From Multiple-Choice to Extractive QA: A Case Study for English and Arabic

Teresa Lynn,[1] Malik H. Altakrori,[2] Samar M. Magdy[1] Rocktim Jyoti Das,[1] Chenyang Lyu,[1] Mohamed Nasr,[2] Younes Samih,[2] Kirill Chirkunov,[1] Alham Fikri Aji,[1] Preslav Nakov,[1] Shantanu Godbole,[2] Salim Roukos,[2] Radu Florian,[2] Nizar Habash[1,3]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]IBM Research AI [3]New York University Abu Dhabi
teresa.lynn@mbzuai.ac.ae, malik.altakrori@ibm.com, nizar.habash@nyu.edu

Abstract

The rapid evolution of Natural Language Processing (NLP) has favoured major languages such as English, leaving a significant gap for many others due to limited resources. This is especially evident in the context of data annotation, a task whose importance cannot be underestimated, but which is time-consuming and costly. Thus, any dataset for resource-poor languages is precious, in particular when it is task-specific. Here, we explore the feasibility of repurposing an existing multilingual dataset for a new NLP task: we repurpose a subset of the BELEBELE dataset (Bandarkar et al., 2023), which was designed for *multiple-choice* question answering (MCQA), to enable the more practical task of *extractive* QA (EQA) in the style of machine reading comprehension. We present annotation guidelines and a parallel EQA dataset for English and Modern Standard Arabic (MSA). We also present QA evaluation results for several monolingual and cross-lingual QA pairs including English, MSA, and five Arabic dialects. We aim to help others adapt our approach for the remaining 120 BELEBELE language variants, many of which are deemed under-resourced. We also provide a thorough analysis and share insights to deepen understanding of the challenges and opportunities in NLP task reformulation.

1 Introduction

Recent years have brought about very fast developments in Natural Language Processing (NLP). However, this progress has not been equal for all languages, and most research has focused on English and a handful of other languages, with the vast majority of other languages being overlooked (Joshi et al., 2020; Rehm and Way, 2023). This is mainly due to the lack of data resources, which hampers the development of NLP tools for these languages. While many resources have been developed for some languages, they are often for very specific tasks, and for very specific formulations of these tasks.

Starting from a motivation of creating cross-lingual Question Answering (QA) datasets for an under-resourced pair, Dialectal Arabic (DA) – Modern Standard Arabic (MSA), we explore the possibility of re-purposing the multilingual BELEBELE dataset (Bandarkar et al., 2023), which was created for the task of multiple-choice question answering, to be suitable for extractive question answering in the style of machine reading comprehension.

In the field of Question Answering, there are a number of different ways to answer a question (Wang, 2022), e.g., (a) **multiple-choice QA** (**MCQA**) *with answer choices, and context provided*, (b) **extractive QA** (**EQA**) *with context provided*, (c) **abstractive QA** *with context provided*, and (d) **open-ended QA** *with no context provided*. An interesting research question is whether we can traverse between them, i.e., if we have a resource for one QA task, can we repurpose it easily for a different reformulation of the task? Here we study this research question, in the context of transferring from MCQA to EQA, as we deem EQA more useful in real-world applications where the correct answer is often not known. Our contributions are:

- We explore the possibility for repurposing a MCQA dataset as an EQA dataset.
- We create and release a new parallel EQA dataset for MSA and English (EN).
- We provide a dialectal Arabic QA benchmark by evaluating our EQA system's performance on monolingual and cross-lingual QA pairs of {EN, MSA}–{EN, MSA, five DAs}.
- We provide guidelines and a strong foundation for creating EQA datasets for the remaining 120 BELEBELE languages.
- We perform careful analysis and discuss the lessons learned as a guide for future research in repurposing NLP datasets.

Proceedings of the 31st International Conference on Computational Linguistics, pages 2456–2477
January 19–24, 2025. ©2025 Association for Computational Linguistics

2 Related Work

2.1 Multilingual QA Datasets

Numerous multilingual QA datasets have already been developed e.g., XQUAD (Artetxe et al., 2020), TyDi QA (Clark et al., 2020), and MLQA (Lewis et al., 2020), which focus on extractive-based QA. The recent ArabicaQA dataset (Abdallah et al., 2024) represents a significant leap forward, being the first large-scale dataset specifically designed for Arabic QA: it has over 89k questions, derived from a diverse set of Standard Arabic documents, providing a robust platform for advancing Arabic NLP, particularly in the context of large language models. Additionally, alternative formats of multilingual QA have been explored, including open-ended QA with Mintaka (Sen et al., 2022) and MCQA with BELEBELE (Bandarkar et al., 2023). With the exception of BELEBELE, these datasets encompass a relatively narrow range of languages and do not include Arabic dialects. To address this gap, we propose leveraging the BELEBELE dataset, to enable the inclusion of its dialectal content into a new parallel DA – MSA/EN dataset.

2.2 Converting MCQA Datasets

Some attempts have been made to convert existing datasets into a different format. Specifically for QA, there has been conversion of MCQA datasets *into* natural language inference (NLI) (Demszky et al., 2018; Khot et al., 2018): the premise is taken from the context paragraph, the correct choice is used as an entailment label, and the other choices are used as neutral or contradictory. Both approaches require effort to rephrase the question into a statement, e.g., using rule-based methods (Demszky et al., 2018) or human annotation (Khot et al., 2018). Hadifar et al. (2023) introduced a new dataset specifically designed for the educational setting, focusing on converting between multiple-choice and open-ended question formats. Automatically generating MCQA has also been explored (Mitkov and Ha, 2003; Kurdi et al., 2020; Ai et al., 2015; Karamanis et al., 2006), where the challenge is to generate distractor choices.

To our knowledge, there is no previous work on converting MCQA to EQA datasets. In this paper, we investigate the feasibility of such a task with an eye towards future automation of the process.

2.3 A Note on Arabic and its Dialects

The Arabic language is a family of variants, among which Modern Standard Arabic (MSA) is the Arab world's shared language of culture, media, and education. However, MSA is not the native language of any speaker of Arabic, whose day-to-day language is typically a local variety, i.e., Dialectal Arabic (DA) that can be quite different from MSA and other dialects (Salameh et al., 2018). MSA and DA coexist in what is called a diglossic relationship where different variants are used in different contexts (Ferguson, 1959). In the context of Arabic QA, there are many efforts that focus primarily on MSA (Shaheen and Ezzeldin, 2014; Mozannar et al., 2019; Biltawi et al., 2021; Alwaneen et al., 2022). A recent exception is the work of Faisal et al. (2021), which works on spoken dialectal QA and includes Arabic among other languages.

The common wisdom in Arabic NLP is that systems should be robust to handling DA forms, but should generate output primarily in MSA. This is reasonable for QA in particular, given the relatively larger trusted content in MSA compared to DA which dominates social media. As such, in this paper, we focus on modeling QA that accepts questions in a range of dialects and provides answers in standard languages (MSA and EN). We selected the following five representative dialects from across the Arab world to work with: Egyptian (arz), Iraqi (acm), Moroccan (ary), Gulf (Najdi; ars), and Levantine (North; apc).

3 MCQA-to-EQA Conversion Challenges

3.1 Belebele Dataset

BELEBELE (Bandarkar et al., 2023) is an open-source *multiple-choice* machine reading comprehension (MRC) dataset that covers 122 languages and language variants, including English, MSA, and five Arabic dialects. Each question is based on a short passage from the Flores-200 dataset (Team et al., 2022) and has four multiple-choice answers, with the correct answer identified. This benchmark dataset enabled the evaluation of natural language understanding (NLU) across high/medium/low-resource languages and language variants.

With respect to the style of the dataset content, the creators of the BELEBELE dataset indicated that they selected multiple-choice questions (MCQs) because *it would lead to the fairest evaluation across languages* and *MCQs enable the ability to scale to many languages when translating from English*.

		Question / Multiple Choice Answers	Paragraph / Span
(a)	**Exact**	Binary digits are also referred to as what? * **Bits** * Jargon * Values * Forms	[...] A binary number can have only one of two values, i.e. 0 or 1, and these numbers are referred to as binary digits - or **$!bits!$,** to use computer jargon.
(b)	**Overlap**	Who discovered nuclear magnetic resonance? * Purcell * Damadian * **Bloch and Purcell** * Bloch and Damadian	MRI is based on a physics phenomenon called nuclear magnetic resonance (NMR), which was discovered in the 1930s by **$!Felix Bloch (working at Stanford University) and Edward Purcell (from Harvard University)!$**. In this resonance, magnetic field and radio waves cause atoms to give off tiny radio signals. [...]
(c)	**Morph**	In response to the protests, which country did not move forward with their signed ACTA agreement? * Germany * Scotland * **Poland** * Lithuania	[...] Last month, there were major protests in Poland when that country signed ACTA, which has led to the **$!Polish government!$** deciding not to ratify the agreement, for now. Latvia and Slovakia have both delayed the process of joining ACTA.
(d) *BB*	**Paraphrase**	What do performers encourage the audience to do during Camille Saint-Saens' opera? * **Partake in the use of cannabis** * Take a trip to Japan * Join them onstage for the performance * Allow their lives to be dictated by what they love	The story presented in the French opera, by Camille Saint-Saens, is of an artist "whose life is dictated by a love for drugs and Japan." As a result, the performers **$!smoke cannabis joints!$** on stage, and the theatre itself is encouraging the audience to join in.
(e) *BB*	**Entailment**	What was the Greek poet Homer unable to do? * Hear * Walk * **See** * Talk	We know many Greek politicians, scientists, and artists. Possibly the most known person of this culture is Homer, the legendary **$!blind poet!$,** who composed two masterpieces of Greek literature: the poems Iliad and Odyssey. [...]
(f)	**Semantic Similarity**	What information regarding the attack have the authorities confirmed? * The identities of any accomplices * **The ethnicity of the suspect** * The motivation behind the attack * The first and last name of the suspect	The man allegedly drove a three-wheeled vehicle armed with explosives into a crowd. The man suspected of detonating the bomb was detained, after sustaining injuries from the blast. His name is still unknown to authorities, although they do know **$!he is a member of the Uighur ethnic group!$**.
(g) *BB*	**Coreference**	Who provides judicial services for FATA? * The Pakistani government * **Political Agents** * Pakistan's president * The British government	Since Pakistani independence from British rule in 1947, the Pakistani President has appointed "Political Agents" to govern FATA, who exercise near-complete autonomous control over the areas. **$!These agents!$** are responsible for providing government and judicial services under Article 247 of the Pakistani Constitution.
(h)	**Context Awareness**	When did Peter Lenz die ? * During the warm-up lap * After falling off his bike * **At the hospital** * While with on-track medical staff	Peter Lenz, a 13-year-old motorcycle racer, has died **$!after being involved in a crash!$** at the Indianapolis Motor Speedway. While on his warm-up lap, Lenz fell off his bike, and was then struck by fellow racer Xavier Zayat. He was immediately attended to by the on-track medical staff and transported to a local hospital where he later died. Zayat was unhurt in the accident.
(i) *BB*	**Pragmatic Knowledge**	At what time in history did Germany exert a strong cultural influence on Estonia? * Around 200 years ago * Around 400 years ago * **Around 600 years ago** * Around 800 years ago	**$!Around the 15th century!$**, northern Estonia was under great cultural influence of Germany. Some German monks wanted to bring God closer to the native people, so they invented the Estonian literal language. It was based on the German alphabet [...] This was the beginning of enlightenment.
(j)	**Restricted Answer**	Where did Cristina Fernandez de Kirchner announce her intention to run? * At a theatre 31 miles away from La Plata * At the Buenos Aires theatre in La Plata * **At the Argentine Theatre 31 miles away from Buenos Aires** * At the La Plata theatre in Buenos Aires	Current senator and Argentine First Lady Cristina Fernandez de Kirchner announced her presidential candidacy yesterday evening **$!in La Plata$!**, a city 50 kilometers (31 miles) away from Buenos Aires. Mrs. Kirchner announced her intention to run for president at the Argentine Theatre, [...]
(k)	**X True/False**	According to the passage, which statement about the suspect is not true? * He used a vehicle during the attack * He allegedly detonated an explosive * His ethnicity is known by authorities * **He was uninjured**	The man allegedly drove a three-wheeled vehicle armed with explosives into a crowd. The man suspected of detonating the bomb was detained, after sustaining injuries from the blast. His name is still unknown to authorities, although they do know he is a member of the Uighur ethnic group.
(l)	**X Entailment**	In what country did Ma study law? * **United States of America** * China * Australia * Hong Kong	Born in Hong Kong, Ma studied at New York University and Harvard Law School and once held an American permanent resident "green card". Hsieh implied during the election that Ma might flee the country during a time of crisis. Hsieh also argued that the photogenic Ma was more style than substance. [...]

