Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya

Fitsum Gaim **Wonsuk Yang** **Hancheol Park** **Jong C. Park**[*]
School of Computing
Korea Advanced Institute of Science and Technology
`{fitsum.gaim,dirrick0511,hancheol.park,jongpark}@kaist.ac.kr`

Abstract

Question-Answering (QA) has seen significant advances recently, achieving near human-level performance over some benchmarks. However, these advances focus on high-resourced languages such as English, while the task remains unexplored for most other languages, mainly due to the lack of annotated datasets. This work presents a native QA dataset for an East African language, Tigrinya. The dataset contains 10.6K question-answer pairs spanning 572 paragraphs extracted from 290 news articles on various topics. The dataset construction method is discussed, which is applicable to constructing similar resources for related languages. We present comprehensive experiments and analyses of several resource-efficient approaches to QA, including monolingual, cross-lingual, and multilingual setups, along with comparisons against machine-translated silver data. Our strong baseline models reach 76% in the F1 score, while the estimated human performance is 92%, indicating that the benchmark presents a good challenge for future work. We make the dataset, models, and leaderboard publicly available.[1]

1 Introduction

Question Answering (QA) and Machine Reading Comprehension (MRC) have seen significant advances in recent years, achieving human-level performance on large-scale benchmarks (Rajpurkar et al., 2016, 2018). The main factors driving the progress are the adaption of large pre-trained large language models and the proliferation of QA datasets (Rogers et al., 2022). However, most studies focus on high-resourced languages, while the task remains unexplored for most of the World's diverse languages. The primary challenge for non-English QA is the lack of native annotated datasets. In particular, there is little to no study done on scarcely resourced languages such as Tigrinya that are markedly different from English in terms of linguistic properties including syntax, morphology, and typography.

Article: ቀይሕ ባሕሪ [The Red Sea]

Paragraph:
ቀይሕ ባሕሪ ሓደ ካብቶም ብኣርባ‘ዕተ ሕብሪ ዝጽው‘ዑ ባሕርታት ‘ዓለምና ማለት ብጫ ባሕሪ፣ ጸሊም ባሕሪ፣ ቀይሕ ባሕሪ ከምኡ‘ውን ጻ‘ዕዳ ባሕሪ እዩ ። መበቆል ስሙ ድማ ካብቶም ኣብ ግዜ ክረምቲ ብብዝሒ ዝራብሑ ብፍላይ ድማ ትራይኮድዝሚም ኤሪትሪን ዝተባህሉ መርዛማት ዝኾኑ ተሃዋስያን ብዝህብዎ ቀይሕ ሕብሪ ምዃኑ ይንገረሉ ።

[The Red Sea is one of the four seas in the world that are named after common colors: the Yellow Sea, the Black Sea, the Red Sea and the White Sea. The origin of its name is attributed to the red color given by the poisonous bacteria, especially the Trichodesmium Erythraeum, which breed in large numbers during the Summer season.]

Question 1: ኣብ ‘ዓለም ክንደይ ብሕብሪ ዝጽው‘ዑ ባሕርታት ኣለዉ? [How many seas in the World are named after common colors?]

Answer: ኣርባ‘ዕተ [four]

Question 2: ኣብ ‘ዓለም ብሕብሪ ዝጽው‘ዑ ባሕርታት ጡዕቐስ? [List all the seas in the World that are named after common colors?]

Answer: ብጫ ባሕሪ፣ ጸሊም ባሕሪ፣ ቀይሕ ባሕሪ ከምኡ‘ውን ጻ‘ዕዳ ባሕሪ [Yellow Sea, the Black Sea, the Red Sea and the White Sea]

Question 3: ትራይኮድዝሚም ኤሪትሪን ኣበየናይ ወቕቲ ይራብሑ? [In which season do the Trichodesmium Erythraeum reproduce?]

Answer: ክረምቲ [Summer]

Question 4: ትራይኮድዝሚም ኤሪትሪን እንታይ ሕብሪ ይህቡ? [What is the color of Trichodesmium Erythraeum?]

Answer: ቀይሕ [red]

Figure 1: Example entry from TiQuAD: A paragraph as context and the corresponding annotated question-answer pairs. Some context was redacted for brevity.

This work presents TiQuAD, the first publicly available **Qu**estion-**A**nswering **D**ataset for **Ti**grinya; see Figure 1 for an example entry. We collaborate with native Tigrinya speakers to collect documents and annotate the dataset, yielding a total of 10.6K question-answer pairs with 6.5K unique

[*] Coressponding author
[1] TiQuAD: `https://github.com/fgaim/tiquad`

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11857–11870
July 9-14, 2023 ©2023 Association for Computational Linguistics

questions over 572 paragraphs gathered from 290 news articles.

We assess the quality of annotations and explore strong baselines by fine-tuning TiRoBERTa and TiELECTRA (Gaim et al., 2021) as monolingual models of Tigrinya and XLM-R (Conneau et al., 2020) and AfriBERTa (Ogueji et al., 2021) as representative multilingual models. In addition to the monolingual QA setup, we perform three scenarios of cross-lingual and multilingual experiments. First, we translate SQuAD1.1 to Tigrinya and evaluate the performance in conjunction with the native TiQuAD. Second, we assess a zero-shot cross-lingual transfer learning approach (Artetxe et al., 2020; Lewis et al., 2020) by evaluating on the new dataset. Third, we explore the performance of a multilingual setup by jointly fine-tuning the models on English and Tigrinya datasets. The experimental settings are depicted in Figure 2. The best-performing baseline model achieves up to 76% in F1 score in the multilingual setup, while the estimated human performance is 92%. Considering the challenges of constructing annotated datasets for under-represented languages, we believe this work could serve as a reference case for similar languages. In particular, the TiQuAD benchmark is an important milestone in the advancement of question-answering for the Tigrinya language.

