ArTrivia: Harvesting Arabic Wikipedia to Build A New Arabic Question Answering Dataset

Sultan Alrowili
Department of Computer Science
University of Delaware
Newark, Delaware, USA
alrowili@udel.edu

K.Vijay-Shanker
Department of Computer Science
University of Delaware
Newark, Delaware, USA
vijay@udel.edu

Abstract

We present ArTrivia, a new Arabic question-answering dataset consisting of more than 10,000 question-answer pairs along with relevant passages, covering a wide range of 18 diverse topics in Arabic. We created our dataset using a newly proposed pipeline that leverages diverse structured data sources from Arabic Wikipedia. Moreover, we conducted a comprehensive statistical analysis of ArTrivia and assessed the performance of each component in our pipeline. Additionally, we compared the performance of ArTrivia against the existing TyDi QA dataset using various experimental setups. Our analysis highlights the significance of often overlooked aspects in dataset creation, such as answer normalization, in enhancing the quality of QA datasets. Our evaluation also shows that ArTrivia presents more challenging and out-of-distribution questions to TyDi, raising questions about the feasibility of using ArTrivia as a complementary dataset to TyDi.

1 Introduction

In recent years, the field of question-answering (QA) in Arabic NLP has witnessed more attention with the introduction of several Arabic QA datasets, such as TyDi QA (Clark et al., 2020), the Arabic Reading Comprehension Dataset (ARCD) (Mozannar et al., 2019), and the Arabic Question-Answer Dataset (AQAD) (Atef et al., 2020). However, existing Arabic QA datasets have several issues, such as having limited topic diversity, picking common question patterns, and the limited size of the dataset.

First, although having a variety of topics is one of the objectives of TyDi QA creators (Clark et al., 2020) [1], many subjects such as classical Arabic poetry are less represented in TyDi. This issue also exists in both ARCD and AQAD datasets since they are generated from a limited number of articles.

Second, when crowd workers are given passages and asked to formulate questions with less defined guidelines, they tend to pick common patterns, which can compromise the quality of the dataset. Our analysis reveals that approximately 33% of the questions in Arabic TyDi are about explaining entities such as "Who is Alfred Nobel?" or "What is graphic design?" [2]. This is in contrast to both ARCD and English TyDi QA, where such questions consist of only about 4-5% of the dataset.

Third, when compared to the SQuAD dataset (Rajpurkar et al., 2016), which consists of 120,000 examples, Arabic QA datasets still have a limited dataset size, mainly due to the constraints imposed by the cost of crowd-sourcing. However, few studies in Arabic NLP explore alternative approaches to the crowd-sourcing method. Most of these approaches rely on Machine Translation, a method that has been criticized for it is poor performance in Arabic Question Answering (Antoun et al., 2021).

One suggested solution to address the existing challenges in Arabic Question Answering is to utilize Large Language Models (LLMs). These LLM models often use a zero-shot learning technique to address QA tasks, eliminating the need to fine-tune the Language Model on a specific question answering dataset. The zero-shot approach with LLMs in English QA tasks has shown promising results, often matching the supervised methods that require a finetuning dataset (Lai et al., 2023). However, recent studies in Arabic NLP show that the performance of the zero-shot approach with LLMs lags behind the supervised approach (Khondaker et al., 2023). The variation in performance between English and Arabic is derived from the fact that the English corpora represent a large portion of LLM's pre-training data, resulting in an inherent bias to-

[1] TyDi QA states that "The prompts are provided merely as inspiration to generate questions on a wide variety of topics"

[2] These questions are easy to formulate by appending the phrase "What/Who is" to the article title.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 191–207

Category	ARCD	ArabicSQuAD	AQAD	TyDi QA	ArTrivia
Number of Questions	1,395	48,344	17,911	15,726	10,045
Number of Passages	465	10,364	3381	11,319	7,982
Number of Articles	155	231	299	9,166	7,594
Questions Per Article	9.0	209.3	59.9	1.7	1.3
Crowd Workers	✓	✗	✗	✓	✗
Machine Translation	✗	✓	✗	✗	✗

Table 1: Summary of existing Arabic QA datasets compared to ArTrivia. Table adapted from (Atef et al., 2020)

ward English NLP tasks(Lai et al., 2023).

The existing issues in the Arabic QA dataset, which we previously discussed, motivate us to introduce our ArTrivia dataset. The ArTrivia dataset adopts two distinguished approaches. First, we rely on structured datasets from Wikipedia and a new proposed pipeline to generate our dataset, thereby mitigating the cost of crowd-sourcing and the issue of picking common patterns in question formulation. Second, we prioritize having a variety of topics in our dataset, including underrepresented topics, such as classical Arabic poetry.

Thus, the contributions of our paper are summarized in the following:

- We introduce a new novel pipeline to generate question-answer-passage triplets, which leverage various structured data sources from Arabic Wikipedia
- We introduce ArTrivia, a new Arabic Question Answering dataset comprising +10,000 question-answer-passage triplets, covering a wide range of 18 diverse topics in Arabic. We released ArTrivia dataset to the public at https://github.com/salrowili/ArTrivia.
- We conduct a statistical analysis of our dataset and a detailed evaluation of each component in our pipeline. In addition, we provide a detailed evaluation of our dataset against TyDi, using different setups to investigate the impact of out-of-distribution issue in TyDi QA dataset.

2 Related Work

In this section, we will provide an overview of existing Arabic datasets and Arabic Language Models, all of which are part of our evaluation setup.

2.1 Arabic Question Answering Datasets

Several Arabic Question Answering datasets have been introduced recently, including TyDi QA, AQAD, TyDi, ARCD, and ArabicSQuAD: a machine translation of the English SQuAD dataset (Mozannar et al., 2019). Table 1 provides a summary of these datasets. The table shows that ARCD, ArabicSQuAD, and AQAD utilize fewer than 300 articles for generating questions, with a higher ratio of questions per article. This higher question-per-article ratio suggests that despite having many questions, the diversity of articles and topics covered is limited. In addition, while the AQAD dataset does not rely on machine translation, it uses an algorithm to find a matched article in Arabic Wikipedia to those on the English SQuAD dataset. Thus, it includes the same topics covered in the English SQuAD dataset.

