correction Guidelines

We have used Google NMT for translating triples. Although Google NMT is efficient and the quality of the translations is good, the present translation machines are not smart enough to generate accurate translations for low-resource languages like Telugu. After a translation is generated by the machine, there has to be a human correction to ensure the piece of translation is grammatically correct, comprehensible and carries the exact information present in the English text (by deducting/appending the knowledge obtained/lost from the translation of the English text.) Data in TeQuAD is in triple form. All three components of triples (para, question, and answer) are translated from English to Telugu. To correct a translated triple, one must correct the translated Telugu para, translated Telugu question, and then extract the Telugu answer from the translated para, as present in the triple parallelly:

- Correct the Telugu paragraph.
- Correct the Telugu question according to the Telugu paragraph.
- Extract the Telugu answer according to the Telugu question from the Telugu paragraph.

While correcting a paragraph or a question, two essential aspects should be considered:

1. **Adequacy:** The meaning/knowledge provided in the English para/question should be preserved in the Telugu para/questions.

   Example: The translation shown below didn't preserve all the meaning/knowledge from the original sentence (English) i.e, lower adequacy.

   Enligsh Sentence: "Dogs can smell and hear better than humans, but cannot see well in color because they are color blind."

Translated Telugu Sentence: కుక్కలు మనుషులకన్నా మంచిగా వినగలవు, కానీ రంగులో బాగా చూడలేవు ఎందుకంటే అవి కలర్ బెల్ండ్. (Kukkalu manuṣulakannā mañcigā vinagalavu, kānī raṅgulō bāgā cūḍalēvu endukaṇṭē avi kalar blaiṇḍ.)

Corrected Telugu Sentence: కుక్కలు మనుషులకన్నా మంచిగా వాసన చూడగలవు మరియు వినగలవు, కానీ రంగులను బాగా చూడలేవు ఎందుకంటే అవి కలర్ బెల్ండ్. (Kukkalu manuṣulakannā mañcigā vāsana cūḍagalavu mariyu vinagalavu, kānī raṅgulanu bāgā cūḍalēvu endukaṇṭē avi kalar blaiṇḍ)

2. **Fluency:** The structure/syntax of the Telugu para/question should be proper/readable.

   Example: The translation shown below suffers from Fluency.

   Enligsh Sentence: "Not only teeth start shining with this . But one also gets relief from stinking breath."

   Translated Telugu Sentence: పళ్ళు మాత్రమే దీనితో మెరుస్తూ ఉండవు. కానీ దుర్వాసన నుండి ఒకరు కూడా ఉపశమనం పొందుతారు. (Paḷḷu mātramē dīnitō merustū uṇḍavu. Kānī durvāsana nuṇḍi okaru kūḍā upaśamana pondutāru.)

   Corrected Telugu Sentence: దీని వల్ల పళ్ళు మెరవడం మాత్రమే కాదు. ఒకరు దుర్వాసన నుండి కూడా ఉపశమనం పొందుతారు. (Dīni valla paḷḷu meravaḍa mātramē kādu. Okaru durvāsana nuṇḍi kūḍā upaśamana pondutāru.)

3.2.4.1 Correcting Paras

- Read and understand the English paragraph.
- Then again read each sentence in the English para and correct the corresponding sentence in Translated Telugu para.
- Then scan the whole Telugu Para and Correct it, If necessary. For example,

English Paragraph: "If one wants to grow a certain type of apple it is not possible to do this by planting a seed from the wanted type. The seed will have DNA from the apple that the seeds came from, but it will also have DNA from the apple flower that pollinated the seeds, which may well be a different type. And Apple farming needs low-temperature conditions. Because of such reasons, Apple farming is tougher than it seems."