The contributions of this work are summarized as follows: (1) We build the first question-answering dataset for Tigrinya and make it publicly available. (2) We present an in-depth analysis of the challenges of question answering in Tigrinya based on the dataset. (3) We apply transformer-based language models to the question-answering task in Tigrinya and compare it with datasets of other languages. (4) We investigate various resource-efficient cross-lingual and multilingual approaches to QA and assess the utility of the native dataset.

2 Related Work

2.1 Tigrinya Language

Tigrinya (ISOv3: `tir`) is a Semitic language, part of the Afro-Asiatic family with over 10 million native speakers in the East African regions of Eritrea and Northern Ethiopia. Tigrinya is closely related to Amharic and Tigre languages that are also spoken in similar regions and share the same ancestor, the now extinct Ge'ez language. In recent years, there is a growing research body and interest in Tigrinya. Gasser (2011) developed `HornMorph`, a morphological analysis and generation framework for Tigrinya, Amharic, and Oromo by employing Finite State Transducers (FSTs). Later, Tedla and Yamamoto (2018) employed a manually constructed dataset to train a Long Short-Term Memory (LSTM) model for morphological segmentation in Tigrinya. Osman and Mikami (2012) proposed a rule-based stemmer for a Lucene based Tigrinya information retrieval. Tedla et al. (2016) presented a part-of-speech (POS) corpus for Tigrinya with over 72K annotated tokens across 4.6K sentences. A few studies explored statistical and neural machine translation, between English and Tigrinya, by exploiting morphological segmentation (Tedla and Yamamoto, 2016; Gaim, 2017; Tedla and Yamamoto, 2018) and data augmentation via back-translation (Öktem et al., 2020; Kidane et al., 2021). More recent studies applied pre-trained language models to various downstream tasks such as part-of-speech tagging, sentiment analysis, and named entity recognition (Tela et al., 2020; Gaim et al., 2021; Yohannes and Amagasa, 2022). Moreover, Gaim et al. (2022) presented a dataset and method for the automatic identification of five typologically related East African languages that include Tigrinya. However, despite the recent progress, Tigrinya still lacks basic computational resources for most downstream tasks with very limited availability of annotated datasets.

2.2 Question-Answering beyond English

Native reading comprehension datasets beyond the English language are relatively rare. Efforts have been made to build MRC datasets in Chinese, French, German, and Korean, among others, all of which are designed following the formulation of SQuAD. The SberQuAD dataset (Efimov et al., 2020) is a Russian native reading comprehension dataset made up of 50K samples. The CMRC 2018 (Cui et al., 2019) dataset is a Chinese reading comprehension dataset that gathers 20K question and answer pairs. The KorQuAD dataset (Lim et al., 2019) is a Korean native reading comprehension dataset containing 70K samples. On the end of low-resourced languages, Mozannar et al. (2019) developed ARCD for Arabic with 1.3K samples. Keren and Levy (2021) presented ParaShoot, a reading comprehension dataset for Hebrew with a size of 3.8K question-answer pairs. More recently, Kazemi et al. (2022) built PersianQuAD, a native MRC dataset for Persian with over 20K samples.

Cross-lingual Question Answering Language-specific datasets are costly and challenging to build, and one alternative is to develop cross-lingual models that can transfer to a target without requiring training data in that language (Lewis et al., 2020). It has been shown that unsupervised multilingual models generalize well in a zero-shot cross-lingual setting (Artetxe et al., 2020). For this reason, cross-lingual question answering has recently gained traction with the availability of a few benchmarks. Artetxe et al. (2020) built XQuAD by translating 1190 question-answer pairs from the SQuAD1.1 development set by professional translators into ten other languages.

Multilingual Question Answering The MLQA dataset (Lewis et al., 2020) consists of over 12K question and answer samples in English and 5000 samples in six other languages such as Arabic, German and Spanish. More recently, Clark et al. (2020) presented TyDiQA, a dataset particularly designed to address information-seeking and natural questions covering 11 typologically diverse languages with a total of 204K samples. Longpre et al. (2021) presented an open domain dataset comprising 10K question-answer pairs aligned across 26 typologically diverse languages, yielding a total of 260K samples. Hu et al. (2020) presented XTREME, a multi-task benchmark for nine prominent NLP tasks including question-answering across 40 languages. Ruder et al. (2021) further extended the benchmark to XTREME-R, covering ten tasks across 50 typologically diverse languages. Xue et al. (2021) proposed a large multilingual pre-trained model that handles 101 languages. Note that none of the aforementioned datasets and models include the Tigrinya language.

Translated QA datasets Another relatively inexpensive alternative to building a native annotated QA dataset is translating an existing English dataset to the target language. Carrino et al. (2020) explored this by proposing the Translate-Align-Retrieve (TAR) method to translate the English SQuAD1.1 dataset to Spanish. Then the resulting dataset was used to fine-tune a multilingual model achieving a performance of 68.1/48.3% F1/EM on MLQA (Lewis et al., 2020) and 77.6/61.8% F1/EM on XQuAD (Artetxe et al., 2020). Similar approaches were also adapted for the Japanese and French languages (Asai et al., 2018; Siblini et al., 2019), where a multilingual version of BERT (Devlin et al., 2019) is trained on the English SQuAD1.1 and evaluated on the small translated corpus, reaching promising scores of 76.7% in F1 and 61.8% in EM.