On the other hand, TyDi relies on crowd workers for dataset creation and is also part of multi-language datasets. Thus, the TyDi dataset may have a limited representation of specific topics related to the Arabic language, such as classical Arabic poetry. In contrast, we can observe from the table that ArTrivia stands out among other datasets as the only dataset that employs many articles for question generation without depending on Crowd Workers or Machine Translation. Furthermore, despite both the TyDi and ARCD datasets relying on crowd-sourcing for dataset generation, ArTrivia still maintains a lower question-per-article ratio of 1.3 in comparison to TyDi and ARCD datasets.

2.2 Arabic Language Models

The introduction of the BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019), has shown impressive results on English question-answering tasks. Consequently, several Arabic Language Models have adopted BERT-like models, such as AraELECTRA (Antoun et al., 2021), AraBERT (Antoun

et al., 2020), and ArabicTransformer (Alrowili and Shanker, 2021). These models represent the state-of-the-art models in Arabic QA for both TyDi and ARCD datasets. Recently, the advent of Generative Large Language Models (LLMs) like ChatGPT (OpenAI, 2023) has also demonstrated considerable potential in English QA tasks, especially with a zero-shot approach. However, the performance of LLMs such as ChatGPT and Google PaLM 2 still lags behind the typical supervised approach with BERT-like models on Arabic QA tasks as shown by Khondaker et al. (2023) and Anil et al. (2023).

3 Building ArTrivia Dataset

Our approach to build our ArTrivia dataset consists of of two components: (1) generate question-answer pairs from various structured data from Wikipedia, and (2) build a new pipeline that consists of multiple functions to generate our question-answer-passage triplets. First, we will explain in section 3.1, our method to generate question-answer pairs. Then, in section 3.2, we will explain our proposed pipeline to generate our question-answer-passage triplets.

3.1 Question-Answer Pairs Collection

In Figure 1, we illustrate the data collection process of our ArTrivia question-answer pairs from different Wikipedia sources including Wiki Tables, Wiki-Data, WikiList and Wiki Entity Description.

Wikipedia Tables The first method of creating our question-answer pairs is derived from tables within Arabic Wikipedia articles. These tables have a set of relationships between two or more items in the table (e.g., a list of capital cities). We exploit these relations to formulate our question-answer pairs. The first part of the relationship will form the question and keyword (A), and the second part will serve as the answer (B). Then, we will use a fixed term for each set of relations to form our question (e.g. What is the capital of (A) country? Answer: (B). The selection of these tables is based on two criteria: (a) questions can be answered by trivia enthusiasts, (b) covers a wide variety of topics (e.g., history, poetry), and a variety of question types (e.g., numbers, dates, persons, and places).

Wiki Data The second method shares similarities with the WikiTables but leverages structured datasets related to specific entities within Wikipedia, utilizing a knowledge base known as WikiData. The WikiData stores valuable relationships for each entity. For example, Rome's entity in WikiData includes relationships like capital city, inception, nickname, and "founded by." Similarly, Thomas Edison's entity has information like country of citizenship, date of birth, and notable work. Thus, by utilizing these relationships, we generate additional QA pairs in our dataset. Our choice of these entities depends on the entity's popularity, measured by the number of languages to which this particular entity has been translated to.

Wikipedia List We observe that the TyDi QA dataset has a limited number of long questions (e.g., terms in economics). To address this gap, we generate 591 question-answer pairs from Wikipedia lists using a simple parser as illustrated in Figure 1.

Wikipedia Entity Description While we were able to generate over 12.7K question-answer pairs using both WikiData and Wiki Tables, we still have a challenge in generating certain types of questions that require more complexity (e.g., smallest planet, second largest country). We observe that Wikipedia annotators populate valuable information for each entity (e.g., persons, places, novels) in the central description of the article title, as shown in figure 6. This information provides a short summary of each entity (article title), highlighting important information related to this entity. For example, an article with the title "Mercury" has a central description that says, "smallest and closest planet to the sun in the Solar System." By using the ChatGPT prompt as illustrated in Figure 1, we can generate a related question for this entity.

Our selection process for entities for this type of question depends on the entity's popularity, following these steps: First, we use the Arabic Wikipedia dump to extract all article titles, selecting only those with central descriptions. Second, we sort these entities based on the number of languages each has been translated into. Finally, we exclude entities that lack a sufficient description to form a question that can be answered by trivia enthusiasts.

3.2 Building ArTrivia Pipeline

We explained earlier our methods to generate question-answer pairs from WikiTable, WikiData, Wiki List, and Wiki Entity Description. However, to build question-answer-passage triplets, we need to find the relevant passage for each question-answer pair. To address this part, we propose a

