TeQuAD: Telugu Question Answering Dataset

Rakesh Kumar Vemula, Manikanta Sai Nuthi and **Manish Shrivastava**
Language Technologies Research Center
International Institute of Information Technology
Hyderabad, India
{rakesh.kumar,mani.kanta}@research.iiit.ac.in and
m.shrivastava@iiit.ac.in

Abstract

Recent state of the art models and new datasets have advanced many Natural Language Processing areas, especially, Machine Reading Comprehension 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. In this paper, 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, the best of them resulting in an F1 score of 83 % and Exact Match (EM) score of 61 %.

1 Introduction

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. A popular large-scale annotated reading comprehension dataset - SQuAD Rajpurkar et al. (2016), revolutionised the research interest in this area for English. And though decent research work has been done in MRC for a few Indian languages, for languages like Telugu, which is a Dravidian language, still need similar resources for such 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 few matching techniques to extract the span indices of answers in the target language (Carrino et al. (2019), Abadani et al. (2021), Artetxe et al. (2019)). For others, the dataset is created by using the methodology followed in the creation of SQuAD (Lim et al. (2019), Efimov et al. (2020), Cui et al. (2018), d'Hoffschmidt et al. (2020)).

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, a 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 (Clark et al., 2020) and on two other test datasets, which we created manually by correcting a few samples from translated SQuAD and by using Wikipedia articles respectively.

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 300 - 307
December 15-18, 2022 ©2022 Association for Computational Linguistics

Our dataset and code are available here[1].

2 Related Work

Several datasets such as SQuAD(Rajpurkar et al., 2016), NewsQA dataset (Trischler et al., 2016) and CNN/Dailymail (Chen et al., 2016), etc fulfills 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. (2019), Efimov et al. (2020), Cui et al. (2018), d'Hoffschmidt et al. (2020) developed RC datasets natively. Clark et al. (2020) presented a question answering dataset covering 11 typologically diverse languages including Telugu.

Few others propose methods to boost the functioning of the model in low-resource settings. Hsu et al. (2019) 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. (2020) presents translation-based data augmentation mechanism to improve multilingual transfer learning.

Liu et al. (2020) and Cui et al. (2019) talks about leveraging translated information from the high resource languages to perform well in low resource languages. Cui et al. (2019) 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 has the ability to learn semantic information from bilingual QA pairs and utilize the learned knowledge to improve MRC in low resource languages.

Yuan et al. (2020) 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. (2020). They added additional layers on top of a pre-trained transformer-based language model to re-examine and modify the predicted answers.

3 Corpus 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 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[2], obtaining translated triples consisting of translated Telugu paragraphs, questions, and answers.

After translation, a well-known issue is difficulty in extraction of the span indices of the translated answers. Considering the different possibilities of 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.

3.1 Matching

We used matching algorithms like cosine similarity 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, 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 English context and selected the corresponding index as the answer from repetitions of the Telugu answer in 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

[1] https://github.com/rakeshvemula1157/TeQuAD

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

	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. Vencuvān 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 1: Representation of QA pairs in parallel corpora

second repetition of the Telugu word (Āpil) as the answer in the translated Telugu context.

3.2 Explicit Position Indicator

We managed to extract the answer span for few ignored samples by marking the English answer phrase before translating to Telugu. English answer phrase in the English context is marked by the special symbol ('|') and then translated to the Telugu language. The marked symbol ('|') remains unchanged, making it easier to find the translated Telugu answer phrase in the translated Telugu context.