Split	Articles	#Parags	#Qs	#As
Train	205	408	4,452	4,454
Development	43	76	934	2,805
Test	42	96	1,122	3,378
Total	290	572	6,508	10,637

Table 1: Data Statistics: Articles, Paragraphs, Questions, and Answers. The dataset is partitioned by articles.

3 Dataset Annotation

TiQuAD is designed following the task formulation of SQuAD (Rajpurkar et al., 2016), where each entry in the dataset is a triple consisting of a paragraph, a question, and the corresponding answer. The answer is a contiguous span of text in the paragraph, a typical setup of extractive question-answering.

The dataset was constructed in four stages: First, a diverse set of articles are collected from which we extract paragraphs that will serve as contexts. Second, the initial question and answer pairs are annotated for all the extracted paragraphs. Third, additional answers are annotated for all the questions in the development and test sets. Fourth, we post-process the annotations for quality control and remove noisy examples. The final dataset contains over 10.6K question-answer pairs across 572 paragraphs. While the size is on the smaller end compared to the English datasets, it reflects a realistic amount of data that researchers of low-resourced languages can acquire with a limited annotation budget.[2] The dataset characteristics are presented in Table 1. In the following sections, we present the data collection and annotation processes.

3.1 Collecting Articles

In the absence of sufficient Tigrinya content on Wikipedia[3], the *Haddas Ertra*[4] newspaper provides a large body of professionally edited Tigrinya text, covering diverse domains and has been used as the main source in previous research (Tedla et al., 2016;

[2] We paid an hourly salary of 150 ERN (10.00 USD, at the time of writing) to the annotators.

[3] Tigrinya Wikipedia `https://ti.wikipedia.org`

[4] The newspaper is made available at `www.shabait.com`, the official website of the Eritrean Ministry of Information.

Question Type		Proportion	Example
Which	ኣየነይቲ	31.37%	ኣብ መወዳእታ ወርሒ 1987፣ ሞጆ ናብ ኣየነይቲ ጋንታ ተሰጋጊሩ? [In the last month of 1987, which team did Moje moved to?]
How many/much	ክንደይ	26.23%	ኣብቲ ከባቢያዊ ብከላ ብምስታፉን ዝተቐጽ‘ዓ ኩባንያታት ክንደይ እየን? [How many companies have been fined for participating in the environmental pollution?]
What	እንታይ	14.78%	ምኽትል ተለንተ ሓምድ እድሪስ እንታይ ሓላፍነት ኣለዎ? [What are the responsibilities of Deputy Lieutenant Hamid Idris?]
Who	መን	11.99%	ጸህያይ ባሕሪ ቅድሚ 1866 ‘ዓ.ም ምስ መን ይምደቡ ኔሮም? [What were sea weeds classified into before 1866?]
When	መዓስ	7.60%	ዋንጫ ሃገራት ኣፍሪቃ 2015 መ‘ዓስ ተዛዚሙ? [When did the 2015 Africa Cup of Nations end?]
Where	ኣበይ	4.82%	ኣምባሳደር ‘ዓሊን ሚኒስተር ሓመድን ኣበይ ተራኺቦም? [When did Ambassador Ali and Minister Ahmed meet?]
Why	ስለምንታይ	1.28%	ግራፋይት ስለምንታይ እዩ ተረኹማሽን መጭላቑን ጠባይ ዘለዎ? [Why does graphite have a malleable and slippery character?]
How	ብኸመይ	1.18%	እቶም ናይ ገበን ጉጅለታት ከመይ ጌሮም ነዳዲ ክሰርቁ ክኢሎም? [How did the crime groups manage to steal the oil?]
Other	ጥቐስ/‘ዶ	0.75%	ሰባት ዝጸልእዎም ግን ገንዘብ ደልዮም ካብ ዝገብርዎም ነገራት ኣብነት ጥቐስ? [Give examples of things that people hate but do for money?]

Table 2: Question-type distributions in the development set of TiQuAD, grouped by question words.

Gaim et al., 2021). We collected 550 Tigrinya articles from *Haddas Ertra* covering a wide range of topics, including science, health, business, history, culture, and sports, published in a period of seven years, 2015-2021. The articles that contain at least 500 characters of plain text are kept after filtering out images and tabular content. We split the dataset randomly into training, development, and test sets of 205, 43, and 42 articles, respectively.

3.2 Annotating Question-Answer Pairs

In the first round of annotation, we recruited eight native speakers of Tigrinya [4 female, 4 male] with ages ranging from 20 to 46. Each annotator is presented with a random paragraph from the collection and tasked to write questions that can be explicitly answered by a contiguous segment of text in the provided context. The annotators were encouraged to phrase questions in their own words instead of copying words from the context and to highlight the minimal span of characters that answer the respective question. The annotators were asked to spend on average one minute for each question and answer pair. The end result of this stage is the set of 6,674 unique questions across all documents.

3.3 Collecting Additional Answers

In the second round of annotation, we asked four of the original annotators to provide a second answer to questions in the development and test parts of the dataset. Our annotation tool ensures that annotators cannot give a second answer to the questions they contributed already in the second stage. Finally, we recruited two new annotators to provide a third reference answer to all the questions in the evaluation sets. These annotators were not involved in the first round of the annotation; with no prior exposure to the task, they are expected to show less bias towards the question formulation. We ensure that all entries in the test and development sets have at least three answers from different annotators, resulting in 6,205 answers for 2,056 questions.