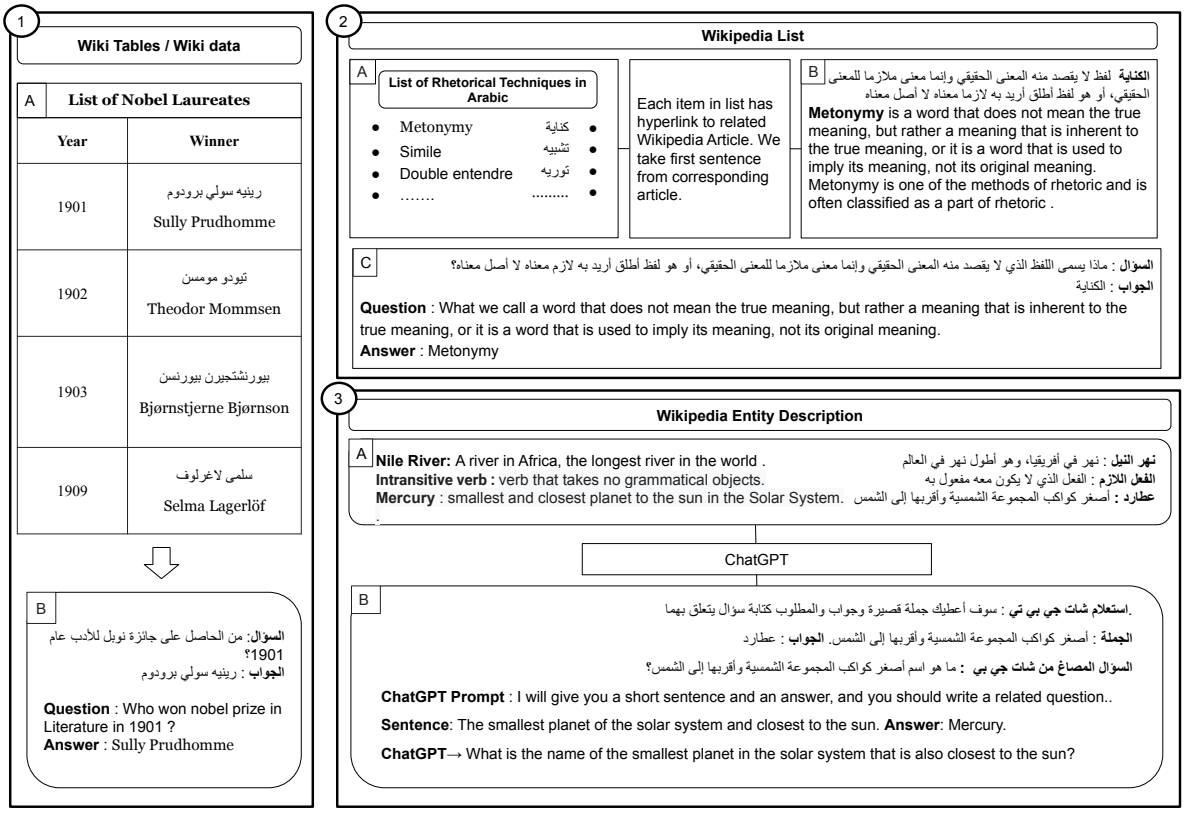


Figure 1: Overview of our method to build the ArTrivia question-answer pairs from different Wikipedia sources.

new novel pipeline that consists of (1) a BM25 retriever (Robertson and Zaragoza, 2009) (2) fuzzy match and approximation functions, (3) a places parser, (4) answer normalization functions, and (5) ChatGPT as an annotation tool to filter irrelevant passages. An overview of our proposed pipeline is illustrated in Figure 2.

Finding Relevant Passage First, it is important to highlight that this stage is not necessary for questions sourced from the Wiki List and Wiki Entity description. In the case of Wiki List, each list item has a hyperlink to the associated related article. Thus, we can consider the first passage in this related article as the relevant passage. Similarly, for questions derived from Wiki Entity descriptions, our entity corresponds to article titles, as mentioned earlier. In the majority of cases, the essential information required to address the question is adequately present in the first passage.

However, for both WikiTable and Wiki data question answer pairs, we need to find the relevant passage using a retrieval model. To build our retrieval component, we first split articles from the Arabic Wikipedia (June 2023) into 100 words, each representing a passage. Then, we use the sparse-based retrieval BM25 with the Pyserini tool (Lin et al., 2021) to build our indexed Arabic Wikipedia [3].

Fuzzy Matching To control the quality of retrieved passages, we use the first elements in the WikiTables and WikiData as a keyword. For example, a related passage for a question-answer pair says, "Who was the winner of the Nobel Prize for Literature in 1901? Sully Prudhommem", should have the following keywords:(1) Nobel, (2) 1901, and (3) Sully Prudhommem.

However, relying on the exact match of keywords to control the retrieved message will eliminate many passages that have the keywords but with different forms. This case will be worse with morphologically rich language such as Arabic. Thus, we integrate an approximate string matching (fuzzy match) function with our pipeline, which is based on the "thefuzz" library (Adam, 2023).

In addition, to handle measurement-related questions, we use an approximation function that accepts answers within +/- 10% of the actual value. The reason to include an approximation function

[3]While alternative approaches like DPR: Dense Passage Retrieval (Karpukhin et al., 2020) could be considered, we use BM25 to maintain simplicity in our pipeline.