Table 1: English examples demonstrating some interesting types of MCQs in the BELEBELE dataset, and the respective issues they pose for extractive QA in this work. The Arabic version is shown in Appendix C.

They worked with a Language Service Provider to create the questions and the multi-choice answers in an iterative process, by developing guidelines that included rules such as *no use of double negatives*, only use answers that *are decently plausible to require the test-taker to fully read and understand the passage* (Bandarkar et al., 2023).

Our effort is part of a larger project on Arabic QA; as such we deemed the BELEBELE dataset to be a suitable starting point for establishing an EQA benchmark for Arabic dialect QA when paired with MSA and English passages. Our five dialects of focus represent the full spectrum of variations across the Arab world: Gulf, Levantine, Iraqi, Egyptian, and Moroccan. We have questions in English, MSA and Arabic dialects, with the answers in MSA and English only.

Broadly defined, our first task in the conversion from MCQA to EQA is to identify the spans in the passages that answer the MCQs.

3.2 The MCQ Zoo

When examining the different BELEBELE MCQs and their associated passages, we recognized a considerable number of MCQ types that vary on a continuum from simple word matching to required passage and world understanding. Table 1 presents a set of English examples along this continuum to highlight the challenges that hindered us from employing a fully automated process in identifying text spans in the associated passages. The Arabic version is in Appendix C.

The simplest MCQ types are perhaps **Exact Match**, Partial **Overlap** Match and **Morph**ological Match, Table 1.(a, b, c), respectively, where the MCQ answer is "present" in the passage. Relatively more complex are cases that may be resolvable through **Paraphrasing**, **Entailment**, or **Semantic Similarity**, in Table 1.(d, e, f), respectively. Next comes the cases that require higher levels of awareness of context, from that of **Coreference** in the text all the way to **Pragmatic Knowledge**, in Table 1.(g-i). Finally we present examples of complex kinds of MCQ, where wording of the question is highly dependent on the provided multi-choice answers. While a span can be identified in principle in Table 1.(j), the question may not stand alone with the passage in Table 1.(k, l).

These challenges made it apparent that (a) not all MCQAs are usable for the EQA task; and (b) that human annotation was a more realistic approach than automating the conversion of the dataset. While the first few MCQ types suggest that an automatic process could be used to identify the QA span, the rest, in increasingly difficulty, show that the MCQ answers are not as helpful. The last two types are not usable in our assessment for this task. We took all of these insights into consideration as we developed the guidelines and annotated the spans, which we discuss next.

4 Annotation Process

In this study, *annotation* refers to marking the appropriate 'answer span' in the original passage in response to a question, or labelling a question as unanswerable ('X'), as shown in Table 1. Our annotations focused on EN and MSA parallel passages from the BELEBELE dataset and we call our annotated set BELEBELE-EQA. Our repurposing approach involved the following steps.

4.1 Automatic Filtering

The original BELEBELE dataset contains 900 question-answer pairs. However, as highlighted in Section 3.2, many of these multiple-choice questions did not suit the EQA task. As a first pass, we automatically removed QA pairs that were deemed unsuitable based on the following set of keywords: "*Which of the following*", "*Select all that apply*", "*Choose the correct*", "*According to the passage*", "*Based on the information given*". Given that the dataset is parallel, we carried out this automatic filtering on the English content first and subsequently removed the same QA pairs in the corresponding MSA content. After this first filtering pass, we were left with 415 QA pairs, whose EQA answer spans we annotated manually.

4.2 Pilot study

We first carried out a pilot annotation study in order to fully understand the nature of the BELEBELE dataset and develop our annotation guidelines. The pilot annotation study was carried out on the first 100 QAs for both EN and MSA. The annotators were presented with the BELEBELE passage, the original BELEBELE question, the 'correct answer' from the multiple-choice answer list[1] and a column in which to write the span. Two native Arabic speakers annotated 50 MSA QAs each, and two fluent English speakers each annotated 50 EN QAs. It was during this pilot study that the extent of the challenges highlighted in Section 3 became known.

4.3 Annotation Guidelines

Informed by the findings of the initial pilot study, we created two separate versions of the guidelines for English and Arabic, accounting for linguistic differences that influenced the wording of the instructions and the need for language-specific examples. See Appendices D and E for the complete EN and MSA guidelines, respectively.

Also informed by the pilot study, the original multiple choice answer was hidden during the annotation, as it had been deemed to be more distracting than helpful in the pilot round (see discrepancies between MC answers and spans in Table 1).

As part of the annotation task, the annotators would mark the span in a copy of the passage using `$!(text)!$` as a span delimiter, while keeping the span as short as possible – in the style of

[1] Available from the BELEBELE data release https://github.com/facebookresearch/belebele

SQUAD (Rajpurkar et al., 2016).

Annotators were also given clear guidance on what to do with articles, prepositions, punctuation, and instances where the answer appeared more than once in the passage (Appendices D and E). Equipped with better awareness of the issues described in Section 3.2, for each QA pair, the annotators could either: (i) annotate the span and if necessary, label the QA as 'BB' to indicate it as being complex and problematic, as per Table 1.(d,e,g,i) or (ii) label the QA as 'X' to be removed, due to the answer being impossible to find, as per Table 1.(k,l). Finally, annotators were advised that it was more important to find a relevant span (in which evidence of the answer could be found) than finding an exact span that may not exist. This guidance was a significant help given the many differences between the nature of MCQA and EQA.

4.4 English Data Annotation

We took a bootstrapping approach to annotation across the two languages. Given the parallel nature of the dataset, the EN annotation was carried out first, followed by a semi-automatic approach to annotating the MSA data (Section 4.5). The annotations were done on the remaining 415 QA pairs in the BELEBELE dataset (including a review and correction of the first 100 from the pilot round). Eleven annotators, both fluent and native English speakers, annotated the QA pairs, with the annotations subsequently reviewed by independent reviewers.

The distribution of the QA pairs was as follows: 50 QA pairs were used in our IAA study (see Section 4.6) and were annotated by three annotators. The remaining 365 pairs were distributed amongst the other eight annotators, where six annotators had 50 pairs each, and the remaining two annotators had 42 and 23, respectively. The annotators were given the English BELEBELE passage, the original question and a copy of the passage to mark the answer span. They could also add comments for problematic instances that required broader discussion or verification.

As a result of the annotation process, 86 out of the 415 QA pairs were labelled as 'X' and excluded. Out of the remaining 329 pairs, 44 were labelled as 'BB'. These BB QAs remain part of the final dataset, as we are also interested in establishing what difficulties they pose for a QA system, given the need for reasoning and resolving linguistic phenomena like paraphrasing, pragmatics and coreference resolution.

4.5 Arabic Data Annotation

For the MSA data, we exploited the parallel nature of the dataset to project pre-annotations from the EN passages to the MSA passages as part of a bootstrapping approach. The goal was to maximize the alignment between the EN and the MSA spans that provide answers to the questions. Our semi-automated approach followed the following steps:

1. Marking as 'X' the MSA questions labelled 'X' in English to avoid re-annotation. In our experience, this approach proved consistent for both English and MSA.
2. Translation of the EN span to MSA using Helsinki MT tool (Tiedemann and Thottingal, 2020)
3. Using a sliding-window approach to identify the MSA span with the same length and highest n-gram overlap with the EN span.
4. Marking an approximate location of the EN span in the MSA passage
5. If no equivalent translation was found, an approximate MSA location based on the EN span's word offset was used.[2]
6. Native Arabic speakers reviewed the pre-annotated spans according to the MSA-based annotation guidelines and made corrections where necessary.

The 415 QA pairs were divided amongst four native Arabic speaking annotators, with the annotations subsequently reviewed by independent reviewers. The annotators were presented with the MSA BELEBELE passage, the original MSA BELEBELE question and a copy of the passage where the span was *pre-annotated*. During the review process, the annotators could either: leave the span unchanged or *modify* the span markers, and label as 'BB' where necessary, or label the QA as 'X' to be removed.

The MSA guidelines are generally similar to the EN guidelines except for containing Arabic examples instead of English. The MSA guidelines match EN in using white-space boundaries as the primary word delimiter. This unavoidably leads to a difference in span scope from EN since Arabic is a morphologically rich language with numerous clitics. As a result some proclitic prepositions, conjunctions, and the definite article, as well as pronominal enclitics are kept inside the spans.