3.4 Post-processing Annotations

Throughout the annotation campaign, we collected over 6,674 unique questions and 10,600 answers, i.e., 2,056 of the questions had at least three ground-truth answers by different annotators. From these annotations, we discarded 166 entries (2.5%) that either contained apparent errors, were incomplete, unanswerable by the context, or had a wrong question formulation such as verification (yes/no) and cloze type. For instance, the question “ሽማንጉስ ታሕታይ ካብ ኣስመራ ኣስታት 28 ኪሎ ሜተር ንሸነኽ ___ ርሒቓ ትርከብ ። [*Lower Shmangus is located about 28 km from Asmara on the ___ side.*]" is in cloze format, hence deleted. We also removed outlier entries that had answers with more than 250 characters.

4 Dataset Analysis

To assess the quality and diversity of the dataset, we perform various analyses of the annotations.

4.1 Question-type Analysis

We clustered all questions in the development set into nine types using a manually curated list of question words. As presented in Table 2, the top three types are *which* [ኣየነይቲ/ኣየናይ], *how many/much* [ክንደይ], and *what* [እንታይ], accounting

Reasoning Type	Example	Frequency
Synonymy	Question: ኣብ ርሳስ ዘሎ ንመጽሓፊ ዝጠቅም ማዕድን እንታይ ይበሃል? [What is the mineral in pencils useful for writing?] Context: ...ኣብ ውሽጢ ርሳስ ዝርከብ ንጽሕፈት ዝሕግዝ እምኒ ወይ ማዕድን ግራፋይት ተባሂሉ ይጽዋዕ።... [The stone or mineral inside a pencil that is used for writing is called graphite.]	35.1%
World knowledge	Question: ግራፋይት ኣበየኖት ሃገራት ብዝያዳ ይርከብ? [In which countries is Graphite found?] Context: ...ግራፋይት ኣብ መላእ ዓለም ዳርጋ ብምዕሩይ ዝርጋሐ'ዩ ዝርከብ። ብዝያዳ ግን ኣብ ቻይና፡ ህንዲ፡ ሰሜን ኮርያ፡ ሜክሲኮ፡ ብራዚል፡ ቼክ ሪፓብሊክን ቱርክን ይዝውተር።... [Graphite is almost evenly distributed worldwide. But it is most common in China, India, North Korea, Mexico, Brazil, Czech Republic and Turkey.]	11.1%
Syntactic/ Morphological variation	Question: ኣብ ርሳስ ዘሎ ንመጽሓፊ ዝጠቅም ማዕድን እንታይ ይበሃል? [What is the mineral in pencils useful for writing?] Context: ...ኣብ ውሽጢ ርሳስ ዝርከብ ንጽሕፈት ዝሕግዝ እምኒ ወይ ማዕድን ግራፋይት ተባሂሉ ይጽዋዕ።... [The stone or mineral inside a pencil that is used for writing is called graphite.]	71.4%
Multi-sentence reasoning	Question: ግራፋይት ኣብ ፈኲስ ሓመድ ዝርከብ ምኽኑ እንታይ ጠባይ የስዕቦ? [What characteristics does graphite inherit from existing on the surface of light soil?] Context: ...ግራፋይት ካብ ካልኦት ዓይነታት ማዕድን ዝፈልዮ ነገር እንተሎ፡ ኣብ ተሪር እምኒ ወይ ከውሒ ኣይኮነን ዝርከብ። ኣብ ፈኲስ ሓመድ ብቐጸላታት ተጸፍጺፉ ይርከብ። በዚ ድማ'ዩ ግራፋይት ተረኽማሽን መጭላቕን ጠባይ ዘርኢ።... [One thing that distinguishes graphite from other minerals is that it is not found on hard stone or rock. It is found laid as solid layers on the surface of light soil. That's why graphite exhibits a malleable and slippery character.]	8.5%

Table 3: Reasoning relationships between 100 randomly selected question-answer pairs from the TiQuAD development set. The annotated answer is underlined and the spans that correspond to the reasoning type are colored. Note that samples can exhibit multiple reasoning types; hence the frequency is computed independently.

for ≈72% of all the questions. These types of questions also make up the largest proportions in other datasets (Keren and Levy, 2021; d'Hoffschmidt et al., 2020). Question types that lead to named entity answers such as *who* [መን], *when* [መዓስ], and *where* [ኣበይ] comprise 24%. While there are only 3.2% of the *why* [ስለምንታይ], *how* [ብኸመይ], and *Other* types, they generally necessitate more complex reasoning and are challenging to create during annotation.

4.2 Question-Context Lexical Overlap

The degree of lexical overlap between questions and paragraphs might affect the difficulty of a dataset. To assess this behavior in TiQuAD, we analyzed 100 random samples from the development set and assigned them to four categories of question-context-answer linguistic relationships proposed by Rajpurkar et al. (2016): (1) *Synonymy* implies that key terms in the question are synonyms of words in the context; (2) *World knowledge* implies that the question requires world knowledge to find the corresponding answer in the context; (3) *Syntactic/Morphological variation* implies a difference in the structure between the question and the answer in the context; (4) *Multi-sentence reasoning* implies that answering a question requires combining knowledge from multiple sentences in the context. We observe that syntax and morphology variations are the most common type in TiQuAD. The results of our findings are presented in Table 3.

4.3 Answer correctness and length

We randomly selected 100 question-answer pairs from the validation set to assess the accuracy and length of the answers manually. We specifically check whether each annotated answer is *correct* and has a *minimal* length in answering the corresponding question. We observe that 74% of the answers are accurate and with a minimum span length, while a significant minority, 23%, contain extra information and are longer by a factor of 1.5 on average than the desired span. Only 3% were shorter than the optimal span length, such as partial annotation of the answer.