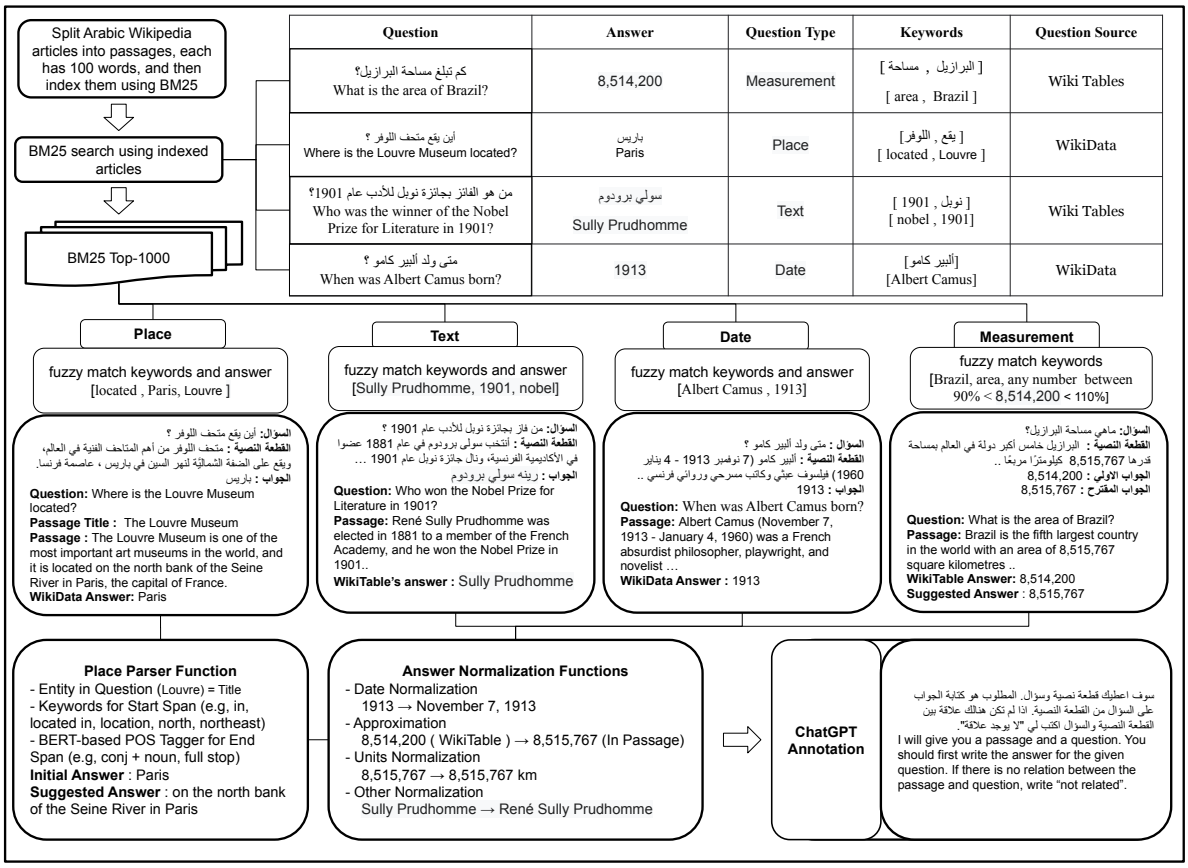


Figure 2: Overview of our proposed pipeline to build our ArTrivia question-answer and relevant passage triplets.

is to mitigate the disputes related to measurements (e.g., rivers' length, country areas, and populations) between structured data (WikiData, WikiTables) and the related passage.

Place Parser Function We find that in questions related to places (e.g., museums and cities locations), in most cases, the retrieved passage will have a better ground truth answer than the initial answer derived from Wiki Tables and WikiData as shown in our example in Figure 2. Thus, we construct a parser to revise answers related to questions about places according to the related passage.

Our Place Parser Function consists of three key components. The first step is to choose the top retrieved passages where the passage's title matches the corresponding place name mentioned in the question. For example, if we have a question inquiring about the location of London City, with "London" as the keyword, we will specifically select passages whose titles match "London."

The second step employs a fuzzy keyword-matching approach to check if any word of our predefined list of places keywords in the passage. These keywords include terms such as "located in," "north," "south," and "northeast". These keywords help us determine the starting point of the answer span within the passage.

Finally, in the third step, we utilize an Arabic BERT-based POS (Part-of-Speech) tagger (Inoue et al., 2021) to determine the end span of the answer. This tagger employs specific POS tags and follows a set of predefined conditions, including, for example, the presence of punctuation marks or conjunctions followed by Proper Nouns.

Answer Normalizing Function To address discrepancies between initial answers from our QA pairs and corresponding ground truths in related passages, we have added an answer-normalizing function to our pipeline. This normalization function targets three aspects: (1) variations in date formats (e.g., "1913" to "November 7, 1913"), (2) differences in units and formatting (e.g., "2400" to "2400 km"), and (3) entities' alternative names (e.g., "Thomas Edison" to "Thomas Alva Edison").

To tackle inconsistencies in dates and units, we first create a reference file containing a list of words

related to dates and units. This list includes months, possible years (e.g., 1-3000), and a list of units (e.g., km, mile, square km). Our normalization function operates by starting from the index of the original answer's location and scanning adjacent terms on both sides. Then, It appends matching terms from our list until it encounters an unlisted term.

For example, consider the sentence "Albert Camus (November 7, 1913 - January 4, 1960) was a French absurdist philosopher, playwright, and novelist". If the initial answer from WikiTable is "1913", our Answer Normalizing function scans adjacent terms, identifying "November" and "7" as valid matches from our list. However, it stops upon encountering symbols like "-" or "(", which are not part of our list of units.

Furthermore, we have improved our normalization function to handle questions involving date ranges (e.g., When did the Macedonian Empire rule?). To address the date ranges question, we simply include symbols and terms associated with date ranges in our reference file, such as "-", "till", "between", "until", and "continued till". For example, when changing the query from "When was Albert Camus born?" to "When did Albert Camus live?", our normalizing function will not stop at the "-" symbol. Instead, it continues scanning left and right until encountering a token not in our list, such as the "(" and ")" symbols. As a result, the normalized ground truth for this example would be "November 7, 1913 - January 4, 1960". We manually flag questions related to date and date range, which is a trivial task since our dataset consists of a set of relationships sharing similar answer types.

Data Annotation with ChatGPT Several studies have suggested that LLM models such as ChatGPT could be used as data annotation tools (Gilardi et al., 2023), (Huang et al., 2023). Drawing from these encouraging findings, we added an additional phase into our pipeline. This phase leverages a ChatGPT prompt, as shown in Figure 2, to filter irrelevant passages generated from our pipeline and validate the suggested answer from our pipeline.

Quality Control This stage is to assess the quality of our pipeline and ChatGPT as an annotation tool. In this stage we manually examine question triplets from our pipeline for the following points (1) if the selected passage is related, (2) if the suggested answer from our pipeline represents the ground truth and revise the ground truth if needed

Stage	#Question
Question-Answer Pairs	15,149
Pipeline (QA Pairs + Passage)	11,527
Filtering Non-Relevant Passages	-1,482
Final ArTrivia Dataset	10,045
- WikiTables	4,916
- WikiData	2,987
- Wiki Entity Description	1,579
- WikiList	563

Table 2: Detailed statistics about our ArTrivia dataset.

(3) For the development set, we add any other possible alternative answers in the related passage.

4 Dataset Analysis

In this section, we will focus on the quantitative analysis of our dataset. Then, in section 5, we will focus on evaluating the performance of ArTrivia as a training and evaluation dataset.

QA Pairs Collection The data collection is quantitatively summarized in Tables 2. This collection includes 15,149 question-answer pairs sourced from a variety of Wikipedia origins, with contributions from WikiTables (49%), WikiData (35%), WikiList (5%), and Wiki Entity Description (11%).

Results of our Pipeline The employment of our pipeline, as illustrated in Figure 2, successfully identified a relevant passage for 11,527 question-answer pairs. In contrast, the pipeline could not find a related passage for 3,622 question-answer pairs. This outcome was in line with our expectations, knowing that Arabic Wikipedia consists of only 2.1 million articles compared to English Wikipedia, which has more than 16 million articles.