[2] An example of this case is when the span is 1-2 words long and its translation is not found in the MSA passage.

4.6 Inter-Annotator Agreement

We carried out an inter-annotation (IAA) study in order to assess the level of agreement across annotators, and get an indication of how reliable the guidelines were.[3] For the IAA study, we selected 50 English QA pairs for annotation by three separate annotators. The three annotations for each QA pair were then compared. Instead of calculating an F-score measure, which would give indication of quality of annotations, we utilize the γ measure for Inter-Annotator Agreement and Alignment (Mathet et al., 2015).[4] This measure involves identifying the optimal alignment among annotations provided by multiple annotators, aiming to align annotations with high similarity and compute the average dissimilarity between them. Specifically, the average dissimilarity is calculated by averaging dissimilarities among aligned annotations, considering both their categories and positions. The alignment with the minimum mean dissimilarity, selected from all potential alignments initially evaluated during computation, is employed. The final inter-annotator agreement (γ measure) was computed on the three sets of 50 QA pairs yielding a high γ value of 0.81, indicating that the designed annotation guidelines are reliable. Given that the MSA annotation guidelines were based on the EN guidelines, we posit that their quality was just as high. The three sets of 50 EN IAA QAs were merged and reviewed to produce a final version for inclusion in our parallel dataset. See Appendix A for more details and visualization of the span combination process.

5 Experiments and Results

5.1 Experimental Setup

Datasets We use two datasets in our experiments: the commonly used state-of-the art QA dataset, SQuAD (Rajpurkar et al., 2016) and our BELEBELE-EQA, both of which are Wikipedia-based. For SQuAD, we only used the validation portion of the dataset which contains 10,657 English EQA pairs. For BELEBELE-EQA, we have two versions: (i) BELEBELE-EQA-**All** contains 329 QA pairs for EN, MSA and five dialects, and (ii) BELEBELE-EQA-**Sub** excludes 44 BB-annotated questions, leaving 285 QA pairs.

QA Evaluation Tool To run the QA task on the above datasets, we used PrimeQA (Sil et al., 2023), an open-source[5] tool that is built on top of the HuggingFace (Wolf et al., 2020) library to support research on running different QA tasks. We used two existing models, namely **PrimeQA/NQ+TyDi**[6] and **PrimeQA/NQ+TyDi+SQuAD**,[7] which are 550m parameters XLM-R$_{Large}$ (Conneau et al., 2020) based models that were finetuned for the multilingual TyDi QA task (Clark et al., 2020). Both models were finetuned on the Natural Questions dataset (Kwiatkowski et al., 2019) and TyDi (Clark et al., 2020). The latter, however, was further finetuned on the training portion of the SQuAD dataset.

Evaluation Metrics We report the results using the commonly used F1-score and Exact Match (EM) score. F1-score is the harmonic mean of Precision and the Recall on word-level uni-grams. The default F1 and EM scores as implemented in PrimeQA toolkit preprocess references and predictions by dropping punctuation, extra white space, and for EN, lower casing and removing the articles *a/an/the*. We also introduce two normalized variants, F1^n and EM^n, where we normalize references and predictions by removing the respective language stopwords as defined in NLTK (Bird and Loper, 2004).

5.2 Results and Analysis

In this section, we present the evaluation results for the BELEBELE-EQA and SQuAD datasets.

SQuAD vs. BELEBELE-EQA We evaluate the SQuAD dataset as well as the EN-EN QA pairs from the BELEBELE-EQA dataset using the PrimeQA toolkit (See Appendix B for details) This experiment allows us to (a) verify that the chosen pretrained models are capable of solving the QA task at-hand without further finetuning, and (b) establish a sense of *difficulty* for the BELEBELE-EQA datasets compared to the commonly used SQuAD. We observe the following based on the results in Table 2: First, on SQuAD, the F1 shows a negligible difference between the two pretrained models, while EM is surprisingly lower by 3.4% for the model that was finetuned on the SQuAD dataset.[8]

[3] As the dataset was small and fully manually reviewed, the IAA study did not serve to evaluate the data quality.

[4] Gamma ranges between 0 and 1 where a higher value indicates higher agreement between annotators.

[5] https://github.com/primeqa/primeqa

[6] PrimeQA/nq_tydi-reader-xlmr_large-20221210

[7] PrimeQA/nq_tydi_sq1-reader-xlmr_large-20221110

[8] The state-of-the-art F1 and EM scores on SQuAD are 93.2 and 90.9, respectively, as per the public PapersWithCode leaderboard: https://paperswithcode.com/sota/question-answering-on-squad20

Dataset	Questions	F1	EM
PrimeQA/NQ+TyDi			
SQuAD	10,570	90.5	82.5
BELEBELE-EQA-All	329	68.1	46.5
BELEBELE-EQA-Sub	285	71.3	50.5
PrimeQA/NQ+TyDi+SQuAD			
SQuAD	10,570	90.6	79.1
BELEBELE-EQA-All	329	71.0	50.5
BELEBELE-EQA-Sub	285	76.6	56.1

Table 2: Evaluating the SQuAD and BELEBELE-EQA datasets using the PrimeQA dataset. (EM: Exact Match. Scores are percentages. Higher is better.)

In contrast, the results on BELEBELE-EQA-All show that additional finetuning on SQuAD yielded better performance both in F1 and EM.

Second, the best BELEBELE-EQA (All and Sub) F1 scores (71.0 and 76.6) are 18.6 and 14 points less than SQuAD's F1 score, respectively. Similarly, the best respective EM scores are 50.5 and 56.1, or 28.6 and 23 points less than that of the SQuAD dataset. Finally, and as expected, removing the BELEBELE-EQA hard questions (BELEBELE-EQA-Sub) resulted in better performance on both F1 and EM measures.

BELEBELE-EQA Results Based on the baseline results presented above, we report in Table 3 on the different question language setups using the best identified model: **PrimeQA/NQ+TyDi+SQuAD**. For specific experiments, we refer to the pair of languages in **passage-question**, e.g., EN-EN, or MSA-Iraqi.

Monolingual Experiments (EN-EN, MSA-MSA) The basic standard monolingual setups have the highest performing F-1 scores, with EN being higher than MSA by 9% absolute. Despite being a multilingual model, this performance difference is not unexpected given the relative resource richness of EN compared with MSA.

Cross-Lingual Experiments (MSA-EN, EN-MSA) The basic cross-lingual setups are next in rank of performance with MSA-EN F1 being only slightly better than EN-MSA (0.5% absolute). The performance drop due to question language is much higher for EN passages (11.1% absolute) compared to MSA passages (1.6% absolute). This suggests that the model struggles in cross-lingual settings in ways independent of its monolingual performance.

Dialectal Experiments (MSA-DA, EN-DA) The DA questions consistently lead to lower performance compare to the above-mentioned settings, and with performance with MSA passages being better on average than on EN passages (by 6.3% absolute). The specific dialects seem to follow a common order of Gulf>(Egyptian|Levantine)>Iraqi>Moroccan, which we speculate may reflect the order of their question sets' similarity to MSA. Automatically translating dialectal questions into the target passage language using Google Translate[9] results in notable improvements: an average increase of 18% for English passages and 4.4% for MSA passages (All setting). Moroccan saw the greatest benefit from this translation method, with F1 increases of 26.7% for EN and 11.4% for MSA.

BELEBELE-EQA All vs Sub Across all experiments, the easier subset (Sub) has higher scores than (All): on average, 2.6% absolute F1 increase for EN passages, and 1.5% for MSA passages. The highest difference is for EN-EN (5.4% absolute F1) and the lowest for MSA-Gulf (0.6%).

Normalized F1 and EM The normalization results are generally consistent in rank order; but the effect of normalization in EN is much higher than MSA: F1 (2.3% vs 1.7%) and EM (5.8% 4.3%). We note that the normalization process reduces the length of references and predictions by 30% for EN, but only 10% for MSA.

6 Discussion

Our motivation for this study was to create a practical and reliable multilingual EQA benchmark dataset, without the need to start from scratch or resort to creating synthetic data. Given our languages of interest, repurposing the BELEBELE MCQA dataset seemed like a feasible solution. However, answering the research question on how feasible this approach is involves considering the various insights and findings presented in this paper.

We found that while repurposing is possible, our exploratory study showed that it required significant manual effort. Contrary to initial assumptions, we discovered that the original answers from the multiple-choice question pairs were not suitable for automated approaches. In fact, the only multiple-choice types in Table 1 that resemble non-complex SQUAD-style QA pairs are *(a) Exact*.

[9]For DA-to-English, we translated using Google Translate (AR-to-EN), while for DA-to-MSA, we used English as an intermediate language (DA ⇒ EN ⇒ MSA): Google Translate (AR-to-EN ⇒ EN-to-AR).

Passage	Question	All				Sub			
		F1	EM	F1^n	EM^n	F1	EM	F1^n	EM^n
EN	EN	71.0	50.5	74.0	60.2	76.6	56.1	79.5	65.6
	MSA	59.9	39.8	62.7	47.7	62.9	43.5	65.6	51.2
	DA–Iraqi	44.5	27.7	46.2	31.3	46.0	29.5	47.8	33.0
	DA–Levantine	49.5	31.9	51.2	36.5	51.9	35.1	53.7	39.6
	DA–Gulf	50.4	31.3	52.6	37.4	52.1	33.0	54.2	38.9
	DA–Morrocan	36.3	18.5	38.5	21.0	37.0	20.0	39.3	22.1
	DA–Egyptian	48.8	30.1	51.5	36.2	51.3	33.3	54.0	39.3
	DA–Average	45.9	27.9	48.0	32.5	47.7	30.2	49.8	34.6
	DA–Iraqi (T$_{EN}$)	59.8	39.2	62.7	46.8	64.2	42.8	67.2	50.9
	DA–Levantine (T$_{EN}$)	65.4	46.2	68.2	53.5	69.7	50.2	72.3	57.9
	DA–Gulf (T$_{EN}$)	65.4	43.8	68.2	52.9	69.5	48.1	72.2	57.2
	DA–Morrocan (T$_{EN}$)	63.0	41.0	65.7	49.5	67.0	44.9	69.6	53.0
	DA–Egyptian (T$_{EN}$)	65.7	45.0	68.7	53.8	70.0	49.5	72.9	58.2
	DA–Average (T$_{EN}$)	63.9	43.0	66.7	51.3	68.1	47.1	70.8	55.4
MSA	EN	60.4	41.6	62.3	46.8	64.5	46.0	66.7	50.9
	MSA	62.0	40.7	64.3	46.2	65.6	44.2	68.1	49.5
	DA–Iraqi	50.3	32.2	51.2	35.0	54.2	36.1	55.6	39.3
	DA–Levantine	54.2	35.3	55.7	38.6	58.5	39.3	60.4	43.2
	DA–Gulf	56.6	36.2	58.4	40.7	59.3	39.6	61.3	44.2
	DA–Morrocan	45.1	26.7	46.6	30.1	48.1	29.8	49.6	33.3
	DA–Egyptian	54.5	34.0	56.6	39.2	57.3	37.5	59.5	42.5
	DA–Average	52.1	32.9	53.7	36.7	55.5	36.5	57.3	40.5
	DA–Iraqi (T$_{MSA}$)	51.5	31.3	53.1	35.3	53.9	34.0	55.7	37.9
	DA–Levantine (T$_{MSA}$)	58.5	38.3	60.3	43.2	62.1	42.1	64.1	46.7
	DA–Gulf (T$_{MSA}$)	57.9	38.0	59.9	42.9	61.0	40.7	63.3	46.0
	DA–Morrocan (T$_{MSA}$)	56.5	36.2	58.6	42.6	59.4	39.3	61.5	45.3
	DA–Egyptian (T$_{MSA}$)	58.5	37.7	60.4	42.2	62.1	41.1	64.2	46.0
	DA–Average (T$_{MSA}$)	56.6	36.3	58.5	41.2	59.7	39.4	61.8	44.4