4.4 Sequence Lengths

The lengths of paragraphs in TiQuAD range between 39-278 words or 198-1264 characters. Around 60% of the questions have 5-10 words, but we observe some verbose examples such as ኣብ ግጥማት ኵ‘ዕሶ እግሪ ደቂ ተባ‘ዕትዮ፣ ኣብቲ ኣብ መንጎ ኮለጅ ሓልሓለን ኮለጅ ጥበባትን ማሕበራዊ ስነፍልጠትን ‘ዓዲቐይሕን ዝነበረ ግጥም ኮለጅ ጥበባትን ማሕበራዊ ስነፍልጠትን ‘ዓዲቐይሕ ክንደይ ሸቶታት ኣመዝጊባ? [*In the men's soccer match between Halhale College and Adikeih College of Arts and Social Sciences, how many goals did Adikeih College of Arts and Social Sciences score in the match?*]. The shortest questions have three words, for example, ፎሮ ኣበይ ትርከብ? [*Where is Foro located?*]. Over 57% of the answers have three or fewer words, but there are cases with up to 32 words that typically constitute a list of items.

4.5 Estimating Human Performance

We assess the human performance on TiQuAD's development and test sets, where each question has at least three answers. In SQuAD, Rajpurkar et al. (2016) use the second answer as the prediction and the rest as ground truths; while in FQuAD, d'Hoffschmidt et al. (2020) compute the average by successively taking each of the three answers as the prediction. For TiQuAD, the third answer is regarded as a prediction, and it is annotated by a control group who had no prior exposure to the task, as elaborated in Section 3.3. We obtain scores of 84.80% EM and 92.80% F1 in the development set, and 82.80% EM and 92.24% F1 in the test set, which are comparable to those of the SQuAD and FQuAD benchmarks.

We analyzed the cases where the human annotators failed to agree and observed that they are mainly due to extra tokens in the answer spans rather than fundamental differences. For instance, the question ጁቬንቱስ ንሰሪ ኣ ብክንደይ ተመርሖ ኣላ? [With how many *[points]* is Juventus leading the Serie A?] has three different annotations: (1) 10 ነጥቢ [*10 points*]; (2) 10 ነጥቢ ፍልልይ [*10 points difference*]; (3) ብናይ 10 ነጥቢ ፍልልይ [*with a 10 points difference*], resulting in zero EM agreement.

5 Experiments

5.1 Model Training

Given a question Q and a context paragraph P from an entry in a QA dataset, the training objective is to predict the start and end positions of the answer span within the paragraph. Following Devlin et al. (2019), we set the input to the transformer model as a concatenation of Q and P, separated by a special delimiter token, `SEP`. Two linear layers, S and E, are introduced to learn the starting and ending positions of answer spans, respectively. Then the probability distributions of token i being the start or the end of an answer span with respect to all tokens in the context can be computed as follows:

$$P_{\text{start}}(i) = \frac{\exp{(S \cdot T_i)}}{\sum_{j=1} \exp{(S \cdot T_j)}}, \quad (1)$$

$$P_{\text{end}}(i) = \frac{\exp{(E \cdot T_i)}}{\sum_{j=1} \exp{(E \cdot T_j)}}, \quad (2)$$

Model	#L	#AH	Param	#Langs
TiELECTRA$_{\text{SMALL}}$	12	4	14M	1
TiRoBERTa$_{\text{BASE}}$	12	12	125M	1
AfriBERTa$_{\text{BASE}}$	8	6	112M	11
XLM-R$_{\text{BASE}}$	12	12	278M	100

Table 4: Models: The monolingual and multilingual pre-trained models used in our experiments. `#L`, `#AH`, and `Param` denote the number of layers, self-attention heads, and parameters, respectively. `#Langs` is the number of languages in the pre-training data. Except for XLM-R, the other models have seen Tigrinya during pre-training.

where T is the model's output of the context sequence, and T_i is the hidden state of the i-th token.

The score for a candidate span (i, j) is defined as the product of the start and end position probabilities, and then the highest-scoring span where $j \geq i$ is used as the final prediction.

$$Score(i, j) = P_{\text{start}}(i) \cdot P_{\text{end}}(j). \quad (3)$$

The loss function $\mathcal{L}$ is the sum of the negative log-likelihoods of the ground truth start and end positions, denoted as i^* and j^*, respectively.

$$\mathcal{L} = -\log P_{\text{start}}(i^*) - \log P_{\text{end}}(j^*) \quad (4)$$

During training, a gradient-based optimizer minimizes the loss and gradually enables the model to accurately predict the answer spans in the context.

5.2 Evaluation Metrics

We use the standard Exact Match (EM) and F1 metrics for evaluation. EM is the percentage of predictions that exactly match the ground truth. F1 score is the average overlap between the predicted tokens and the ground truth, hence rewards partial matches. For both metrics, when there are multiple ground truth answers for a given question in the test set, the final score represents the highest overlap between the prediction and all the reference answers. To improve the robustness of the evaluation, SQuAD (Rajpurkar et al., 2016) removes the English punctuation and articles before computing the scores. Other non-English datasets have also adapted the metrics (d'Hoffschmidt et al., 2020; Möller et al., 2021). In the case of TiQuAD, we remove Tigrinya's articles, common functional tokens, and the punctuation set of its writing system, the Ge'ez Script (Gaim et al., 2022).