Manual and ChatGPT Filtering By utilizing the ChatGPT prompt against 11,527 triplets from our pipeline, a total of 566 passages were classified by ChatGPT as non-relevant. Subsequently, employing manual filtering against the same 11,527 triplets resulted in the identification of 1,482 irrelevant passages. Next, by comparing the manual filtering process to ChatGPT's filtering, we discovered that 154 question-answer-passage triplets were incorrectly classified as irrelevant by ChatGPT. Thus, despite the promising results in English, our result shows that ChatGPT lags behind human performance in Arabic QA annotation.

Question Word	ArTrivia	TyDi
What (ما)	36.6%	30.4%
When (متى)	22.8%	28.9%
Who (من هو)	23.8%	17.9%
Where (أين)	13.3%	12.0%
How Much / Many (كم)	3.5%	9.61%
YES/NO (هل)	<1%	<1%
How (كيف)	<1%	<1%
Why (لماذا)	-	<1%

Table 3: Distribution of ArTrivia by question word against the Arabic portion of TyDi QA.

Final ArTrivia Dataset The final dataset consists of 10,045 question-answer-passages triplets, which suggests that the accuracy of our pipeline in retrieving related passages is 87% (10,045 out of 11,527). Following the final manual filtering stage, we split our ArTrivia dataset into training and development sets. Our strategy was to select 20% of each relationship (e.g., list of capitals) in WikiTables, Wiki Data as part of our development set. For Wiki List and Entity Description questions, we randomly selected 20% of the dataset for the development set. This split ensures that the development set is representative of the entire dataset.

Topics Distribution In Table 6, we show the distribution of topics in ArTrivia, which shows that ArTrivia covers a wide variety of 18 topics. The distribution of topics also shows how we addressed under-represented topics in existing Arabic QA datasets, such as Arabic Literature, Cartoon Movies, Arabic Cinema, and National dishes. We show examples of each of these categories along with other categories in Figure 3 and Figure 4. The table also shows that History, Geography, and "Dates of Birth/Death of Famous people" are the most represented topics in our dataset. However, many of the questions under History topics are in the grey area of other topics such as politics, geopolitics, and world organization history.

Question Word Distribution In Table 3, we compare the distribution of question words in our dataset against TyDi. While our ArTrivia dataset shows a higher proportion than TyDi for both "what" and "who" questions, TyDi still has a larger overall number of questions for these categories. On the other hand, we can observe that our dataset demonstrates a lower proportion of questions starting with the "when" word compared to Arabic TyDi. It is worth noting that the English subset of the TyDi QA dataset constitutes only 14% of the entire question pool dedicated to "when" questions. It is also important to note that "where" questions are mostly categorized into "Geography" topics. Thus, increasing the questions in "Geography" topics has helped us maintain a similar distribution of "where" questions to the TyDi QA dataset.

Furthermore, the table shows that both the Arabic TyDi dataset and our ArTrivia dataset contain less than 1% of "YesNo" and "How" questions. Most of the questions in these two categories come from WikiData, indicating it is effectiveness in generating these types of questions. It is also worth noting that TyDi includes 31 questions related to "why" questions, which we encountered challenges in generating using our pipeline.

5 Pipeline and Dataset Evaluation

In this section, we will first discuss the evaluation performance of our proposed pipeline to highlight the impact of each stage in our pipeline. Then, we will discuss and compare the performance of ArTrivia and TyDi datasets using different setups for training and evaluation sets. This will help us to study the out-of-distribution and study how ArTrivia can serve as a complementary dataset to TyDi QA.

5.1 ArTrivia Pipeline Evaluation

Table 4 shows a comprehensive evaluation of our pipeline using the AraELECTRA model on the $\text{TyDi}_{\text{short}}$ dataset. The $\text{TyDi}_{\text{short}}$ subset of TyDi eliminates questions that inquire about entity explanations (e.g., What is a space galaxy?). The main objective of this evaluation is to assess the individual contributions of each component in our pipeline against our baseline dataset (TyDi training set), shown in the last row of the table.

Initially, we use a basic approach to retrieve question-answer-passage triplets by checking if answers can be located within passages retrieved by the BM25 model. This basic strategy yields an EM/F1 score of 38.3/57.7 and retrieved 13,407 triplets. Then, we introduce a second strategy that uses the question keywords to reduce the possibility of retrieving irrelevant passages, which resulted in a slight improvement of the EM score to 40.4.

Stage	#Q	EM/F1
ArTrivia QA Pairs	15,149	-
ArTrivia Pipeline		
+ Answer In Passage	13,407	38.3/57.7
+ Question Keywords	10,290	40.4/60.2
+ FuzzyMatch	11,265	40.9/60.1
+ Approximation	11,462	40.6/60.3
+ Date Normalization	11,459	56.8/70.1
+ Other Normalization	11,459	62.2/72.1
+ PlaceParser	11,527	64.9/76.1
ArTrivia Quality Control		
- Irrelevant Passages	10,045	64.9/75.0
+ Revised Answer Span	10,045	70.0/79.9
Baseline Dataset		
TyDi Training Dataset	14,805	74.9/84.3

Table 4: Exact Match (EM) and F1 scores of AraELECTRA with our pipeline on $\text{TyDi}_{\text{Dev-Short}}$.

However, using the exact match of keywords with morphologically rich languages like Arabic, where words can have different forms, causes a reduction for our triplets by 3,117 examples. To address this issue, we incorporate a FuzzyMatch function, which restores over 975 triplets. We then include another function in our pipeline, the approximation function. This function recovers an additional 197 triplets and raises the F1 score to 60.3.

However, the most significant improvement in our pipeline occurs with the introduction of our normalization functions, which increases our pipeline's performance from 40.6/60.3 to 62.2/72.1. These results highlight the often overlooked role of answer normalization in enhancing the overall quality of question-answering datasets.

Furthermore, the PlaceParser function, detailed in section 3.2, substantially improves our pipeline's performance, contributing an additional 2-4 points to the EM/F1 score. Thus, the final performance with our pipeline without the additional manual quality control is 64.9/76.1 compared to the TyDi training set, which achieved 74.9/84.3.

Finally, our quality control stage, as outlined in Section 3.2, had significantly contributed to the overall performance, improving our score to 70.0/79.9. This significant enhancement was primarily attributed to the manual refinement of the start and end spans of the suggested answers generated by our pipeline. In total, we undertook a total of 1,378 answer span revisions ranging from minor to major revisions. These revisions include cases where (1) the suggested answer from our pipeline is a single entity, where ground truth is a multiple entities (2) the presence of unusual date formats and units within the passage, and (3) the absence of essential prefixes, suffixes, and articles.