Table 3: Monolingual and Cross-lingual evaluation of the BELEBELE-EQA dataset using the PrimeQA toolkit with PrimeQA/NQ+TyDi+SQuAD. (EM: Exact Match; F1^n and EM^n are the text-normalized versions of F1 and EM, respectively; DA: Dialectal Arabic; T$_{EN/MSA}$: Question translated to EN/MSA; Scores are %s; Higher is better.)

Rather than offering a scalable solution here, instead our manual efforts and the subsequent insights serve to make this repurposing task much more feasible and potentially automated by other dataset developers. We can see that the repurposing process is much more feasible when the differences between the QA styles are fully understood. We found that, through an effective filtering process, we could easily eliminate many QA pairs that did not lend themselves easily to EQA. While this resulted in reducing the dataset size by more than half, this leaves open new avenues for examining how those 'unsuitable' QA pairs could be reformulated to be useful in EQA. Repurposing a multilingual dataset like BELEBELE means that its parallelism could be exploited through a bootstrapping approach. We have learned that once one language dataset is manually annotated and verified, the spans could be relatively easily projected to a new language from the wider dataset provided a translation system is available (see Section 4.5). Likewise, few changes were required when adapting our annotation guidelines from EN to MSA, suggesting that this adaptability may hold for other languages. Finally, our experimental results confirm that our approach produced a useful data set that can be used for benchmarking cross-lingual EQA for all BELEBELE languages.

7 Conclusion and Future Work

We repurposed the MCQA BELEBELE dataset to create a new EQA dataset for English and MSA, with cross-lingual benchmark results for dialectal Arabic QA.[10] We also provided guidelines for practitioners exploring QA in other BELEBELE languages. Our work, including negative insights, aims to enhance understanding of the challenges and opportunities in reformulating NLP tasks.

For future work, we plan to support repurposing BELEBELE-EQA for other languages, potentially automating the process. The dataset offers a resource for evaluating such methods. We also aim to use it for fine-tuning general-purpose Arabic QA models and improving EQA evaluation metrics, focusing on span length.

[10] https://github.com/mbzuai-nlp/MultiChoice2ExtractiveQA

Limitations

We would like to point to some limitations of our study. First, while we are addressing a general problem, in our study, out of all the other BELEBELE under-resourced languages, we only focused on cross-lingual evaluation of five Arabic dialects. While we believe that our findings are more generally applicable, we would like to note that each language and dataset may present unique challenges and opportunities. Second, our repurposing was based on a single dataset, BELEBELE, and for a single task (extractive question answering in the style of machine reading comprehension), and we note that other tasks may require some specific considerations.

Ethics and Broader Impact

Data Bias. It is important to acknowledge the possibility of biases within our dataset due to subjective human judgments, or due to the influence of the original task's data (e.g. a narrow cultural representation across the questions).

Broader Impact and Potential Use. Our repurposing insights presented here can help speed up the creation of resources and tools for resource-poor languages and dialects. We do not see any immediate threats from its use.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for Arabic question answering. *Preprint*, arXiv:2403.17848.

Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit. 2015. Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 26–33, Beijing, China. Association for Computational Linguistics.

Tahani H Alwaneen, Aqil M Azmi, Hatim A Aboalsamh, Erik Cambria, and Amir Hussain. 2022. Arabic question answering system: a survey. *Artificial Intelligence Review*, 55(1):207–253.

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.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv e-prints*, pages arXiv–2308.

Mariam M Biltawi, Sara Tedmori, and Arafat Awajan. 2021. Arabic question answering systems: gap analysis. *IEEE Access*, 9:63876–63904.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315.

Charles A Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. 2006. Generating multiple-choice test items from medical text: A pilot study. In *Proceedings of the Fourth International Natural Language Generation*

Conference, pages 111–113, Sydney, Australia. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

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

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.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

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.

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.

Georg Rehm and Andy Way. 2023. *European Language Equality: Introduction*, pages 1–10. Springer International Publishing, Cham.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohamed Shaheen and Ahmed Magdy Ezzeldin. 2014. Arabic question answering: systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, 39:4541–4564.

Avirup Sil, Jaydeep Sen, Bhavani Iyer, Martin Franz, Kshitij Fadnis, Mihaela Bornea, Sara Rosenthal, J Scott McCarley, Rong Zhang, Vishwajeet Kumar, et al. 2023. PrimeQA: The prime repository for state-of-the-art multilingual question answering research and development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 51–62, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Zhen Wang. 2022. Modern question answering datasets and benchmarks: A survey. *arXiv e-prints*.

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.

A Processing IAA Spans for Visualization

In Figure 1, we visualize the degree of agreement among the annotators on the 42 QA pairs that remained from the IAA study (See Section 4.6). We combine the annotation spans into one artificial document in order to visualize the results in a meaningful way where this process is explained later in this section.

In this figure, the x-axis represents the artificial document that is 450 words long. The vertical, colored lines represents a span in that document. The three different colors represent the three annotators and the number on the span is the question number. 42 out of the 50 QA pairs remained as eight questions were labelled as 'X' by all three annotators and were excluded. As annotators were allowed to label a questions as 'X', some questions have less than three annotations, e.g. Question 38.

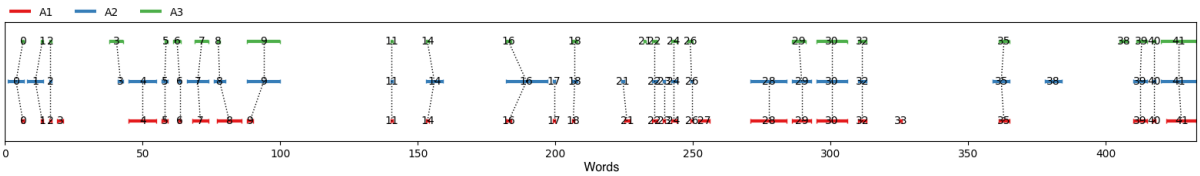

Figure 1: Visualization of the IAA agreement for 42 questions. The colored lines indicate a span. Each color represents an annotator. The number on the colored line is a question ID.

A.1 Combining the Spans Into One Artificial Document.

While we could visualize the agreement between annotators for each QA pairs separately, this would have resulted in 42 different figures, one for each QA pair. Instead, we decided to combine the annotated spans from each annotator to create one large artificial document where the annotated spans would appear in different locations. One simple approach to this is combining the whole passages which will offset the location of the annotated span by the length of all the passages that came before it. This, however, will result in a huge document that is too long to have a meaningful visualization.

A.2 Compressing the Artificial Document.

To overcome this issue, we *compressed* the document by removing the large un-annotated text spans. Note that the annotation are technically a start and end positions in a document and our task is basically to measure (and visualize) the chosen start and end position of each QA pair across the three annotators. This means that the QA pairs are independent, and ordering them in the document or shifting their locations –as long as they do not overlap and they maintain their original length– should not affect them.

To illustrate, assume that annotator (a_1) for passage (p_1) chose the span (3, 7) indicating that the answer is from word (w_3) to (w_7) while annotator (a_2) chose the span (4, 10). We can see that (a_1) and (a_2) *disagree* in the regions: (w_4) and from (w_8) to (w_10). If we add passage (p_0) which contains 20 words before (p_1) to make one document resulting in shifting all the words in (p_1), as well as in the spans (a_1) and (a_2) by 20 words, we see that the disagreement remains the same. After the shift, (a_1) and (a_2) change from (3, 7) and (4, 10) to (23, 27) and (24, 30), respectively. As can be noticed, the disagreement between (a_1) and (a_2) remained in two regions with the same length for each region, that is (w_24) and (w_28) to (w_30).

B PrimeQA Experimental Details

The experiments discusses in Section 5.2 were conducted on an Apple M1 Max MacBook Pro with 32GBs of memory. Since we used the aforementioned models in inference mode, each experiment with 329 questions (E.g. the combination of En Passage and En Questions) took approximately 30 seconds to complete. Note that additional time may be required to download each model for the first time. The 56 core experiments (without translation) in Table 3 (2 Passages x 7 question types x 2 configurations x 2 normalization schemes) are expected to require 28 minutes of running.