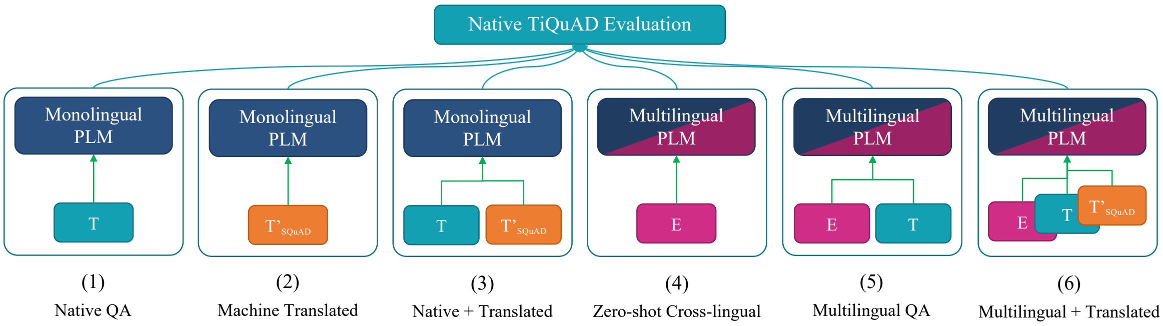

Figure 2: Experimental Setups: Native, Translated, Cross-lingual, and Multilingual Question-Answering settings. T: the native TiQuAD; T'$_{\text{SQuAD}}$: SQuAD1.1 translated to Tigrinya; and E: the English SQuAD1.1.

5.3 Experimental setup

We designed six experimental configurations and evaluated each on four models of varying sizes, ranging from 14 to 278 million parameters. Details of the models are presented in Table 4. The experiments can be grouped into three setups, based on the language of the training data: (1) *Monolingual setting*: We train and evaluate models using the native and machine translated datasets, separately and in combination; (2) *Zero-shot cross-lingual setting*: We investigate transfer learning by training models on an English dataset and evaluating them on Tigrinya – treating QA as a language-independent task; and (3) *Multilingual setting*: We investigate models trained on combined Tigrinya and English QA datasets and evaluated in a native setup. Figure 2 illustrates the experimental settings.

In all experiments, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer with the weight decay parameter set to 0.01 and a learning rate of $3e{-}5$. We set the mini-batch size to 16 and fine-tune for 3 epochs, except in the `Native` settings, where only the small native dataset is used, the batch size and number of epochs are set to 8 and 5, respectively. In the settings where only the small native dataset is used for training, we use a mini-batch size of 8 and fine-tune for 5 epochs; in all other settings, the batch size and the number of epochs are set to 16 and 3, respectively. The experiments were implemented using the HuggingFace Transformers library (Wolf et al., 2020) and ran on a single NVIDIA V100 GPU.

Translation of English dataset For the experiments, we machine translated the training part of SQuAD v1.1 to Tigrinya. The positional information of the answer spans needs to be computed as it is generally lost during translation, making it difficult to retain the original data size. As a remedy, we applied two machine translation services[5] and aggregated the aligned entries, while discarding the spurious ones. This resulted in 46.7K question-answer pairs that we use for model training in our experiments.

	TiQuAD-dev		TiQuAD-test	
Model	EM	F1	EM	F1
Human Performance	84.80	92.80	82.71	92.24
Translated				
TiELECTRA$_{\text{SMALL}}$	38.54	46.04	39.25	48.36
TiRoBERTa$_{\text{BASE}}$	48.50	56.39	48.17	58.81
AfriBERTa$_{\text{BASE}}$	40.36	48.72	40.68	52.96
XLM-R$_{\text{BASE}}$	**51.71**	**59.64**	**53.17**	**62.61**
Native				
TiELECTRA$_{\text{SMALL}}$	36.19	43.06	28.81	37.00
TiRoBERTa$_{\text{BASE}}$	56.21	64.36	53.08	61.82
AfriBERTa$_{\text{BASE}}$	38.01	44.85	35.06	44.24
XLM-R$_{\text{BASE}}$	**56.53**	**65.37**	**55.75**	**65.49**
Translated + Native				
TiELECTRA$_{\text{SMALL}}$	46.36	53.60	47.46	56.64
TiRoBERTa$_{\text{BASE}}$	**62.42**	70.12	62.18	70.42
AfriBERTa$_{\text{BASE}}$	52.68	59.38	47.37	58.35
XLM-R$_{\text{BASE}}$	61.99	**70.44**	**64.76**	**73.53**

Table 5: Performance of models in the Monolingual setups, evaluated on the development and test sets of TiQuAD. The models were trained on TiQuAD (Native) and machine translated SQuAD (Translated).

6 Results and Discussions

In this section, we present and discuss the results of the proposed experimental setups.

6.1 End-to-end Tigrinya QA

In this setup, we train all models on the native and translated Tigrinya datasets then evaluate on the TiQuAD development and test sets. The experimental results are presented in Table 5.

[5] Bing Translator API; Google Translate API

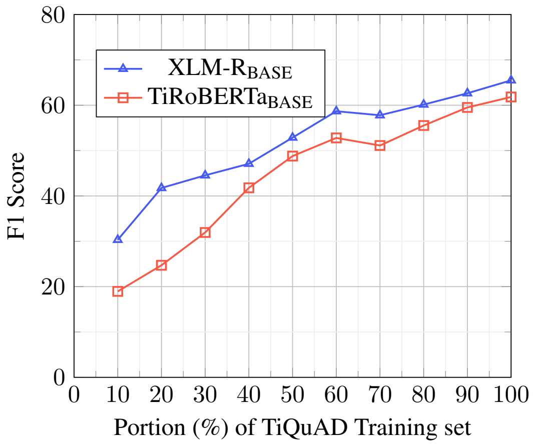

Figure 3: Performance of models with respect to the training data size in the monolingual setup. Evaluated on TiQuAD test set. The x-axis indicates the portion of the training dataset used with an increment of 10%.