On the other hand, we can observe that filtering out irrelevant passages did not yield any additional improvements in our results. This finding implies that the Language Model can tolerate having irrelevant passages in the training set without compromising performance. However, this filtering step remains critical to maintain the quality of our ArTrivia dev set as a reliable evaluation dataset.

5.2 ArTrivia Evaluation

In Table 5, we provide an evaluation of our Final ArTrivia dataset in comparison to TyDi. We finetune the AraELECTRA model with various dataset configurations for this evaluation. Our primary objective is to assess how well our ArTrivia dataset could serve as a complementary dataset to TyDi and to examine the challenge that ArTrivia introduces to a language model trained on TyDi QA.

In the first two rows, we show the evaluation of ArTrivia and TyDi on the complete TyDi dataset. We can observe in row 2 that the ArTrivia dataset shows a lower performance in this setup, with an F1 score of 60.7, against a score of 86.8 with TyDi QA. These results are as expected, given that our ArTrivia dataset only addresses short-answer questions. In contrast, TyDi includes a substantial 33% of its dataset dedicated to long-answer questions, which typically ask to explain an entity.

Next, in rows 4-5, we replicate a similar evaluation setup presented in Table 4 by removing 300 questions that typically have long answers from TyDi QA. With this setup, the gap with TyDi significantly decreased to less than 4.4 in the F1 score. The marginal gap between ArTrivia and the TyDi training set is expected, given that ArTrivia is an out-of-distribution dataset for TyDi. Indeed, this gap is larger when we evaluate TyDi on the ArTrivia development set, as we will discuss next.

In rows 7-8, we replaced the development dataset in our table with our ArTrivia dev dataset. Comparing rows 7 and 8 shows how significantly the TyDi training set underperforms against our ArTrivia with a margin of 8.4 in the F1 score and 10.4 in the EM score. Considering that ArTrivia includes a manual quality control phase that examines ev-

Row	Train Size	Eval Size	Train Data	Eval Data	EM	F1
1	14,805	921	$\text{TyDi}_{\text{Train}}$	$\text{TyDi}_{\text{Dev-Full}}$	74.5	86.8
2	10,045	921	ArTrivia	$\text{TyDi}_{\text{Dev-Full}}$	49.1	60.7
3	24,805	921	ArTrivia + $\text{TyDi}_{\text{Train}}$	$\text{TyDi}_{\text{Dev-Full}}$	74.7	86.6
4	14,805	621	$\text{TyDi}_{\text{Train}}$	$\text{TyDi}_{\text{Dev-Short}}$	74.9	84.3
5	10,045	621	ArTrivia	$\text{TyDi}_{\text{Dev-Short}}$	70.0	79.9
6	24,805	621	ArTrivia + $\text{TyDi}_{\text{Train}}$	$\text{TyDi}_{\text{Dev-Short}}$	75.7	85.1
7	14,805	1,700	$\text{TyDi}_{\text{Train}}$	$\text{ArTrivia}_{\text{Dev}}$	79.0	84.9
8	8,345	1,700	$\text{ArTrivia}_{\text{Train}}$	$\text{ArTrivia}_{\text{Dev}}$	89.4	93.3
9	23,150	1,700	$\text{TyDi}_{\text{Train}}$ + $\text{ArTrivia}_{\text{Train}}$	$\text{ArTrivia}_{\text{Dev}}$	89.5	93.0

Table 5: The Exact Match (EM) and F1 scores of the AraELECTRA model using different setups for Training and Development datasets with ArTrivia and TyDi QA. In the $\text{TyDi}_{\text{Dev-Short}}$ setup, questions that ask about entity descriptions (e.g., "Who is Alfred Nobel?") were excluded since these questions typically have long answers.

ery example in our development set, it is clear that the decline in performance with the TyDi is not attributed to the poor quality of ArTrivia. This suggests that ArTrivia presents a more challenging question to TyDi.

In Figure 5, we present examples that highlight the challenges our dataset poses for the TyDi QA dataset. The Figure illustrates that in the majority of these examples, the Language Model tends to select the first entity corresponding to the question type. For example, when the question asks about a person entity, such as a novel author, and the passage has another person's name before the actual answer, the Language Model often selects the first name mentioned in the passage. This pattern is consistently observed with dates, places, and other types of entities as well. These cases raise important questions about whether the Language Model relies on context to answer a given question or if it simply adapts to common patterns associated with question words (e.g., when, where, who). This also suggests that we could build a more challenging QA dataset in the future based on these observations.

Finally, we should also note that by contrasting rows 5 and 7, we can observe that ArTrivia present more tolerance for out-of-distribution issue than TyDi. Furthermore, upon comparing rows 3, 6, and 9, we can conclude that combining TyDi and ArTrivia as complementary datasets achieves almost the best score against TyDi and ArTrivia devolvement datasets.

6 Conclusion

In this paper, we introduce ArTrivia, a novel dataset consisting of over +10,000 question triplets, covering a wide range of 18 diverse topics. We present a detailed description of our proposed pipeline and conduct a comprehensive analysis of the contribution of each component to overall performance. While most of the existing research on Question Answering datasets primarily focuses on question formulation and passage selection, our work emphasizes the overlooked, yet crucial, role of answer normalization in the quality of QA datasets. Our results also highlight the out-of-distribution issue within TyDi when presented with more challenging questions. In future work, we plan to adapt our proposed pipeline to different domains and languages such as creating a new Multi-Language QA dataset.

Ethics Statement

The ArTrivia dataset is collected from different sources of Arabic Wikipedia structured datasets. These datasets are populated by human annotators (Wikipedia Contributors). Wikipedia adapts the Neutral point of view (NPOV) policy [4], defined as "representing fairly, proportionately, and, as far as possible, without editorial bias, all the significant views that have been published by reliable sources on a topic."

Acknowledgements

The authors would like to acknowledge the ultimate support from Google Research Cloud TRC for providing access to Tensor Processing Unit TPUs. In this paper, we employ TPU VM, which has 96 CPUs and 350 GB of RAM, to efficiently run our QA pipeline in parallel and evaluate our dataset.