Translated Telugu Paragraph: ఒక నిర్దిష్ట రకం ఆపిల్ను పెంచుకోవాలనుకుంటే, కోరుకున్న రకం నుండి ఒక విత్తనాన్ని నాటడం ద్వారా దీన్ని చేయడం సాధ్యం కాదు. విత్తనాలు విత్తనాల నుండి వచ్చిన ఆపిల్ నుండి DNA ను కలిగి ఉంటాయి, కానీ విత్తనాలను పరాగసంపర్కం చేసిన ఆపిల్ పువ్వు నుండి DNA కూడా ఉంటుంది, ఇది వేరే రకం కావచ్చు. ఇటువంటి కారణాల వల్ల వ్యవసాయం ఆపిల్ను కనిపించే దానికంటే కఠినమైనది. (Oka nirdiṣṭa rakaṁ āpilnu peñcukōvālanukuṇṭē, kōrukunna rakaṁ nuṇḍi oka vittanānni nāṭaḍaṁ dvārā dīnni cēyaḍaṁ sādhyaṁ kādu. Vittanālu vittanāla nuṇḍi vaccina āpil nuṇḍi DNA nu kaligi uṇṭāyi, kānī vittanālanu parāgasamparkaṁ

cēsina āpil puvvu nuṇḍi DNA kūḍā uṇṭundi, idi vērē rakaṁ kāvaccu. Iṭuvaṇṭi kāraṇāla valla vyavasāyaṁ āpilnu kanipiñcē dānikaṇṭē kaṭhinamainadi.)

Corrected Telugu Paragraph: ఒక నిర్దిష్ట రకం ఆపిల్ను పెంచుకోవాలనుకుంటే, కావాల్సిన రకం నుండి ఒక విత్తనాన్ని నాటడం ద్వారా సాధ్యం కాదు. విత్తనం ఏ ఆపిల్ నుండి వచ్చిందో ఆ ఆపిల్ యొక్క డీ.ఎన్.ఏ విత్తనంలో ఉంటుంది, కానీ దాంతో పాటు ఏ ఆపిల్ పువ్వు ద్వారా విత్తనాలు పరాగసంపర్కం చెందబడ్డాయో ఆ ఆపిల్ డీ.ఎన్.ఏ కూడా ఉంటుంది, అది వేరే రకానికి చెందినది కావొచ్చు. మరియు ఆపిల్ సాగుకు తక్కువ ఉష్ణోగ్రత పరిస్థితులు అవసరం. ఇటువంటి కారణాల వల్ల ఆపిల్ వ్యవసాయం కనిపించే దానికంటే కఠినమైనది. (Oka nirdiṣṭa rakaṁ āpilnu peñcukōvālanukuṇṭē, kāvālsina rakaṁ nuṇḍi oka vittanānni nāṭaḍaṁ dvārā sādhyaṁ kādu. Vittanaṁ ē āpil nuṇḍi vaccindō ā āpil yokka ḍī.En.Ē vittananlō uṇṭundi, kānī dāntō pāṭu ē āpil puvvu dvārā vittanālu parāgasamparkaṁ cendabaḍḍāyō ā āpil ḍī.En.Ē kūḍā uṇṭundi, adi vērē rakāniki cendinadi kāvoccu. Mariyu āpil sāguku takkuva uṣṇōgrata paristhitulu avasaraṁ. Iṭuvaṇṭi kāraṇāla valla āpil vyavasāyaṁ kanipiñcē dānikaṇṭē kaṭhinamainadi.)

In the above example, Sentence 1, Sentence 2, and Sentence 4 in Translated Telugu Para are suffering in the readability aspect. The readability of Sentence 4 is low and it affects the meaning of the sentence. So, while correcting the sentences, both readability and adequacy are maintained simultaneously. They are interlinked! Some information related to Sentence 3 is lost in the translation and should have to be included in the correction. Observe that in the above-corrected paragraph, both adequacy and fluency are maintained.

3.2.4.2 Correcting Questions

- Read and understand the English Question.
- Modify the Telugu question according to the English question and Telugu para.
- Make sure that the questions are relevant to their respective paras. Telugu Question should have to be properly constructed and have to ask the same query as its corresponding English Query does

Example-1: Even though the Telugu Translated Query is asking the same thing as the English Question, the Readability of Translated Query is bad.

English Query: "What fruit are we talking about?"

Incorrectly Translated Telugu Query: ఏ ఫలం మనం గురించి మాట్లాడుతున్నాం? (Ē phalaṁ manaṁ guriñci māṭlāḍutunnāṁ?)

Corrected Telugu Query: మనం ఏ ఫలం గురించి మాట్లాడుతున్నాం? (Manaṁ ē phalaṁ guriñci māṭlāḍutunnāṁ?)

Example-2: Information loss in the Translated Telugu Query.

English Query: "Who is the ruler of the Turkish empire after the Azzaruddin's death?"