Native vs. Translated QA Datasets For models TiRoBERTa$_{\text{BASE}}$ and XLM-R$_{\text{BASE}}$, we observe significant gains when training on the native dataset over the translated one, despite the latter being 10 times larger. The performance of TiRoBERTa$_{\text{BASE}}$ increases by 5 and 3 points in EM and F1 scores on the test set, respectively. However, we observe that the smaller models TiELECTRA$_{\text{SMALL}}$ and AfriBERTa$_{\text{BASE}}$ perform better when trained on the translated data than on the native one. More consistent performance benefits are observed in all models when the two datasets are combined. For instance, TiRoBERTa$_{\text{BASE}}$ gains up to 10 points in EM and F1 than when it is trained on the datasets separately. Overall, our experiments show: (1) A small native dataset can make a positive impact when augmented with larger low-quality data; (2) Machine translated datasets are useful augmentation but can be suboptimal when used alone depending on the quality; and (3) A native dataset could be a vital resource in the evaluation process.

Monolingual vs. Multilingual QA Models When comparing models of comparable sizes, we observe that the monolingual models achieve better performance than their multilingual counterparts. As shown in Table 5, TiRoBERTa$_{\text{BASE}}$ is consistently better than AfriRoBERTa$_{\text{BASE}}$, with gains of 6-15 points in F1 score. Conversely, the larger multilingual model, XLM-R$_{\text{BASE}}$, outperformed all models despite not being exposed to Tigrinya during its pre-training. While TiELECTRA$_{\text{SMALL}}$ trailed in performance in all settings, confirming the impact of model size on the QA task.

	TiQuAD-Dev		TiQuAD-Test	
Model	EM	F1	EM	F1
Human Performance	84.80	92.80	82.71	92.24
SQuAD				
TiELECTRA$_{\text{SMALL}}$	09.85	20.91	09.81	20.41
TiRoBERTa$_{\text{BASE}}$	10.71	20.88	10.88	20.69
AfriBERTa$_{\text{BASE}}$	**20.24**	**32.05**	20.52	32.95
XLM-R$_{\text{BASE}}$	17.99	27.81	**22.66**	**34.44**
SQuAD + Translated				
TiELECTRA$_{\text{SMALL}}$	37.69	46.06	39.07	49.07
TiRoBERTa$_{\text{BASE}}$	51.28	59.25	51.12	60.75
AfriBERTa$_{\text{BASE}}$	44.33	51.43	45.58	56.36
XLM-R$_{\text{BASE}}$	**52.89**	**61.06**	**57.36**	**66.37**
SQuAD + Native				
TiELECTRA$_{\text{SMALL}}$	33.73	41.51	32.74	40.53
TiRoBERTa$_{\text{BASE}}$	57.07	65.75	59.05	67.30
AfriBERTa$_{\text{BASE}}$	51.93	59.66	51.38	62.13
XLM-R$_{\text{BASE}}$	**62.42**	**69.95**	**63.07**	**71.76**
SQuAD + Translated + Native				
TiELECTRA$_{\text{SMALL}}$	45.72	53.40	47.73	57.10
TiRoBERTa$_{\text{BASE}}$	**65.20**	71.88	62.53	71.08
AfriBERTa$_{\text{BASE}}$	51.93	59.47	53.26	63.22
XLM-R$_{\text{BASE}}$	64.78	**72.80**	**68.06**	**76.58**

Table 6: Performance of the Zero-shot Cross-lingual transfer and Multilingual setups on the development and test sets of TiQuAD. The models were trained on the English SQuAD, TiQuAD (Native), machine translated SQuAD (Translated), and their combination.

Training Sample Efficiency To assess the impact of data size, we fine-tuned the TiRoBERTa$_{\text{BASE}}$ and XLM-R$_{\text{BASE}}$ models on subsets of the TiQuAD train set gradually increased by 10% of randomly selected samples and evaluated every step on the test set. We observe a promising trajectory where the models do not show signs of saturation and can potentially benefit from a larger dataset. The progress in F1 score performance is depicted in Figure 3, and a similar trend was observed for the EM score.

6.2 Zero-shot Cross-lingual QA

We investigate the transferability of QA models in a zero-shot setting by training on the high-resource language English and evaluate them on Tigrinya. The multilingual models, AfriBERTa$_{\text{BASE}}$ and XLM-R$_{\text{BASE}}$, trained on the English SQuAD1.1 achieve 32-34% in F1 score on the TiQuAD test set and outperform their monolingual counterparts. While the models show promising results in transferring the task between two linguistically distant languages, those trained on the small native dataset remain vastly superior. Table 6 presents the results of the cross-lingual and multilingual experiments.

6.3 Multilingual QA

In this setup, we train the models on combined English and Tigrinya training datasets, exposing the models to both languages, then evaluate on the native TiQuAD. We observe a consistent improvement in performance across all models in contrast to the previous setups. For instance, XLM-R$_{\text{BASE}}$ in the multilingual setup obtains an increase of over three points in F1 score, setting the state-of-the-art on the TiQuAD test set at 68.06% EM and 76.58% F1 score. Our experiments show that the transfer of models from high to low resourced languages is a viable approach to mitigate the scarcity of annotated datasets. In our case, the benefit emerges when the native and translated Tigrinya datasets are combined with their English counterpart.