[4] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view/FAQ

Limitations

ArTrivia uses Wiki Tables from Arabic Wikipedia to generate a large proportion of the dataset. One limitation of this method is that it lacks having more complicated questions such as "Why" questions, as shown in Table 3. In the future, we plan to overcome this limitation by finding new methods to retrieve these types of questions from new sources of structured data in Wikipedia.

Moreover, another limitation of this work is that we have a limited number of structured tables in Arabic Wikipedia. However, we plan to overcome this limitation by using machine translation of tables from English Wikipedia. Using machine translation for a dataset like SQuAD may not yield optimal results. However, using machine translation with our method that uses a table from Wikipedia may yield better results since we often in this case translate entities in the table rather than translating complete passages.

Finally, another limitation of our dataset creation method is that we use a fixed term to generate question-answer pairs for each set of relationships from WikiTable and WikiData. However, this limitation can be easily overcome by using LLMs (e.g. ChatGPT) to generate the question phrase for each relationship in the structured dataset.

References

Cohen Adam. 2023. thefuzz. https://github.com/seatgeek/thefuzz.

Sultan Alrowili and Vijay Shanker. 2021. ArabicTransformer: Efficient large Arabic language model with funnel transformer and ELECTRA objective. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1255–1261, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Adel Atef, Bassam Mattar, Sandra Sherif, Eman Elrefai, and Marwan Torki. 2020. Aqad: 17,000+ arabic questions for machine comprehension of text.

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.

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.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for

text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

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.

OpenAI. 2023. `https://chat.openai.com`.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

A Appendix

Topic	#Questions
Inventions and Discoveries (اختراعات واكتشافات)	231
Global Literature (أدب و روايات عالمية)	455
Arabic Literature (أدب ولغة عربية)	469
Cartoon Movies and Manga (أفلام كرتون وقصص مصورة)	284
Economy and Business (اقتصاد وأعمال)	210
History (تاريخ)	1645
Dates of Birth/Death of Famous people (تاريخ ميلاد ووفاة شخصيات شهيرة)	1274
Geography (جغرافيا)	1714
Awards and Prizes (جوائز والقاب)	355
Countries and Currencies (دول وعملات)	339
Sports and Olympic Games (رياضة والعاب اولمبية)	740
Global Cinema (سينما وأفلام عالمية)	107
Arabic Cinema (سينما وأفلام عربية)	409
Press and Media (صحافة وإعلام)	141
Science and Astronomy (علوم وفلك)	955
Food and National Dishes (غذاء واطباق وطنيه)	76
Art (فن)	133
Others (معلومات أخرى)	508
Total	10,045

Table 6: Topics distribution in ArTrivia Dataset. To have an accurate topics distribution of our dataset, we manually annotate the topic category for each question.

السؤال	القسم
من هو مخترع الرادار؟	اختراعات واكتشافات
من هو الشاعر المسرحي العظيم صاحب مسرحيات : يوليوس قيصر، الملك لير، هاملت، ماكبث؟	أدب عالمي
من هو الشاعر العربي الذي اشتهر بعشقه المتبادل مع ليلى الأخيلية؟	أدب عربي
من هي الشخصية الرئيسة في سلسلة المحقق كونان ؟	أفلام كرتون
ماهو اسم المؤشر المالي الذي اخترعته مجلة ذي إيكونوميست عام 1986؟	اقتصاد وأعمال
من هو مؤسس سلالة تانغ الحاكمة؟	تاريخ
متى ولد مايكل أنجلو؟	تاريخ ميلاد ووفاة شخصيات شهيرة
أين تقع صحراء باتاغونيا ؟	جغرافيا
من هو العالم المصري الحائز على جائزة نوبل في الكيمياء ؟	جوائز وألقاب
ما هو اسم الطائر الوطني لدولة الأردن ؟	دول وعملات
من هي صاحبة الرقم القياسي الأولمبي للسيدات في سباق 200 متر؟	رياضة و ألعاب أولمبية
من كان البطل الرئيسي في فيلم فورست غامب؟	سينما عالمية
من مخرج فيلم سواق الأتوبيس؟	سينما عربية
من قام بتأسيس صحيفة الديلي ميل ؟	صحافة وإعلام
هل التسارع كمية متجهة أم قياسية ؟	علوم وفلك
ماهي الدولة التي تشتهر بطبق المروزية ؟	غذاء وأطباق وطنية
من هو الفنان صاحب لوحة معركة الإسكندر في إسوس؟	فن
ماهو الشهر السرياني الذي يقابل شهر يوليو؟	معلومات أخرى

Figure 3: Examples of ArTrivia dataset from 18 different diverse topics.

Question	Category
Who invented the Radar?	Inventions and Discoveries
Who is the great poet and playwright who wrote the plays: Julius Caesar, King Lear, Hamlet, and Macbeth?	Global Literature
Who is the Arab poet who is famous for his mutual love with Laila Al-Akhiliya?	Arabic Literature
Who is the main character in the Detective Conan series?	Cartoon Movies
What is the name of the financial indicator that was invented by The Economist magazine in 1986?	Economy
Who is the founder of the Tang dynasty?	History
When was Michelangelo born?	Dates of Birth/Death of Famous people
Where is the Patagonian desert located?	Geography
Who is the Egyptian scientist who won the Nobel Prize in Chemistry?	Awards and Prizes
What is the name of the national bird of Jordan?	Countries and Currencies
Who is the women's Olympic record holder in the 200 meters?	Sports and Olympic Games
Who plays the hero role in Forrest Gump?	Global Cinema
Who is the director of the movie The Bus Driver?	Arabic Cinema
Is acceleration a vector or scalar quantity?	Science and Astronomy
Who founded the Daily Mail?	Press and Media
Which country is famous for its Marouzia dish?	Food and National Dishes
Who is the artist who painted The Battle of Alexander at Issus ?	Art
What is the Syriac month that corresponds to July?	Others

Figure 4: Examples of ArTrivia dataset from 18 different diverse topics. This is the English Translation of Figure 3.