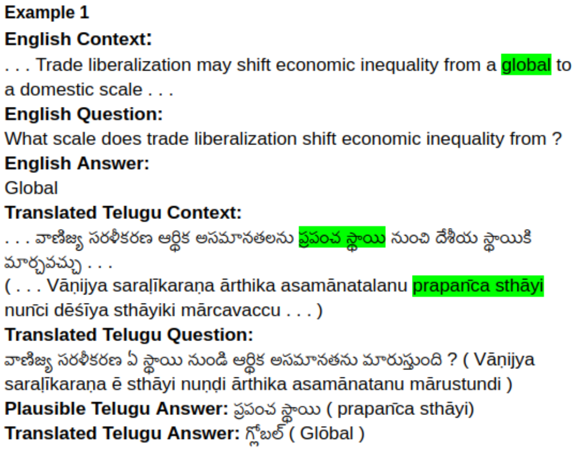
Using these approaches, we were able to obtain 82,605 English-Telugu parallel triples - creating a reading comprehension dataset, TeQuAD. Table 1 shows the representation of parallel corpora of English - Telugu. For evaluation, we have created two different test datasets to analyze the performance of models on Telugu MRC.

- **Translated & Corrected dataset:** 1000 English triples from the dev set of SQuAD1.1 are translated to Telugu and corrected manually. A set of guidelines is prepared and explained in 3.4 to correct the translated Telugu context, questions, and answers.
- **Wiki dataset:** Similar to SQuAD, we created this data from Wikipedia articles. Randomly selected wikipedia articles are splitted 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. Minimum 5 and maximum 10 questions were created for each paragraph/context.

3.3 Span Extractor

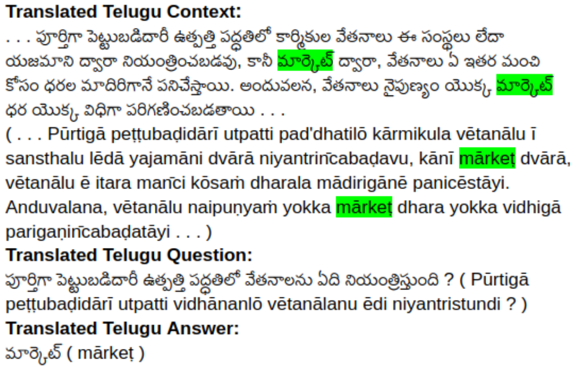
In TeQuAD, we obtained the span indices for the Telugu answers by using the above-mentioned matching techniques. Such rule-based techniques might not provide better results in cases where,

1. The translated Answer might not be present in the translated Context (information about the answer might have lost or different form of the answer generated in translation). See Figure 1 for example.
2. Multiple instances of the translated Answer might be in the translated Context. See Figure 2 for example.
3. A partial answer phrase returns a better matching score with the translated answer phrase than the actual answer phrase. See Figure 3 for example.



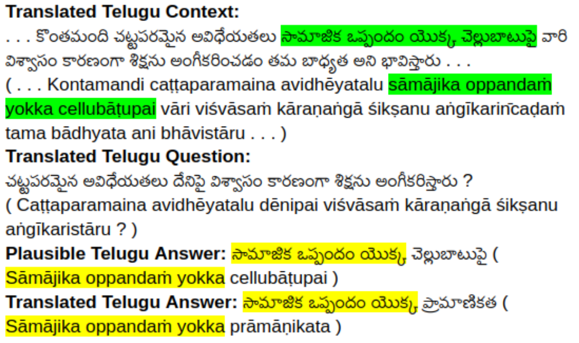
Example 1
English Context:
. . . Trade liberalization may shift economic inequality from a global to a domestic scale . . .
English Question:
What scale does trade liberalization shift economic inequality from ?
English Answer:
Global
Translated Telugu Context:
. . . వాణిజ్య సరళీకరణ ఆర్థిక అసమానతలను 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 1: Example for absence of translated Answer in the translated Context. Both *’prapanca 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 mārkeṭ 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)

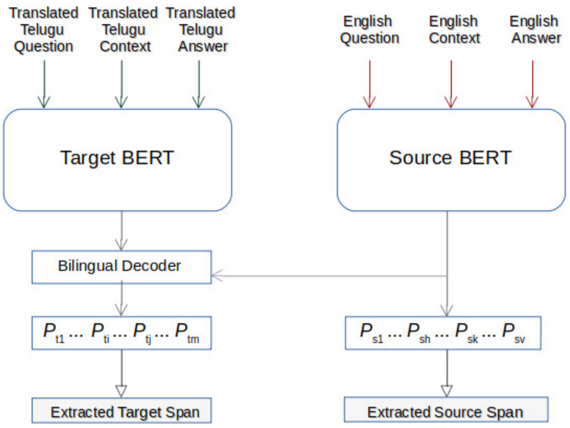
Figure 2: Example for multiple instances of Answer in the Context



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: Example for partial matching answer scenario.