7 Conclusion

In this work, we presented the Tigrinya Question Answering Dataset (TiQuAD). The context paragraphs were collected from high-quality News articles of diverse genres, and we collaborated with native speakers to annotate over 6.5K unique questions and 10.6K answers. The development and test sets were further enriched with additional answers to enable a robust evaluation. We conducted comprehensive experiments in monolingual, cross-lingual, and multilingual settings. The estimated human performance on the test set is 81.3% EM and 92.1% F1 score, while the top performing model achieves 68.06% EM and 76.58% F1, leaving a room for future improvements.

Limitations

There are two known limitations of the SQuAD-like annotation approach we used in this work: (1) It can result in higher lexical-overlap between the context and question pairs. (2) It leads to proportionally fewer truly information-seeking questions (Gururangan et al., 2018; Kaushik and Lipton, 2018). The main reason is that the annotators create questions after reading a paragraph, which can induce bias towards recycling words and phrases observed in the context. Our annotation guidelines advise against this, but it is difficult to avoid entirely. Several approaches have been proposed to mitigate this issue, such as Natural Questions (Kwiatkowski et al., 2019) and TyDiQA (Clark et al., 2020). However, they tend to be expensive, and comparatively, the SQuAD-like method is resource efficient and a more suitable starting point for low-resourced languages such as Tigrinya. Finally, the current dataset does not include adversarial examples to measure the capability of models to abstain from providing an answer when it does not exist in the context; this extension is left for future work.

Ethics Statement

This research adheres to the academic and professional ethics guidelines of our university. Our annotation task was approved by the Institutional Review Board (IRB)[6]. All the data collection and annotation procedures were conducted with respect and the informed consent of the participants, and best effort was made to ensure their privacy and autonomy. All participants of annotation tasks indicated their understanding of the procedure for the annotation and acknowledged their agreement to participate. The data sources are published News articles, and for our dataset, we have made an effort to ensure that (1) no personally identifying sensitive information is included, and (2) there exists a fair representation of various genres of news. Furthermore, we ensure that the dataset is available for public use. There may exist inaccuracies or inconsistencies in the questions or answers that could be misleading or ambiguous, potentially due to mistakes and subjective decisions made by the annotators. Furthermore, a bias in the dataset could lead to wrong answers or answers that are only applicable to specific groups of people. We have made the best effort to avoid such issues, but these types of limitations are difficult to detect and remove entirely and potentially present in all similar datasets. The dataset and models released in this work are for research purposes only and may not be suitable for production services without further scrutiny.

Acknowledgements

We would like to thank all the annotators who participated in this work and the anonymous reviewers for their time and constructive feedback. This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

[6] Approval number: KH2018-080

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Casimiro Pio Carrino, Marta Ruiz Costa-jussà, and José A. R. Fonollosa. 2020. Automatic spanish translation of squad dataset for multi-lingual question answering. In *LREC*.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Pavel Efimov, 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.

Fitsum Gaim. 2017. Applying morphological segmentation to machine translation of low-resourced and morphologically complex languages: The case of tigrinya. Master's thesis, School of Computing, Korea Advanced Institute of Science and Technology (KAIST), July.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2021. Monolingual pre-trained language models for tigrinya. In *5th Widening NLP (WiNLP2021) workshop, co-located with the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2022. Geezswitch: Language identification in typologically related low-resourced east african languages. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

Michael Gasser. 2011. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Arefeh Kazemi, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. 2022. Persianquad: The native question answering dataset for the persian language. *IEEE Access*, 10:26045–26057.

Omri Keren and Omer Levy. 2021. ParaShoot: A Hebrew question answering dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 106–112, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. 2021. An exploration of data augmentation techniques for improving english to tigrinya translation. *ArXiv*, abs/2103.16789.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

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

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alp Öktem, Mirko Plitt, and Grace Tang. 2020. Tigrinya neural machine translation with transfer learning for humanitarian response. *arXiv preprint arXiv:2003.11523*.

Omer Osman and Yoshiki Mikami. 2012. Stemming tigrinya words for information retrieval. In *Proceedings of COLING 2012: Demonstration Papers*, pages 345–352.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.* Just Accepted.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wissam Siblini, Charlotte Pasqual, Axel Lavielle, and Cyril Cauchois. 2019. Multilingual question answering from formatted text applied to conversational agents. *ArXiv*, abs/1910.04659.

Yemane Tedla and Kazuhide Yamamoto. 2016. The effect of shallow segmentation on english-tigrinya statistical machine translation. In *2016 International Conference on Asian Language Processing (IALP)*, pages 79–82. IEEE.

Yemane Tedla and Kazuhide Yamamoto. 2018. Morphological segmentation with lstm neural networks for tigrinya. In *Intenational Journal on Natural Language Computing (JNLC)*, volume 7.

Yemane Tedla, Kazuhide Yamamoto, and A. Marasinghe. 2016. Tigrinya part-of-speech tagging with morphological patterns and the new nagaoka tigrinya corpus. *International Journal of Computer Applications*, 146:33–41.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: The case of tigrinya.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th SIGAPP Symposium on Applied Computing*, SAC '22, page 837–844, New York, NY, USA. Association for Computing Machinery.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7, after the conclusion section.

☑ A2. Did you discuss any potential risks of your work?
Section 7, after the conclusion section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.2