Incorrectly Translated Telugu Query: అజ్జరుద్దీన్ టర్కి సామ్రాజ్యం యొక్క పాలకుడు ఎవరు? (Ajjaruddīn ṭarkiṣ sāmrājya yokka pālakuḍu evaru?)

Corrected Telugu Query: అజ్జరుద్దీన్ మరణం తరువాత టర్కి సామ్రాజ్యం యొక్క పాలకుడు ఎవరు? (Ajjaruddīn maraṇa taruvāta ṭarkiṣ sāmrājya yokka pālakuḍu evaru?)

3.2.4.3 Extracting Answers

- Identify the sentence related to the Corrected Telugu Query in the Corrected Telugu Para.
- Copy and paste the identified sentence.
- Find the answer to the query in the identified sentence.
- Copy and paste the answer phrase.
- If it is not an accurate translation of the English answer, modify the Answer phrase.

Answers obtained by translation are partial, in a few cases incorrect. Such answers should be corrected based on the corresponding Corrected Telugu Questions, Corrected Telugu Paras, and English Answers. The answer should be obtained from its paragraph (Corrected Telugu Para) and must be recorded. In short, for a given query, the answer should have to be in its parallel paragraph. Along with the answer, sentences, where the answers are found, are also recorded. For example,

Corrected Telugu Paragraph: ఒక నిర్దిష్ట రకం ఆపిల్ను పెంచుకోవాలనుకుంటే, కావాల్సిన రకం నుండి ఒక విత్తనాన్ని నాటడం ద్వారా సాధ్యం కాదు. విత్తనం ఏ ఆపిల్ నుండి వచ్చిందో ఆ ఆపిల్ యొక్క డీ.ఎన్.ఏ విత్తనంలో ఉంటుంది, కానీ దాంతో పాటు ఏ ఆపిల్ పువ్వు ద్వారా విత్తనాలు పరాగసంపర్కం చెందబడ్డాయో ఆ ఆపిల్ డీ.ఎన్.ఏ కూడా ఉంటుంది, అది వేరే రకానికి చెందినది కావొచ్చు (Oka nirdiṣṭa rakaṁ āpilnu peñcukōvālanukuṇṭē, kāvālsina rakaṁ nuṇḍi oka vittanānni nāṭaḍaṁ dvārā sādhyaṁ kādu. Vittanaṁ ē āpil nuṇḍi vaccindō ā āpil yokka ḍī.En.Ē vittananlō uṇṭundi, kānī dāntō pāṭu ē āpil puvvu dvārā vittanālu parāgasamparkaṁ cendabaḍḍāyō ā āpil ḍī.En.Ē kūḍā uṇṭundi, adi vērē rakāniki cendinadi kāvoccu)

Corrected Telugu Query: మనం ఏ ఫలం గురించి మాట్లాడుతున్నాం? (Manaṁ ē phalaṁ guriñci māṭlāḍutunnāṁ?)

English Answer: Apple

Incorrect Telugu Answers:

- Partial Answers
  - Subset of answer is correct: ఆపిల్ పువ్వు(Āpil puvvu)
  - Incomplete answer: ఆపి (Āpi)
- Completely Incorrect Answer: డీ.ఎన్.ఏ (Ḍī.En.Ē)
- Correct but out-of-the-context Answer: ఆపిల్ పండు (Āpil paṇḍu)

Extracted Correct Telugu Answer: ఆపిల్ ను (Āpil nu)

Modified Telugu Answer: ఆపిల్ (Āpil)

3.2.4.4 General Corrections

In addition to rectifying paragraphs, questions, and answers, certain rare instances require generic correction, as outlined below. After addressing these rare instances individually as per Para, Question, and Answer, we applied these rules to further enhance the overall quality of the data.

1. **Transliterate Acronyms**

   Transliterating an acronym ensures that its meaning and pronunciation remain consistent across different languages or writing systems, facilitating effective communication and comprehension.

   English Word: CBI

   Translated Telugu Word: CBI

   Corrected Telugu Word: సి.బి.ఐ (Si.Bi.Ai)

2. **Transliterate Proper Nouns**

   Converting proper nouns, such as names of people, places, or organizations, from one language or script to another while maintaining their pronunciation as closely as possible. This process is essential for ensuring accuracy and consistency when referring to proper nouns in different languages or scripts. Transliterating proper nouns enables effective communication across linguistic and cultural boundaries, allowing individuals to recognize and understand familiar names even when written in a different script.