C Arabic Examples of MCQ-Span Types

		Question / Multiple Choice Answers	Paragraph / Span
(a)	Exact	بماذا يُشار أيضاً إلى الأرقام الثنائية؟ * **بتّات** * لغة اصطلاحية * قيم * أشكال	[...] يمكن أن تكون قيمة الرقم الثنائي قيمة واحدة فقط من قيمتين، وهما 0 أو 1، ويُشار إلى هذه الأرقام على أنها أرقام ثنائية - أو **$!بتّات!$**، إذا استخدمنا لغة الحاسوب الاصطلاحية.
(b)	Overlap	من اكتشف الرنين المغناطيسي النووي؟ * بورسيل * داماديان * **بلوخ وبورسيل** * بلوخ وداماديان	"يعتمد التصوير بالرنين المغناطيسي على ظاهرة فيزيائية تسمى الرنين المغناطيسي النووي (NMR)، والتي اكتشفت في ثلاثينيات القرن الماضي عن طريق **$!فيليكس بلوخ (من جامعة ستانفورد) وإدوارد بورسيل (من جامعة هارفارد)!$** في هذا الرنين، تجعل المجالات المغناطيسيّة والموجات الراديويّة الذرّات تبعث إشارات راديو دقيقة. [...]
(c)	Morph	رداً على الاحتجاجات، أي دولة لم تمضي قدماً في اتفاقية مكافحة التزييف التجارية الموقعة ؟ * ألمانيا * **بولندا** * إسكتلندا * ليتوانيا	[...] في الشهر المنقضي، شهدت بولندا احتجاجات كبرى عندما وقعت تلك الدولة على اتفاقية مكافحة التزييف التجارية، الأمر الذي أدى إلى اتخاذ **$!الحكومة البولندية!$** قراراً بعدم التصديق على الاتفاق في الوقت الراهن. أجلتا كلا لاتفيا وسلوفاكيا انضمامهما إلى اتفاقية مكافحة التزييف التجارية."
(d) BB	Paraphrase	ماذا يشجع فناني الأداء الجمهور على القيام به خلال أوبرا كاميل سان سايِنز؟ * **المشاركة في تعاطي القنب** * رحلة إلى اليابان * انضم إليهم على خشبة المسرح للأداء * يدعوا ما يحبونه يحدد حياتهم	القصة التي قُدمت بالأوبرا الفرنسية، لكاميل سان سايِنز، هي قصة فنان "حياته محكومة بحب المخدرات واليابان." ونتيجة لذلك، **$!يدخن الفنانون لفائف القنب !$** على خشبة المسرح، ويشجع المسرح نفسه الجمهور على مشاركتهم في ذلك.
(e) BB	Entailment	ما الذي كان الشاعر اليوناني هومر غير قادر على القيام به؟ * السمع * **الرؤية** * المشي * الكلام	نعرف الكثير من السياسيين والعلماء والفنانين اليونانيين . قد يكون الشخص الأكثر شهرة الذين نعرفهم من هذه الثقافة هو هوميروس، الشاعر الأسطوري الذي **$! فقد بصره،!$** والذي ألف رائعتين من روائع الأدب اليوناني، ألا وهما الإلياذة والأوديسا . [...]
(f)	Semantic Similarity	ما هي المعلومات المتعلقة بالهجوم التي أكدتها السلطات؟ * هويات جميع المشاركين * **عرق المشتبه به** * الدافع وراء الهجوم * الاسم الأول والأخير للمشتبه به	زُعم أن الرجل قاد سيارة ذات ثلاث عجلات ملغمة بالمتفجرات وسط حشد من الناس. تم اعتقال الرجل المشتبه في أنه قام بتفجير القنبلة بعد إصابته بجروح جراء الانفجار. ولم تعرف السلطات اسمه حتى الآن، بالرغم من أنهم يعرفون **$!أنه من مجموعة الإيغور العرقية!$**.
(g) BB	Coreference	من يقدم الخدمات القضائية للمناطق القبلية الخاضعة للإدارة الاتحادية؟ * الحكومة الباكستانية * **عملاء سياسيون** * الرئيس الباكستاني * الحكومة البريطانية	"ومنذ استقلال باكستان عن الحكم البريطاني في عام 1947، عيّن الرئيس الباكستاني "عملاء سياسيين" لحكم المناطق القبلية الخاضعة للإدارة الاتحادية، والذين يمارسون سيطرة شبه كاملة على المناطق. **$!هؤلاء الوكلاء!$** مسؤولون عن تقديم الخدمات الحكومية و القضائية بموجب المادة 247 من الدستور الباكستاني."
(h)	Context Awareness	متى مات بيتر لينز؟ * خلال دورة الإحماء * بعد سقوطه من دراجته * **في المستشفى** * أثناء وجوده مع الطاقم الطبي على الطريق	توفي راكب الدراجات النارية ذا ال13 عام بيتر لينز **$!بعد كونه جزءاً في حادث!$** حلبة إنديانابوليس موتور سبيدواي. بينما كان لينز في جولة الإحماء، سقط من دراجته، ثم صدمه المتسابق زافيير زيات زميله. عولج على الفور من قبل الطاقم الطبي على الطريق ونُقل إلى مستشفى محلي حيث توفي لاحقاً. ولم يتعرض "الزيات" لأي إصابة في الحادث.
(i) BB	Pragmatic Knowledge	في أي وقت من التاريخ مارست ألمانيا تأثيراً ثقافياً قوياً على إستونيا؟ * منذ حوالي 200 عام * منذ حوالي 400 عام * **منذ حوالي 600 عام** * منذ حوالي 800 عام	كان شمال إستونيا في **$!حوالي القرن الخامس عشر!$** تحت تأثير ثقافي كبير من ألمانيا. أراد بعض الرهبان الألمان تقريب السكان الأصليين من الآله، لذلك اخترعوا اللغة الحرفية الإستونية. ولقد استند إلى الأبجدية الألمانية وأضيف حرف واحد هو ""Õ/õ"". مع مرور الزمن، العديد من الكلمات التي أُخِذَت من اللغة الألمانية قد اندمجت. كانت هذه بداية حركة التنوير.
(j)	Restricted Answer	أين أعلنت كريستينا فرنانديز دي كيرشنر عن نيتها الترشح؟ * في مسرح على بعد 31 ميلاً من لابلاتا * في مسرح بوينس آيرس في لابلاتا * **في المسرح الأرجنتيني على بعد 31 ميلاً من بوينس آيرس** * في مسرح لابلاتا في بوينس آيرس	لقد أعلنت كريستينا فرنانديز دي كيرشنر عضوة مجلس الشيوخ الحالية وسيدة الأرجنتين الأولى ترشحها للرئاسة الأمس مساءً **$!في لابلاتا!$**، وهي مدينة تبعد 50 كيلومتراً (أو 31 ميلاً) عن بوينس أيرس أعلنت السيدة كيرشنر عزمها الترشح للرئاسة في المسرح الأرجنتيني، وهو نفس الموقع الذي استخدمته لبدأ حملتها الانتخابية عام ٢٠٠٥ لمجلس الشيوخ بصفتها عضواً في وفد مقاطعة بوينس أيرس.
(k)	X True/False	وفقاً للفقرة، أي جملة عن المشتبه به غير صحيحة؟ * استخدم سيارة أثناء الهجوم * زُعم أنه فجر عبوة ناسفة * السلطات على دراية بعرقه * **لم يصب بأذى**	زُعم أن الرجل قاد سيارة ذات ثلاث عجلات ملغمة بالمتفجرات وسط حشد من الناس. تم اعتقال الرجل المشتبه في أنه قام بتفجير القنبلة بعد إصابته بجروح جراء الانفجار. ولم تعرف السلطات اسمه حتى الآن، بالرغم من أنهم يعرفون أنه من مجموعة الإيغور العرقية.
(l)	X Entailment	في أي بلد درس ما القانون؟ * **الولايات المتحدة الأمريكية** * الصين * أستراليا * هونغ كونغ	وُلد "ما" في هونغ كونغ، ودرس في جامعة نيويورك وكلية هارفارد للحقوق، وحصل ذات مرة على "البطاقة الخضراء" كمقيم دائم في الولايات المتحدة. هسيه أشار أثناء الانتخابات إلى أن ما قد يهرب من البلاد في وقت المصائب. هسيه جادل أيضاً أن ما الجذاب فوتوغرافياً كان يهتم بالشكل أكثر من الجوهر. [...]

Table 4: This table shows examples in Arabic demonstrating part of the range of types of MCQs in BELEBELE, and the respective issues they pose for extractive QA in this work. These examples are parallel to the English examples in Table 1.

D Guidelines for English Answer Span Annotation

Belebele Answer Span: English Annotation Guidelines

You'll be provided with a context paragraph and a question related to that paragraph. Your task is to find the *answer span* from the given paragraph where either the exact answer or the evidence of the answer can be found. The span should be as **concise as possible.**

The original Belebele multi-choice answer has been hidden in your file. However, the examples provided below include the multi-choice answer simply to illustrate the differences in the nature of the original answers and the "answer span" - which may help explain some annotation guideline decisions.

****

Please follow the annotation guidelines below carefully.

****

1. Identify the span of text in which the answer to the question or evidence of the answer can be found (Answer Span column)
2. Mark the start of the span with **$!** and the end of the span with **!$**
3. The answer span needs to be as short as possible.
4. Finding a relevant span is more important than finding an exact answer.
5. Don't include non-relevant punctuation.
6. Include the articles "the", "a" and "an".
7. Include the preposition if the question is a When/ Where/ How question.
8. In cases of multiple instances of the answer string, choose the appropriate one only.
9. Mark problematic questions with BB or X.

Examples expanding on guidelines 3-9 are presented next.

3. The answer span needs to be as short as possible.

3 (a) Example of correct annotation:

Paragraph	Question	Multi-choice Answer	Answer Span
The American plan relied on launching coordinated attacks from three different directions. General John Cadwalder would launch a diversionary attack against the British garrison at Bordentown, in order to block off any reinforcements. General James Ewing would take 700 militia across the river at Trenton Ferry, seize the bridge over the Assunpink Creek and prevent any enemy troops from escaping.	Where was there a British garrison located?	Bordentown	The American plan relied on launching coordinated attacks from three different directions. General John Cadwalder would launch a diversionary attack against the British garrison **$!**at Bordentown**!$**, in order to block off any reinforcements. General James Ewing would take 700 militia across the river at Trenton Ferry, seize the bridge over the Assunpink Creek and prevent any enemy troops from escaping.

3 (b) Example of incorrect annotation: (Note: You don't need to add more context to the span)

Paragraph	Question	Multi-choice Answer	Answer Span
Twentieth century research has shown that there are two pools of genetic variation: hidden and expressed. Mutation adds new genetic variation, and selection removes it from the pool of expressed variation. Segregation and recombination shuffle variation back and forth between the two pools with each generation.	Which process is responsible for adding genetic variation?	Mutation	Twentieth century research has shown that there are two pools of genetic variation: hidden and expressed. $!Mutation adds new genetic variation!$, and selection removes it from the pool of expressed variation. Segregation and recombination shuffle variation back and forth between the two pools with each generation.