In order 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. (2019), 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 4). Due to its ability to exploit semantic information from both Telugu-English parallel triples, it can identify a modified variant of answer phrase in the translated context, even if the translated answer phrase does not present in the translated context completely.



Translated Telugu Question
Translated Telugu Context
Translated Telugu Answer
English Question
English Context
English Answer
Target BERT
Source BERT
Bilingual Decoder
$P_{t1} \cdots P_{ti} \cdots P_{tj} \cdots P_{tm}$
$P_{s1} \cdots P_{sh} \cdots P_{sk} \cdots P_{sv}$
Extracted Target Span
Extracted Source Span

Figure 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 needs sufficient resources to perform well and predicts a span even if the answer information is not present in the context (e.g. might have been lost in machine translation).

3.4 Manual QA Correction

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.)

3.4.1 Correction Guidelines

Data in SQuAD is in triples form. All three components of triples (para, question, and answer) are translated from English to Telugu. The order followed to correct a translated triple is:

- 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:

- **Adequacy:** The meaning/knowledge provided in the English para/question should be preserved in the Telugu para/questions.
- **Fluency:** The structure/syntax of the Telugu para/question should be proper/readable.

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. By using these guidelines, we corrected 1000 samples which are used as test set.

4 MRC Experiments

We experimented on 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 nltk tokenizer followed by 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.

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

Cross-lingual setup: The dual BERT approach proposed in Cui et al. (2018) 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 Cui et al. (2019) and trained it for 3 epochs with the learning rate of 2e-5.

Both the Cross-Lingual and Monolingual fine-tuned models are evaluated on three test datasets. Along with Translated & Corrected (1000) and Wiki (947) test datasets, Telugu samples of Gold Passage task (Span Extractive QA task) from TyDiQA dev (667) dataset are considered for evaluation.

F1 score and EM score are used as evaluation metrics. Results of the evaluation for monolingual and cross-lingual setups are shown in Table 2.

5 Results and Observations

The most important observation from the results is that the models fine tuned on TeQuAD performed way better than the zero-shot mBERT model. On average, a 40% increase in F1 and EM scores were obtained in all setups.

Our experiments also show that performance on the Wiki test dataset is better than others. It must be noted that Wiki test dataset is well annotated, created from Telugu Wikipedia articles, and has better quality than the Translated & Corrected test dataset. In contrast, the quality of the TyDiQA Telugu samples is low and not recommended for fair Telugu MRC evaluation. Most of the queries in TyDiQA revolve around the area of lands, zip codes, date of births/deaths *etc.* Learning on a set of similar types of questions more often will make the model over-

Model	Test Dataset	Mono-Lingual		Cross-Lingual	
		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 2: Experimental results of MRC on Test Datasets. Performance (in terms of %) F1 : F1 Score and EM: Exact Match Score

Test Dataset	**TyDi QA**		**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 3: Comparison b/w TeQuAD and TyDi QA for Telugu MRC. Performance (in terms of %) F1: F1 Score and EM: Exact Match Score

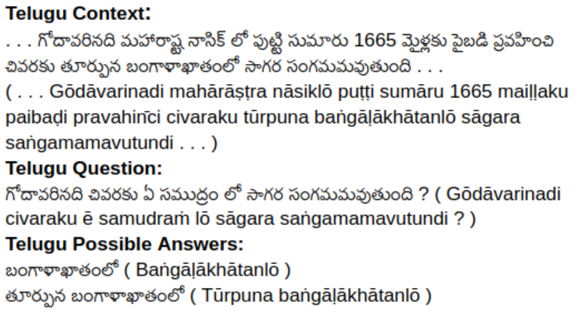
fit on these types of questions, but would lack the ability to comprehend other types. As expected, the model trained on TyDiQA train dataset achieved good performance when evaluated on TyDiQA dev dataset, but the performance fell behind compared to TeQuAD model when evaluated on Translated-&-Corrected and Wiki test datasets.