   English Word: Rocky Mountains

   Translated Telugu Word: రాతి పర్వతాలు (Rāti parvatālu)

   Corrected Telugu Word: రాకీ పర్వతాలు (Rākī parvatālu)

   English Sentence: Good Friday is one of the most celebrated festivals.

   Translated Telugu Sentence: మంచి శుక్రవారం అత్యంత ప్రసిద్ధ పండుగలలో ఒకటి. (Mañci śukravāraṁ atyanta prasid'dha paṇḍugalalō okaṭi.)

   Corrected Telugu Sentence: గుడ్ ఫెర్డే అత్యంత ప్రసిద్ధ పండుగలలో ఒకటి. (Guḍ phraiḍē atyanta prasid'dha paṇḍugalalō okaṭi.)

3. **Contexually relatable words**

   When translating a sentence, it is essential to maintain the contextual meaning rather than translating word-for-word. In below example, both the translated and corrected sentences are correct. But try to choose more appropriate words in the correction.

   English Sentence: you spoiled my food by tasting it.

Translated Telugu Sentence: నువ్వు నా ఆహారాన్ని రుచి చూడటం ద్వారా పాడుచేసావు. (Nuvvu nā āhārānni ruci cūḍaṭa dvārā pāḍucēsāvu.)

Corrected Telugu Sentence: నువ్వు నా ఆహారాన్ని ఎంగిలి చేసి పాడుచేసావు. (Nuvvu nā āhārānni eṅgili cēsi pāḍucēsāvu.)

4. **Linguistic nature**

When translating from English to Telugu, understanding the linguistic nature of both languages is crucial. Understanding the linguistic characteristics of both languages helps translators choose appropriate words and phrases that resonate with the target audience and convey the message in a culturally appropriate manner. Acknowledging the linguistic nature of both languages is essential for producing high-quality translations that are faithful to the original text while resonating with the target audience.

English Sentence: Queen Elizabeth celebrates two birthdays.

Translated Telugu Sentence: క్వీన్ ఎలిజబెత్ రెండు పుట్టినరోజులను జరుపుకుంటాడు. (Kvīn elijabet reṇḍu puṭṭinarōjulanu jarupukuṇṭāḍu)

Corrected Telugu Sentence: క్వీన్ ఎలిజబెత్ రెండు పుట్టినరోజులను జరుపుకుంటుంది. (Kvīn elijabet reṇḍu puṭṭinarōjulanu jarupukuṇṭundi)

English Sentence: Tesla finally reached India.

Translated Telugu Sentence: టెస్లా ఎట్టకేలకు భారత్ కు చేరుకున్నాడు. (Ṭeslā eṭṭakēlaku bhārat ku cērukunnāḍu.)

Corrected Telugu Sentence: టెస్లా ఎట్టకేలకు భారత్ కు చేరుకుంది. (Ṭeslā eṭṭakēlaku bhārat ku cērukundi.)

Chapter 4

Experiments and Results

4.1 Experiments

We experimented with TeQuAD in monolingual and cross-lingual setups. The pre-trained Multilingual-BERT (mBERT) trained in 104 languages, including Telugu and English, is employed for obtaining encoded representations for both languages. We use the NLTK tokenizer followed by the BERT Word Piece tokenizer to sub-tokenize the tokens in all the experiments. Experimented with a batch size of 64 and sequence length of 512. As in Google's TensorFlow implementation of BERT, ADAM with weight decay optimizer is considered with different learning rates for different experimental setups. Our models have been trained on Google Cloud TPU v2.

4.1.1 Monolingual setup

In the monolingual setup, 82k Telugu triples from TeQuAD are considered for fine-tuning the mBERT model for the MRC task. We used Google's Tensorflow implementation of BERT for running SQuAD tasks and trained it for 3 epochs with a learning rate of 1e-4.