3 (b) In some cases, the answer can't be found in a concise form in the paragraph, thus requiring a longer span. Find the answer span where one can find the answer to the question, including the additional information that it may entail. Example of correct annotation where longer span is necessary:

Paragraph	**Question**	**Multi-choice Answer**	**Answer Span**
This is called a chemical's pH. You can make an indicator using red cabbage juice. The cabbage juice changes color depending on how acidic or basic (alkaline) the chemical is. The pH level is indicated by the amount of Hydrogen (the H in pH) ions in the tested chemical.	How is the pH level of a chemical measured?	The amount of Hydrogen ions in the chemical	This is called a chemical's pH. You can make an indicator using red cabbage juice. The cabbage juice changes color depending on how acidic or basic (alkaline) the chemical is. The pH level is indicated **$!**by the amount of Hydrogen (the H in pH) ions in the tested chemical**!$**.

4. Finding a relevant span is more important than finding an exact answer.

4 (a) In some cases, there is no exact answer in the paragraph, and therefore reasoning or deduction would be required to give a fully comprehensive answer. Yet keep in mind that the task is finding the span which points to *evidence for the answer*. In these cases, simply identify the span of text where an answer can be found. In the example below, the referent "this" can later be resolved by the reader by reading the earlier text in the passage (movement of men and materials).

Paragraph	**Question**	**Multi-choice Answer**	**Answer Span**
Using ships to transport goods is by far the most efficient way to move large amounts of people and goods across oceans. The job of navies has traditionally been to ensure that your country maintains the ability to move your people and goods, while at the same time, interfering with your enemy's ability to move his people and goods. One of the most noteworthy recent examples of this was the North Atlantic campaign of WWII. The Americans were trying to move men and materials across the Atlantic Ocean to help Britain. At the same time, the German navy, using mainly U-boats, was trying to stop this traffic. Had the Allies failed, Germany probably would have been able to conquer Britain as it had the rest of Europe.	What was the German navy trying to accomplish during WWII?	Preventing Britain from receiving people and goods	Using ships to transport goods is by far the most efficient way to move large amounts of people and goods across oceans. The job of navies has traditionally been to ensure that your country maintains the ability to move your people and goods, while at the same time, interfering with your enemy's ability to move his people and goods. One of the most noteworthy recent examples of this was the North Atlantic campaign of WWII. The Americans were trying to move men and materials across the Atlantic Ocean to help Britain. At the same time, the German navy, using mainly U-boats, was trying to $!stop this traffic!$. Had the Allies failed, Germany probably would have been able to conquer Britain as it had the rest of Europe.

5. Don't include non-relevant punctuation

5 (a) Final punctuation is usually not considered relevant

Paragraph	**Question**	**Multi-choice Answer**	**Answer Span**
Things did not go well for the Italians in North Africa almost from the start. Within a week of Italy's declaration of war on June 10, 1940, the British 11th Hussars had seized Fort Capuzzo in Libya. In an ambush east of Bardia, the British captured the Italian Tenth Army's Engineer-in-Chief, General Lastucci. On June 28, Marshal Italo Balbo, the Governor-General of Libya and apparent heir to Mussolini, was killed by friendly fire while landing in Tobruk.	Where was Italo Balbo killed?	Tobruk	Things did not go well for the Italians in North Africa almost from the start. Within a week of Italy's declaration of war on June 10, 1940, the British 11th Hussars had seized Fort Capuzzo in Libya. In an ambush east of Bardia, the British captured the Italian Tenth Army's Engineer-in-Chief, General Lastucci. On June 28, Marshal Italo Balbo, the Governor-General of Libya and apparent heir to Mussolini, was killed by friendly fire while landing $!in Tobruk!$.

5 (b) Closing punctuation is more likely to be relevant. ( ) “ “ [ ]
The example below includes the closing bracket in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
MRI is based on a physics phenomenon called nuclear magnetic resonance (NMR), which was discovered in the 1930s by Felix Bloch (working at Stanford University) and Edward Purcell (from Harvard University). In this resonance, magnetic field and radio waves cause atoms to give off tiny radio signals. In the year 1970, Raymond Damadian, a medical doctor and research scientist, discovered the basis for using magnetic resonance imaging as a tool for medical diagnosis. Four years later a patent was granted, which was the world's first patent issued in the field of MRI. In 1977, Dr. Damadian completed the construction of the first “whole-body” MRI scanner, which he called the ”Indomitable”.	Who discovered nuclear magnetic resonance?	Bloch and Purcell	MRI is based on a physics phenomenon called nuclear magnetic resonance (NMR), which was discovered in the 1930s by $!Felix Bloch (working at Stanford University) and Edward Purcell (from Harvard University)!$. In this resonance, magnetic field and radio waves cause atoms to give off tiny radio signals. In the year 1970, Raymond Damadian, a medical doctor and research scientist, discovered the basis for using magnetic resonance imaging as a tool for medical diagnosis. Four years later a patent was granted, which was the world's first patent issued in the field of MRI. In 1977, Dr. Damadian completed the construction of the first “whole-body” MRI scanner, which he called the ”Indomitable”.

6. In order to maintain consistency across annotators, the span will include the articles “the”, “a” and “an”.

6 (a) An example of the inclusion of the article “the” in a span:

Paragraph	Question	Multi-choice Answer	Answer Span
Golf is a game in which players use clubs to hit balls into holes. Eighteen holes are played during a regular round, with players usually starting on the first hole on the course and finishing on the eighteenth. The player who takes the fewest strokes, or swings of the club, to complete the course wins. The game is played on grass, and the grass around the hole is mown shorter and called the green.	On a golf course, where is the grass cut shorter?	On the green	Golf is a game in which players use clubs to hit balls into holes. Eighteen holes are played during a regular round, with players usually starting on the first hole on the course and finishing on the eighteenth. The player who takes the fewest strokes, or swings of the club, to complete the course wins. The game is played on grass, and the grass around the hole is mown shorter and called $!the green!$.

6 (b) Example of inclusion of the article “a” in a span:

Paragraph	Question	Multi-choice Answer	Answer Span
The atom can be considered to be one of the fundamental building blocks of all matter. Its a very complex entity which consists, according to a simplified Bohr model, of a central nucleus orbited by electrons, somewhat similar to planets orbiting the sun - see Figure 1.1. The nucleus consists of two particles - neutrons and protons. Protons have a positive electric charge while neutrons have no charge. The electrons have a negative electric charge.	The particles that orbit the nucleus have which type of charge?	Negative charge	The atom can be considered to be one of the fundamental building blocks of all matter. Its a very complex entity which consists, according to a simplified Bohr model, of a central nucleus orbited by electrons, somewhat similar to planets orbiting the sun - see Figure 1.1. The nucleus consists of two particles - neutrons and protons. Protons have a positive electric charge while neutrons have no charge. The electrons have $!a negative electric charge!$.

7. If the question is a When/ Where/ How question – then include the preposition or adverb that would normally form part of an appropriate answer.

7 (a) An example of how-question where the preposition should be included in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
Subcultures bring together like-minded individuals who feel neglected by societal standards and allow them to develop a sense of identity. Subcultures can be distinctive because of the age, ethnicity, class, location, and/or gender of the members. The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors. Members of a subculture often signal their membership through a distinctive and symbolic use of style, which includes fashions, mannerisms, and argot.	How do members of a particular subculture often signify their association with the group?	By using style as a form of symbolism	Subcultures bring together like-minded individuals who feel neglected by societal standards and allow them to develop a sense of identity. Subcultures can be distinctive because of the age, ethnicity, class, location, and/or gender of the members. The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors. Members of a subculture often signal their membership $!through a distinctive and symbolic use of style!$, which includes fashions, mannerisms, and argot.

7 (b) An example of a when-question where the preposition should be included in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
Gothic style peaked in the period between the 10th - 11th centuries and the 14th century. At the beginning dress was heavily influenced by the Byzantine culture in the east. However, due to the slow communication channels, styles in the west could lag behind by 25 to 30 year. Towards the end of the Middle Ages western Europe began to develop their own style. one of the biggest developments of the time as a result of the crusades people began to use buttons to fasten clothing.	When did western Europe stop relying heavily on influences and start developing its own style?	Around the end of the Middle Ages	Gothic style peaked in the period between the 10th - 11th centuries and the 14th century. At the beginning dress was heavily influenced by the Byzantine culture in the east. However, due to the slow communication channels, styles in the west could lag behind by 25 to 30 year. $!towards the end of the Middle Ages!$ western Europe began to develop their own style. one of the biggest developments of the time as a result of the crusades people began to use buttons to fasten clothing.

7 (c) An example of a how-question case where the preposition should be included in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
The invention of spoke wheels made Assyrian chariots lighter, faster, and better prepared to outrun soldiers and other chariots. Arrows from their deadly crossbows could penetrate the armor of rival soldiers. About 1000 B.C., the Assyrians introduced the first cavalry. A cavalry is an army that fights on horseback. The saddle had not yet been invented, so the Assyrian cavalry fought on the bare backs of their horses.	How are battles that utilize a calvary fought?	on horseback	The invention of spoke wheels made Assyrian chariots lighter, faster, and better prepared to outrun soldiers and other chariots. Arrows from their deadly crossbows could penetrate the armor of rival soldiers. About 1000 B.C., the Assyrians introduced the first cavalry. A cavalry is an army that fights $!**on horseback**!$. The saddle had not yet been invented, so the Assyrian cavalry fought on the bare backs of their horses.

8. In some cases, the exact original answer string can be found multiple times in the paragraph. Choose the appropriate one only.

Paragraph	Question	Multi-choice Answer	Answer Span
Eighteen percent of Venezuelans are unemployed, and most of those who are employed work in the informal economy. Two thirds of Venezuelans who work do so in the service sector, nearly a quarter work in industry and a fifth work in agriculture. An important industry for Venezuelans is oil, where the country is a net exporter, even though only one percent work in the oil industry.	Which line of work mentioned in the passage employs the least number of Venezuelans?	Oil	Eighteen percent of Venezuelans are unemployed, and most of those who are employed work in the informal economy. Two thirds of Venezuelans who work do so in the service sector, nearly a quarter work in industry and a fifth work in agriculture. An important industry for Venezuelans is oil, where the country is a net exporter, even though only one percent work in the $!oil!$ industry.

9. The Belebele dataset was initially curated as a multiple-choice reading comprehension dataset. As a result, the questions were not crafted in a way that is suitable for extractive QA, and as such it may not be possible to find an exact answer or *any* answer:

9 (a) You may come across some answers that cannot be found simply in a span of text i.e. reasoning or deduction is needed through reading the rest of the passage. In these cases, mark the span where evidence for the answer can be found. Then mark it as BB to indicate that it's problematic due to the nature of the Belebele dataset.