Why low EM scores ?

Although decent F1 scores have been registered, the gap between the two metrics is notable. The difference between the metrics is approximately 20% across all the setups. We tried to analyze the reasons for error predictions. One reason for the low Exact Match score is multiple possible answers for a query. Different answers, all that seems to be correct might affect the MRC model generating the exact answer. See Figure 5 for Example.

Another obvious reason for faulty answer predictions is the low-to-moderate resources available for the language. Pre-trained models exposed to such fewer 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 Yuan et al. (2020), 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.



Telugu Context:
. . . గోదావరినది మహారాష్ట్ర నాసిక్ లో పుట్టి సుమారు 1665 మైళ్లకు పైబడి ప్రవహించి చివరకు తూర్పున బంగాళాఖాతంలో సాగర సంగమమవుతుంది . . .
(. . . Gōdāvarinadi mahārāṣṭra nāsiklō puṭṭi sumāru 1665 maiḷḷaku paibaḍi pravahiñci civaraku tūrpuna baṅgāḷākhātanlō sāgara saṅgamamavutundi . . .)
Telugu Question:
గోదావరినది చివరకు ఏ సముద్రం లో సాగర సంగమమవుతుంది ? (Gōdāvarinadi civaraku ē samudraṁ lō sāgara saṅgamamavutundi ?)
Telugu Possible Answers:
బంగాళాఖాతంలో (Baṅgāḷākhātanlō)
తూర్పున బంగాళాఖాతంలో (Tūrpuna baṅgāḷākhātanlō)

Figure 5: Example for multiple possible answers

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

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

Why Cross-lingual Experimentation?

As Discussed, Cui et al. (2019) proposed 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). But with the creation of large synthetic data, the effect of CLMRC setup is negligible. In table 2, results obtained by training the model on 82k data in mono-lingual setup are identical to the results of CLMRC setup. Creation of such resources helps the machine to learn from the target language itself instead of relying on High resource languages.

Comparison with TyDiQA

Clark et al. (2020) presents the performance of Gold-Passage MRC (Similar to SQuAD style QA) in Telugu. They train the model on approximately 49k Multilingual QA pairs and evaluate it on the Telugu test dataset. We also experimented by fine-tuning mBERT on TyDIQA 49k QA pairs and evaluated it on the above-mentioned test datasets. See table 3 for a comparison between models trained on TyDiQA and TeQuAD. The TeQuAD based model outperformed the TyDiQA trained model in Telugu MRC.

6 Conclusion and Future work

As a move towards the creation of the QA dataset for Indian languages, this work took a step forward in the MRC corpus creation using translation. To record decent performances in NLP tasks for low resource languages, sufficient resources are necessary. As we discussed, the creation of such resources is difficult. Resources from high resource languages like English might be considered for the creation of datasets for low-resource languages like Telugu to create abundant data for different NLP tasks.

We introduce creation/correction techniques for such datasets improving the quality along with the quantity of the datasets along with providing mechanisms for further augmenting data. In the future, we would like to improve the MRC task for Telugu, provide a collection of pre-trained models trained on openly available resources in the Telugu language, as well as create additional data resources for the Telugu language.

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. 2021. Parsquad: Machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168. IEEE.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2020. Multilingual transfer learning for qa using translation as data augmentation. *arXiv preprint arXiv:2012.05958*.

Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. Fquad: French question answering dataset. *arXiv preprint arXiv:2002.06071*.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv:1909.09587*.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. 2020. Cross-lingual machine reading comprehension with language branch knowledge distillation. *arXiv preprint arXiv:2010.14271*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Revanth Gangi Reddy, Md Arafat Sultan, Efsun Sarioglu Kayi, Rong Zhang, Vittorio Castelli, and Avirup Sil. 2020. Answer span correction in machine reading comprehension. *arXiv preprint arXiv:2011.03435*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. *arXiv preprint arXiv:2004.14069*.