4.1.2 Cross-lingual setup

The dual BERT approach proposed in [15] is used for the CLMRC setup. In this approach, deep contextualized representations of the inputs from both languages are considered, and 'Bilingual Context' is computed, which will be used to exploit the semantic relations among the English and Telugu QA pairs. Parallel QA pairs of English and Telugu are passed as inputs to the model, and span indices of the Telugu answer phrases are predicted. 82k Parallel Telugu-English triples from TeQuAD are considered for fine-tuning the pre-trained mBERT model. We used the implementation in and trained it for 3 epochs with a learning rate of 2e-5. Both the Cross-Lingual and Monolingual fine-tuned models are evaluated on three test datasets. Along with Translated & Corrected (2000) and Wiki (947) test datasets, Telugu samples of the Gold Passage task (Span Extractive QA task) from the TyDiQA dev

Model	Test Dataset	Monolingual		Crosslingual	
		F1	EM	F1	EM
mBERT(Zero shot)	Translated & Corrected	28.4	0.0	27.1	0.01
	Wiki QA	27.1	0.0	27.6	0.0
	TyDi dev QA	21.0	0.0	21.3	0.0
mBERT(TeQuAD)	Translated & Corrected	69.4	43.7	69.4	43.5
	Wiki QA	**83.0**	**61.0**	**83.3**	**61.9**
	TyDi dev QA	61.0	41.6	69.1	43.3

Table 4.1: Experimental results of MRC on Test Datasets. Performance (in terms of %) F1 : F1 Score and EM: Exact Match Score

(667) dataset are considered for evaluation. Results of the evaluation for monolingual and cross-lingual setups are shown in Table 4.1.

4.2 Results and Observations

From the results attained, the key observation is that compared to the zero-shot mBERT model, the models finetuned on TeQuAD performed better for the Telugu MRC task. On average, a 40% increase in F1 and EM Scores was registered across all setups.

From the experiments, we observed that the performance of the finetuned model on the Wiki test dataset is much better than on other test datasets. Notice that the Wiki dataset is created from original Telugu Wikipedia articles, followed by the manual effort to produce QA pairs, hence has higher quality than the Translated & Corrected dataset. On the other hand, the TiDyQA Telugu samples are of low quality and not preferable for evaluating Telugu MRC. Most of the QA pairs in TyDiQA revolve around numbers such as zip codes, dates of birth/death, area of land, etc. MRC model exposed with such data resources will overfit to learn and answer just these types of questions and lack the ability to comprehend other QA types. So, the model trained on the TyDiQA train dataset produced good results when evaluated on the TyDiQA dev dataset. However, compared to the TeQuAD model, the performance of the TyDiQA-trained model fell behind when evaluated on Translated & Corrected and Wiki test datasets.

4.2.1 Why low EM scores?

In the MRC experiments, though TeQuAD registered decent F1 scores, Exact Match scores are observed to be noticeably lower than F1 scores. Approximately 20% gap can be seen between these two metrics across all the setups. Partial answer predictions will affect the Exact Match score, and we tried to analyze the causes for such error predictions. The primary reason for the low EM scores is the multiple possible answers for a query. The existence of different answer

Experimental Setup	**Translated & Corrected**		**Wiki QA**	
	F1	EM	F1	EM
Mono-lingual	65.9	39.1	79.3	50.5
Cross-lingual	67.4	39.7	82.2	54.0

Table 4.2: Results of the Experimental setups trained on less corpora : 34k QA pairs. Performance (in terms of %) F1: F1 Score and EM: Exact Match Score

phrases in the paragraph, all of which seem to be correct, will affect the ability of the MRC model to predict the exact answer.

Another obvious reason for faulty answer predictions is the low-to-moderate resources available for the language. Pre-trained models exposed to such few data resources might not be able to reason the context leading to false answer predictions. And even though such models leverage the information from high resource language(s), due to the linguistic divergences between the languages (here Telugu and English), answer boundary detection capability in the low resource language is poorer, failing to identify the complete answer phrase in the context.

In [49], they discussed the deficient answer boundary detection capability of MRC models for low-resource languages. Their work suggested improving the detection capability by training the MRC model on phrases in low-resource language mined from the internet. We experimented by mining approximately 32k Telugu phrases from Wikipedia and trained the model with the phrase masking prediction task. Results don't show any noticeable improvement in the EM scores.

On the other hand, several MRC works employ character-level span indices to point to the answer phrase specifically. This might lead to worse EM scores in Telugu, considering the rich morphology of the language. So, instead, we stuck to word-level span indices for the answer phrases.