In some cases, you may need to unhide the original answer (column H) to help you decide whether or not it should be labelled it as BB.

Example of BB case:

Paragraph	Question	Original Answer	Multi-choice Answer	Belebele Problem?
The balance of power was a system in which European nations sought to maintain the national sovereignty of all European states. The concept was that all European nations had to seek to prevent one nation from becoming powerful, and thus national governments often changed their alliances in order to maintain the balance. The War of Spanish Succession marked the first war whose central issue was the balance of power. This marked an important change, as European powers would no longer have the pretext of being religious wars. Thus, the Thirty Years' War would be the last war to be labeled a religious war.	In which country was the first war in Europe whose central issue was said to relate to the balance of power rather than having a religious context?	Spain	The balance of power was a system in which European nations sought to maintain the national sovereignty of all European states. The concept was that all European nations had to seek to prevent one nation from becoming powerful, and thus national governments often changed their alliances in order to maintain the balance. The War of Spanish Succession marked the first war whose central issue was the balance of power. This marked an important change, as European powers would no longer have the pretext of being religious wars. Thus, the Thirty Years' War would be the last war to be labeled a religious war.	BB

9 (b) You may also be asked a true or false question. These are also not useful for our QA dataset. Remove the passage text and mark with X.

Paragraph	**Question**	**Multi-choice Answer**	**Answer Span**
Virtual teams are held to the same standards of excellence as conventional teams, but there are subtle differences. Virtual team members often function as the point of contact for their immediate physical group. They often have more autonomy than conventional team members as their teams may meet according to varying time zones which may not be understood by their local management. The presence of a true “invisible team” (Larson and LaFasto, 1989, p109) is also a unique component of a virtual team. The “invisible team” is the management team to which each of the members report. The invisible team sets the standards for each member.	Based on the passage, which statement regarding physical and virtual teams is not true?	Conventional teams are usually held to a higher standard	X

9 (c) In some cases, inference or calculation is required to establish the answer to the question. In other words, there is no particular span in which the answer can be found. Remove the passage text and mark as X.

Paragraph	**Question**	**Multi-choice Answer**	**Answer Span**
MRI is based on a physics phenomenon called nuclear magnetic resonance (NMR), which was discovered in the 1930s by Felix Bloch (working at Stanford University) and Edward Purcell (from Harvard University). In this resonance, magnetic field and radio waves cause atoms to give off tiny radio signals. In the year 1970, Raymond Damadian, a medical doctor and research scientist, discovered the basis for using magnetic resonance imaging as a tool for medical diagnosis. Four years later a patent was granted, which was the world's first patent issued in the field of MRI. In 1977, Dr. Damadian completed the construction of the first “whole-body” MRI scanner, which he called the ”Indomitable”.	In what year was the first patent granted for medical imaging resonance?	1974	X

E Guidelines for Modern Standard Arabic Answer Span Annotation

Belebele Answer Span: Arabic Annotation Guidelines

You'll be provided with a context paragraph and a question related to that paragraph. Your task is review and validate the pre-annotated *answer span* in the given paragraph. You should edit the span where necessary. The span indicates either the exact answer or where evidence of the answer can be found. The span should be as **concise as possible.**

The original Belebele multi-choice answer has been hidden in your file. However, the examples provided below include the multi-choice answer simply to illustrate the differences in the nature of the original answers and the "answer span" - which may help explain some annotation guideline decisions.

****

Please follow the annotation guidelines below carefully.

****

1. The span should identify the part of the text in which the answer to the question or evidence of the answer can be found
2. Where necessary, edit the start of the span with **$!** and the end of the span with **!$** . (Answer Span column V)
3. The answer span needs to be as short as possible.
4. Finding a relevant span is more important than finding an exact answer.
5. Don't include non-relevant punctuation.
6. Include the preposition if the question is a When/ Where/ How question.
7. In cases of multiple instances of the answer string, choose the appropriate one only.
8. Mark problematic questions with BB or X.

<u>Examples expanding on guidelines 3-8 are presented next.</u>

3. The answer span needs to be as short as possible.

3 (a) Example of correct annotation:

Paragraph	Question	Multi-choice Answer	Answer Span
تنتج النباتات الأكسجين الذي يتنفسه البشر، ويمتصون ثاني أكسيد الكربون الذي يخلفه البشر (أي الزفير). تصنع النباتات طعامها من الشمس من خلال التمثيل الضوئي. كما يوفرون الظل. نصنع بيوتنا وملابسنا من النباتات . معظم الطعام الذي نتناوله من النباتات كذلك. بدون النباتات، لا يمكن للحيوانات البقاء.	ما الذي يساعد النباتات في عملية التمثيل الضوئي؟	الشمس	تنتج النباتات الأكسجين الذي يتنفسه البشر، ويمتصون ثاني أكسيد الكربون الذي يخلفه البشر (أي الزفير). تصنع النباتات طعامها من **$!الشمس!$** من خلال التمثيل الضوئي. كما يوفرون الظل. نصنع بيوتنا وملابسنا من النباتات . معظم الطعام الذي نتناوله من النباتات كذلك. بدون النباتات، لا يمكن للحيوانات البقاء.

3 (b) Example of incorrect annotation: (Note: You don't need to add more context to the span)

Paragraph	Question	Multi-choice Answer	Answer Span
العضو الرئيسي للجهاز الدوري هو القلب الذي يضخ الدم. يتحرك الدم بعيداً عن القلب في أنابيب تسمى الشرايين ويعود إلى القلب في أنابيب تسمى الأوردة. أصغر الأنابيب تسمى الشعيرات الدموية.	ما هي الأنابيب الأصغر في جهاز الدورة الدموية؟	الشعيرات الدموية	العضو الرئيسي للجهاز الدوري هو القلب الذي يضخ الدم. يتحرك الدم بعيداً عن القلب في أنابيب تسمى الشرايين ويعود إلى القلب في أنابيب تسمى الأوردة. أصغر الأنابيب تسمى **$!الشعيرات الدموية.!$**

3 (c) In some cases, the answer can't be found in a concise form in the paragraph, thus requiring a longer span. Find the answer span where one can find the answer to the question, including the additional information that it may entail. Example of correct annotation where longer span is necessary:

Paragraph	Question	Multi-choice Answer	Answer Span
ضرب زلزال متوسط القوة مونتانا الغربية الساعة ١٠:٠٨ مساءً. يوم الاثنين. ولم ترد تقارير فورية عن وقوع أضرار من قبل هيئة المسح الجيولوجي الأمريكي (USGS) ومركز معلومات الزلازل الوطني التابع لها. وكان مركز الزلزال على بعد حوالي 20 كم (15 ميلاً) بين شمال وشمال شرق ديلون، وحوالي 65 كم (40 ميلاً) جنوب بوتي.	أين تمركز الزلزال بالنسبة لبوتي؟	40ميلا جنوباً	ضرب زلزال متوسط القوة مونتانا الغربية الساعة ١٠:٠٨ مساءً. يوم الاثنين. ولم ترد تقارير فورية عن وقوع أضرار من قبل هيئة المسح الجيولوجي الأمريكي (USGS) ومركز معلومات الزلازل الوطني التابع لها. وكان مركز الزلزال على بعد حوالي 20 كم (15 ميلاً) بين شمال وشمال شرق ديلون، **$!وحوالي 65 كم (40 ميلاً) جنوب بوتي!$.**

4. Finding a relevant span is more important than finding an exact answer.

4 (a) In some cases, there is no exact answer in the paragraph, and therefore reasoning or deduction would be required to give a fully comprehensive answer. Yet keep in mind that the task is finding the span which points to *evidence for the answer*. In these cases, simply identify the span of text where an answer can be found. In the example below, the referent "this" can later be resolved by the reader by reading the earlier text in the passage (movement of men and materials).

Paragraph	Question	Multi-choice Answer	Answer Span
قد يكون للطفرات مجموعة متنوعة من الآثار المختلفة، على حسب نوع الطفرة، وأهميّة قطعة المادّة الوراثيّة المتأثّرة، وما إذا كانت الخلايا المتأثرة من خلايا الخط الإنتاشي. يمكن فقط نقل الطفرات في خلايا الخط الجرثومي إلى الأطفال، بينما يمكن أن تسبب الطفرات في أماكن أخرى موت الخلايا أو السرطان.	على ماذا تعتمد قدرة الطفرة في الانتقال إلى الذرية؟	إذا كانت الخلايا من خلايا الخط الإنتاشي	قد يكون للطفرات مجموعة متنوعة من الآثار المختلفة، على حسب نوع الطفرة، وأهميّة قطعة المادّة الوراثيّة المتأثّرة، وما إذا كانت الخلايا المتأثرة من خلايا الخط الإنتاشي. **$!يمكن فقط نقل الطفرات في خلايا الخط الجرثومي إلى الأطفال!$**، بينما يمكن أن تسبب الطفرات في أماكن أخرى موت الخلايا أو السرطان.

5. Don't include non-relevant punctuation

5 (a) Final punctuation is usually not considered relevant

Paragraph	Question	Multi-choice Answer	Answer Span
العضو الرئيسي للجهاز الدوري هو القلب الذي يضخ الدم. يتحرك الدم بعيداً عن القلب في أنابيب تسمى الشرايين ويعود إلى القلب في أنابيب تسمى الأوردة. أصغر الأنابيب تسمى الشعيرات الدموية.	أي جزء من الدورة الدموية يجلب الدم نحو القلب؟	الأوردة	العضو الرئيسي للجهاز الدوري هو القلب الذي يضخ الدم. يتحرك الدم بعيداً عن القلب في أنابيب تسمى الشرايين ويعود إلى القلب في أنابيب تسمى **$!الأوردة!$**. أصغر الأنابيب تسمى الشعيرات الدموية.