4.2.2 Cross Lingual Experimentation

As discussed, [16] proposed the Dual BERT approach to improve the MRC for low-resource languages by utilizing cross-lingual knowledge. With experiments, we observed that CLMRC setup helps in boosting the performance of the model when the size of the corpora is low (See 4.2). However, with the creation of large synthetic data, the effect of the CLMRC setup is negligible. In table 4.1, results obtained by training the model on 82k data in the mono-lingual setup are identical to the results of the CLMRC setup. The creation of such resources helps the machine to learn from the target language itself instead of relying on high-resource languages.

Test Dataset	**TyDiQA**		**TeQuAD**	
	F1	EM	F1	EM
Translated-&-Corrected	57.7	29.5	69.4	43.7
Wiki QA	77.3	48.4	83.0	61.0

Table 4.3: Comparison b/w TeQuAD and TyDi QA for Telugu MRC. Performance (in terms of %) F1: F1 Score and EM: Exact Match Score

4.2.3 Comparison with TyDiQA

Clark et al. [14] demonstrated the performance of the Gold-Passage MRC task in Telugu, which is similar to the SQuAD style Question Answering task. They trained the model on approximately 49k multilingual QA pairs and evaluated it on the Telugu test dataset. For comparison, we also considered 49k multilingual QA pairs from the TyDiQA dataset and finetuned the mBERT model for the MRC task. Then we evaluated this model on the discussed Telugu test datasets - The wiki test dataset and the Translated & Corrected test dataset. See 4.3 for a comparison of Telugu MRC performance between TyDiQA and TeQuAD models. The MRC model finetuned on TeQuAD was observed to be outperforming the TiDyQA model in the Telugu MRC task.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we have addressed the challenges involved in creating resources and systems for low-resource languages such as Telugu. The main contributions of the thesis are focused on resource creation and system development for question answering in the Telugu language.

In the realm of QA, numerous endeavors concentrate on utilizing English resources to establish materials for low-resource languages through translation. However, a primary challenge associated with this approach lies in ensuring the precision of the translated answers and their respective span positions. To address this challenge, we introduced several techniques for correcting and extracting answer spans, aiming to enhance both the quality and quantity of translated QA datasets, particularly those following the SQuAD format. Our methodology involved translating English triples from SQuAD into Telugu and subsequently applying these span correction techniques. Consequently, we constructed the Telugu Question Answering Dataset - TeQuAD, comprising 82,000 parallel Telugu-English triples. Subsequently, we assessed the performance of the Telugu MRC model across various Telugu QA resources and presented our findings.

5.2 Future Work

In the future, we aim to improve the MRC task for the Telugu language by offering a set of pre-trained models that are trained on publicly available resources in Telugu, as well as creating additional QA data resources for the Telugu language. Future work should focus on creating standardized datasets and evaluation metrics specific to Telugu closed domain QA. Further research can explore the integration of multi-lingual pre-trained models with domain-specific fine-tuning to leverage knowledge from multiple languages and domains.

Related Publications

- **Rakesh Kumar Vemula**, Mani Kanta Sai Nuthi and Manish Shrivastava "**TeQuAD: Telugu Question Answering Dataset**". Proceedings of the nineteenth International Conference on Natural Language Processing (ICON-2022)

Bibliography

[1] N. Abadani, J. Mozafari, A. Fatemi, M. A. Nematbakhsh, and A. Kazemi. Parsquad: Machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168. IEEE, 2021.

[2] E. Adebisi, B. Adefowoke Ojokoh, and F. Olubusola Isinkaye. An open domain factoid qa framework with improved validation techniques. *International Journal of Information Science and Management (IJISM)*, 20(1), 2022.

[3] S. Archana, N. Vahab, R. Thankappan, and C. Raseek. A rule based question answering system in malayalam corpus using vibhakthi and pos tag analysis. *Procedia Technology*, 24:1534–1541, 2016.

[4] S. Arora, A. May, J. Zhang, and C. Ré. Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*, 2020.

[5] M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

[6] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay. Bfqa: A bengali factoid question answering system. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17*, pages 217–224. Springer, 2014.

[7] M. R. Bhuiyan, A. K. M. Masum, M. Abdullahil-Oaphy, S. A. Hossain, and S. Abujar. An approach for bengali automatic question answering system using attention mechanism. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2020.

[8] V. Bolotova, V. Blinov, F. Scholer, W. B. Croft, and M. Sanderson. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207, 2022.

[9] M. Bornea, L. Pan, S. Rosenthal, R. Florian, and A. Sil. Multilingual transfer learning for qa using translation as data augmentation. *arXiv preprint arXiv:2012.05958*, 2020.

[10] E. Cabrio, J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. Gandon. Qakis: an open domain qa system based on relational patterns. In *International Semantic Web Conference, ISWC 2012*, 2012.

[11] C. P. Carrino, M. R. Costa-jussà, and J. A. Fonollosa. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*, 2019.

[12] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*, 2016.

[13] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

[14] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470, 2020.

[15] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, and G. Hu. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*, 2018.

[16] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu. Cross-lingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*, 2019.

[17] A. Das, J. Mandal, Z. Danial, A. Pal, and D. Saha. A novel approach for automatic bengali question answering system using semantic similarity analysis. *International Journal of Speech Technology*, 23:873–884, 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] M. d'Hoffschmidt, W. Belblidia, T. Brendlé, Q. Heinrich, and M. Vidal. Fquad: French question answering dataset. *arXiv preprint arXiv:2002.06071*, 2020.

[20] P. Efimov, A. Chertok, L. Boytsov, and P. Braslavski. Sberquad–russian reading comprehension dataset: Description and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer, 2020.

[21] D. Gupta, S. Kumari, A. Ekbal, and P. Bhattacharyya. Mmqa: A multi-domain multilingual question-answering framework for english and hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[22] P. Gupta and V. Gupta. Hybrid approach for punjabi question answering system. In *Advances in Signal Processing and Intelligent Recognition Systems*, pages 133–149. Springer, 2014.

[23] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021.

[24] T.-Y. Hsu, C.-L. Liu, and H.-y. Lee. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv:1909.09587*, 2019.

[25] D. Karani. Introduction to word embedding and word2vec. *Towards Data Science*, 1, 2018.

[26] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[27] S. Lim, M. Kim, and J. Lee. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*, 2019.

[28] J. Liu, Y. Lin, Z. Liu, and M. Sun. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, 2019.

[29] J. Liu, L. Shou, J. Pei, M. Gong, M. Yang, and D. Jiang. Cross-lingual machine reading comprehension with language branch knowledge distillation. *arXiv preprint arXiv:2010.14271*, 2020.

[30] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

[31] D. Meyer. How exactly does word2vec work. *Uoregon. Edu, Brocade. Com*, pages 1–18, 2016.

[32] G. Nanda, M. Dua, and K. Singla. A hindi question answering system using machine learning approach. In *2016 international conference on computational techniques in information and communication technologies (ICCTICT)*, pages 311–314. IEEE, 2016.

[33] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

[34] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[35] G. Puccetti, A. Miaschi, and F. Dell'Orletta. How do bert embeddings organize linguistic knowledge? In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 48–57, 2021.

[36] F. Rahutomo, T. Kitasuka, M. Aritsugi, et al. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1. University of Seoul South Korea, 2012.

[37] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[38] R. G. Reddy, M. A. Sultan, E. S. Kayi, R. Zhang, V. Castelli, and A. Sil. Answer span correction in machine reading comprehension. *arXiv preprint arXiv:2011.03435*, 2020.

[39] X. Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[40] S. Sarker, S. T. A. Monisha, and M. M. H. Nahid. Bengali question answering system for factoid questions: A statistical approach. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE, 2019.

[41] I. Seena, G. Sini, and R. Binu. Malayalam question answering system. *Procedia Technology*, 24:1388–1392, 2016.

[42] M. E. Sucunuta and G. E. Riofrio. Architecture of a question-answering system for a specific repository of documents. In *2010 2nd International Conference on Software Technology and Engineering*, volume 2, pages V2–12. IEEE, 2010.

[43] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019.

[44] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] G. Wiedemann, S. Remus, A. Chawla, and C. Biemann. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

[47] W. Yu, Z. Jiang, Y. Dong, and J. Feng. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*, 2020.

[48] C. Yuan and C. Wang. Parsing model for answer extraction in chinese question answering system. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 238–243. IEEE, 2005.

[49] F. Yuan, L. Shou, X. Bai, M. Gong, Y. Liang, N. Duan, Y. Fu, and D. Jiang. Enhancing answer boundary detection for multilingual machine reading comprehension. *arXiv preprint arXiv:2004.14069*, 2020.