5 (b) Closing punctuation is more likely to be relevant. ( ) “ “ [ ]
The example below includes the closing bracket in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
"يعتمد التصوير بالرنين المغناطيسي على ظاهرة فيزيائية تسمى الرنين المغناطيسي النووي ،(NMR) والتي اكتشفت في ثلاثينيات القرن الماضي عن طريق فيليكس بلوخ (من جامعة ستانفورد) وإدوارد بورسيل (من جامعة هارفارد) في هذا الرنين، تجعل المجالات المغناطيسيّة والموجات الراديويّة الذرّات تبعث إشارات راديو دقيقة. في عام 1970، اكتشف رايموند داماديان، وهو طبيب وعالم أبحاث، الأساس الذي بُني عليه استخدام التصوير بالرنين المغناطيسي كأداة للتشخيص الطبّي. بعد أربع سنوات، تم منح براءة اختراع، والتي كانت أول براءة اختراع في العالم يتم منحها في مجال التصوير بالرنين المغناطيسي. في عام 1977 استكمل الدكتور داماديان إنشاء أول ماسح ضوئي للتصوير بالرنين المغناطيسي "لكامل الجسم"، والذي يسمى اليوم ب"الذي لا يقهر"."	من اكتشف الرنين المغناطيسي النووي؟	بلوخ وبورسيل	"يعتمد التصوير بالرنين المغناطيسي على ظاهرة فيزيائية تسمى الرنين المغناطيسي النووي ،(NMR) والتي اكتشفت في ثلاثينيات القرن الماضي عن طريق **$!فيليكس بلوخ (من جامعة ستانفورد) وإدوارد بورسيل (من جامعة هارفارد)!$** في هذا الرنين، تجعل المجالات المغناطيسيّة والموجات الراديويّة الذرّات تبعث إشارات راديو دقيقة. في عام 1970، اكتشف رايموند داماديان، وهو طبيب وعالم أبحاث، الأساس الذي بُني عليه استخدام التصوير بالرنين المغناطيسي كأداة للتشخيص الطبّي. بعد أربع سنوات، تم منح براءة اختراع، والتي كانت أول براءة اختراع في العالم يتم منحها في مجال التصوير بالرنين المغناطيسي. في عام 1977 استكمل الدكتور داماديان إنشاء أول ماسح ضوئي للتصوير بالرنين المغناطيسي "لكامل الجسم"، والذي يسمى اليوم ب"الذي لا يقهر"."

6. If the question is a When/ How question – then include the preposition or adverb that would normally form part of an appropriate answer. Note that sometimes the preposition is already part of the word in the form of a morpheme.

6 (a) An example of a “how-question” where the preposition should be included in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
ستطاع اختراع العجلات ذات العوارض أن يجعل العربات الأشورية أخف وزناً وأسرع وأكثر استعداداً لتجاوز الجنود والعربات الأخرى. يمكن أن تخترق السهام المُطلقة من الأقواس القاتلة دروع الجنود المتنافسين. قدم الأشوريون أول سلاح فرسان في حوالي عام 1000 قبل الميلاد. سلاح الفرسان هو جيش يقاتل على ظهور الخيل. لم يكن السرج قد اخترع بعد، لذلك قاتل الفرسان الأشوريون على ظهور خيولهم العارية.	**كيف** كانت تُخاض المعارك التي تستخدم الفرسان؟	على ظهر الخيل	استطاع اختراع العجلات ذات العوارض أن يجعل العربات الأشورية أخف وزناً وأسرع وأكثر استعداداً لتجاوز الجنود والعربات الأخرى. يمكن أن تخترق السهام المُطلقة من الأقواس القاتلة دروع الجنود المتنافسين. قدم الأشوريون أول سلاح فرسان في حوالي عام 1000 قبل الميلاد. سلاح الفرسان هو جيش يقاتل **$!على ظهور الخيل!$**. لم يكن السرج قد اخترع بعد، لذلك قاتل الفرسان الأشوريون على ظهور خيولهم العارية..

6 (b) An example of a when-question where the preposition should be included in the span.

Paragraph	Question	Multi-choice Answer	Answer Span
بلغ الطراز القوطي ذروته في فترة ما بين القرنين العاشر أو الحادي عشر إلى القرن الرابع عشر. في البداية، أثرت الثقافة البيزنطية في الشرق على الملابس تأثيرًا شديدًا. وعلى الرغم من ذلك، فنظرا لبطء قنوات التواصل، قد تتأخر الأنماط في الغرب قدر ما يقارب 25 إلى 30 سنة. في نهاية العصور الوسطى، بدأت أوروبا الغربية في تطوير أسلوبها الخاص. واحدة من أكبر التطورات في ذلك الوقت نتيجة للحروب الصليبية بدأ الناس في استخدام الأزرار لربط الملابس.	**متى** توقفت أوروبا الغربية عن الاعتماد الشديد على التأثيرات وبدأت في تطوير أسلوبها الخاص؟	في نهاية العصور الوسطى	بلغ الطراز القوطي ذروته في فترة ما بين القرنين العاشر أو الحادي عشر إلى القرن الرابع عشر. في البداية، أثرت الثقافة البيزنطية في الشرق على الملابس تأثيرًا شديدًا. وعلى الرغم من ذلك، فنظرا لبطء قنوات التواصل، قد تتأخر الأنماط في الغرب قدر ما يقارب 25 إلى 30 سنة. **$!في نهاية العصور الوسطى!$**، بدأت أوروبا الغربية في تطوير أسلوبها الخاص. واحدة من أكبر التطورات في ذلك الوقت نتيجة للحروب الصليبية بدأ الناس في استخدام الأزرار لربط الملابس

7. In some cases, the exact original answer string can be found multiple times in the paragraph. Choose the appropriate one only.

Paragraph	Question	Multi-choice Answer	Answer Span
"أفادت تقارير باستمرار أعمال النهب على نطاق واسع حتى صباح اليوم التالي وذلك لغياب ضباط إنفاذ القانون عن شوارع مدينة "بيشكك". وصف أحد المراقبين مدينة بيشكيك بأنها تغرق في حالة من "فقدان السيطرة"، حيث تجولت عصابات من الناس في الطرقات وقاموا بنهب مخازن السلع الاستهلاكية. حمّل العديد من سكان مدينة "بشكيك" المتظاهرين في الجنوب حالة الفوضى."	من ألقى باللوم على المتظاهرين من الجنوب في أعمال النهب؟	سكان بيشكك	"أفادت تقارير باستمرار أعمال النهب على نطاق واسع حتى صباح اليوم التالي وذلك لغياب ضباط إنفاذ القانون عن شوارع مدينة "بيشكك". وصف أحد المراقبين مدينة بيشكيك بأنها تغرق في حالة من "فقدان السيطرة"، حيث تجولت عصابات من الناس في الطرقات وقاموا بنهب مخازن السلع الاستهلاكية. حمّل العديد من **$!سكان مدينة "بشكيك"!$** المتظاهرين في الجنوب حالة الفوضى."

8. The Belebele dataset was initially curated as a multiple-choice reading comprehension dataset. As a result, the questions were not crafted in a way that is suitable for QA, and as such it may not be possible to find an exact answer or *any* answer:

8 (a) You may come across some answers that cannot be found simply in a span of text i.e. reasoning or deduction is needed through reading the rest of the passage. In these cases, mark the span where evidence for the answer can be found. Then mark it as BB to indicate that it's problematic due to the nature of the Belebele dataset.

Example of BB case:

Paragraph	Question	Original Answer	Answer Span	Belebele Problem?
كانت الموازنة في الطاقة نظاماً حاولت من خلاله الدول الأوروبية أن تُحافظ على السيادة الوطنية الخاصة بجميع الدول الأوروبية. نص المفهوم على أن جميع الدول الأوروبية يجب أن تسعى لمنع دولة واحدة من أن تصبح قوية، وبالتالي عادةً ما غيرت الحكومات الوطنية تحالفاتها للحفاظ على التوازن. كانت حرب الخلافة الإسبانية هي الحرب الأولى التي كانت قضيتها المركزية هي ميزان القوى. كان ذلك بمثابة تغيير مهم، حيث لم يعد لدى القوى الأوروبية ذريعة لكونها حروباً دينية. وهكذا، ستكون حرب الثلاثين عاماً آخر حرب توصف بأنها حرب دينية.	في أي بلد حصلت الحرب الأولى في أوروبا والتي كانت قضيتها المركزية مرتبطة بتوازن القوى بدلاً من أن تكون في سياق ديني؟	إسبانيا	كانت الموازنة في الطاقة نظاماً حاولت من خلاله الدول الأوروبية أن تُحافظ على السيادة الوطنية الخاصة بجميع الدول الأوروبية. نص المفهوم على أن جميع الدول الأوروبية يجب أن تسعى لمنع دولة واحدة من أن تصبح قوية، وبالتالي عادةً ما غيرت الحكومات الوطنية تحالفاتها للحفاظ على التوازن. كانت **$!حرب الخلافة الإسبانية!$** هي الحرب الأولى التي كانت قضيتها المركزية هي ميزان القوى. كان ذلك بمثابة تغيير مهم، حيث لم يعد لدى القوى الأوروبية ذريعة لكونها حروباً دينية. وهكذا، ستكون حرب الثلاثين عاماً آخر حرب توصف بأنها حرب دينية	BB

8 (b) Questions for which an answer (or evidence for an answer) could not be found at all have been marked as X (as per the English annotations)

Example of X case:

Paragraph	Question	Original Answer	Answer Span	Belebele Problem?
كانت الموازنة في الطاقة"يعد الهرم الأكبر في محافظة "الجيزة" هو الوحيد من العجائب السبع التي لا تزال قائمة حتى يومنا هذا. تم بناؤه من قبل المصريين في القرن الثالث من الحقبة الحالية، يعتبر الهرم الأعظم واحداً من الأهرام الضخمة المبنية على شرف الفرعون الميت. تحتوي هضبة الجيزة، أو "مدينة الجيزة الجنائزية" بوادي الموتى المصري، على عدة أهرامات (أكبرها الهرم الأكبر) ومقابر صغيرة ومعابد وأبو الهول. تم إنشاء الهرم الأكبر لتكريم الفرعون خوفو، وتم بناء العديد من الأهرامات والمقابر والمعابد الأصغر لتكريم زوجات خوفو وأفراد أسرته."	ماذا يمكن العثور عليه في هضبة الجيزة؟	جميع العجائب السبع	كانت الموازنة في الطاقة"يعد الهرم الأكبر في محافظة "الجيزة" هو الوحيد من العجائب السبع التي لا تزال قائمة حتى يومنا هذا. تم بناؤه من قبل المصريين في القرن الثالث من الحقبة الحالية، يعتبر الهرم الأعظم واحداً من الأهرام الضخمة المبنية على شرف الفرعون الميت. تحتوي هضبة الجيزة، أو "مدينة الجيزة الجنائزية" بوادي الموتى المصري، على عدة أهرامات (أكبرها الهرم الأكبر) ومقابر صغيرة ومعابد وأبو الهول. تم إنشاء الهرم الأكبر لتكريم الفرعون خوفو، وتم بناء العديد من الأهرامات والمقابر والمعابد الأصغر لتكريم زوجات خوفو وأفراد أسرته."	X