السؤال : ماهي المدينة التي استضافت الالعاب الاولمبية عام 1936؟
القطعة النصية : المانيا في الالعاب الاولمبيه لدي الالعاب الاولمبيه شعبيه عند الالمان، المانيا ايام الامبراطوريه كانت اول الدول التي شاركت في الالعاب الاولمبيه الصيفيه 1896 في اثينا في اليونان وشاركت في اول 5 العاب اولمبيه، لكنها لم تشارك في اولمبياد 1920 و 1924 احتجاجا علي عدم استضافه برلين للاولمبياد والتي اختيرت مسبقا قبل الحرب العالميه الاولي، وثم عادت المانيا للمشاركه في 1928 بعد ان تم السماح لبرلين باستضافه الالعاب الاولمبيه في1936، استضافت مدينه برلين عاصمه المانيا الالعاب الاولمبيه الصيفيه 1936 في ايام حكم ادولف هتلر وانهت المانيا الاولمبياد في صداره جدول الميداليات، بعد انقسام المانيا الي المانيا الشرقيه و المانيا الغربيه .
الجواب الصحيح : برلين
الجواب المقترح : اثينا

Question : What city hosted the Olympic Games in 1936?
Passage : Germany in the Olympic Games The Olympic Games are popular with the Germans. Germany in the days of the empire was the first country to participate in the 1896 Summer Olympics in Athens, Greece, and participated in the first 5 Olympic Games, but it did not participate in the 1920 and 1924 Olympics in protest against Berlin not hosting the Olympics, which It was chosen in advance before World War I, and then Germany returned to participate in 1928 after Berlin was allowed to host the Olympic Games in 1936. The city of Berlin, the capital of Germany, hosted the 1936 Summer Olympics during the days of Adolf Hitler's rule, and Germany finished the Olympics at the top of the medal table, after Germany was divided into East Germany and West Germany .
Actual Answer : Berlin
Prediction : Athens

السؤال : ماهو اسم المسلسل الكرتوني يتحدث المسلسل عن ابن التاجر هيثم أحد التجار المشهورين في العراق، ورحلاته مع صديقيه علي بابا و علاء الدين وطائره ياسمينة؟
القطعة النصية : مغامرات سندباد هو مسلسل رسوم متحركه ياباني من اخراج فوميو كوروكاوا ومن انتاج شركه نيبون انيميشن ويحوي 52 حلقه. تاريخ عرضه لاول مره كان في 1 اكتوبر 1975 واستمر حتي 29 سبتمبر 1976. مسلسل مغامرات سندباد يقوم في الاصل علي القصص القديمه المعروفه الف ليله وليله وفي قصص الف ليله وليله السندباد هو بحار عربي من بغداد يهوي الابحار والمغامرات وتحكي قصصه والمصاعب التي يواجهها ويتغلب عليها وسندباد هنا تاجر مسافر يبحر احيانا واحيانا اخري يسافر بجمله علي البر. القصه. سندباد بطل المسلسل هو ابن التاجر هيثم احد التجار المشهورين في مدينه بغداد، له صديق اسمه حسن (يفترض انه الشاطر حسن) وهو فتي.
الجواب الصحيح : مغامرات سندباد
الجواب المقترح : باباني

Question : What is the name of the cartoon series that talks about the son of the merchant Haitham, one of the famous merchants in Iraq, and his travels with his friends Ali Baba and Aladdin and his pet, Yasmina?
Passage : The Adventures of Sinbad is a Japanese animated series directed by Fumio Kurokawa and produced by Nippon Animation. It contains 52 episodes. The date of its first showing was on October 1, 1975 and continued until September 29, 1976. The series The Adventures of Sinbad is originally based on the well-known old stories One Thousand and One Nights. In the stories of One Thousand and One Nights, Sinbad is an Arab sailor from Baghdad who loves sailing and adventures. His stories are told and the difficulties that he and Sinbad face and overcome. Here is a traveling merchant who sometimes sails and sometimes travels wholesale on land. the story. Sinbad, the hero of the series, is the son of the merchant Haitham, one of the famous merchants in the city of Baghdad. He has a friend named Hassan (presumably the smart one Hassan), who is a young man.
Actual Answer : The Adventures of Sinbad
Prediction : Japanese

السؤال :من هو مبتكر شخصية أرسين لوبين؟
القطعة النصية : الفرنسيين جول رونارد والفونس دوديه، لكن دون تحقيق نجاح جماهيري. خلال تواجده بباريس رافق كبار الكتاب الفرنسيين امثال ستيفان مالارمي والفونس اليه. في عام 1901 نشر كتابه «الحماس». في عام 1905 وبطلب من مدير مجله Je sais tout، بدا لوبلان بكتابه قصص ارسين لوبين ولاقت هذه القصص نجاحا جماهيريا فاجئ الكاتب وصنع له طريق الشهره والثروه. في عام 1907 بدا لوبلان بكتابه روايات كامله حول ارسين لوبين، ونظرا لنسبه المبيعات الجيده التي حققتها هذه الاعمال، قرر الكاتب ان يكرس باقي اعمال مسيرته لهذه الشخصيه، ليبلغ 21 ما بين روايات وقصص قصيره. وتماما مثل كونان دويل وشخصيته شرلوك هولمز، حاول موريس لوبلان التخلص
الجواب الصحيح : [موريس لوبلان, لوبلان]
الجواب المقترح : جول رونارد والفونس دوديه

Question : Who is the creator of the character Arsène Lupine?
Passage : The Frenchmen Jules Renard and Alphonse Daudet, but without achieving mass success. During his stay in Paris, he accompanied major French writers such as Stephane Mallarmé and Alphonse to him. In 1901, he published his book "Enthusiasm." In 1905, at the request of the director of Je sais tout magazine, Leblanc began writing the stories of Arsène Lupine, and these stories met with a popular success that surprised the writer and paved the way for him to fame and fortune. In 1907, Leblanc began writing entire novels about Arsène Lupin, and due to the good sales achieved by these works, the writer decided to devote the rest of his career to this character, amounting to 21 novels and short stories. Just like Conan Doyle and his character Sherlock Holmes, Maurice LeBlanc tried to get away
Actual Answer : [Leblanc , Maurice LeBlanc]
Prediction : Jules Renard and Alphonse Daudet

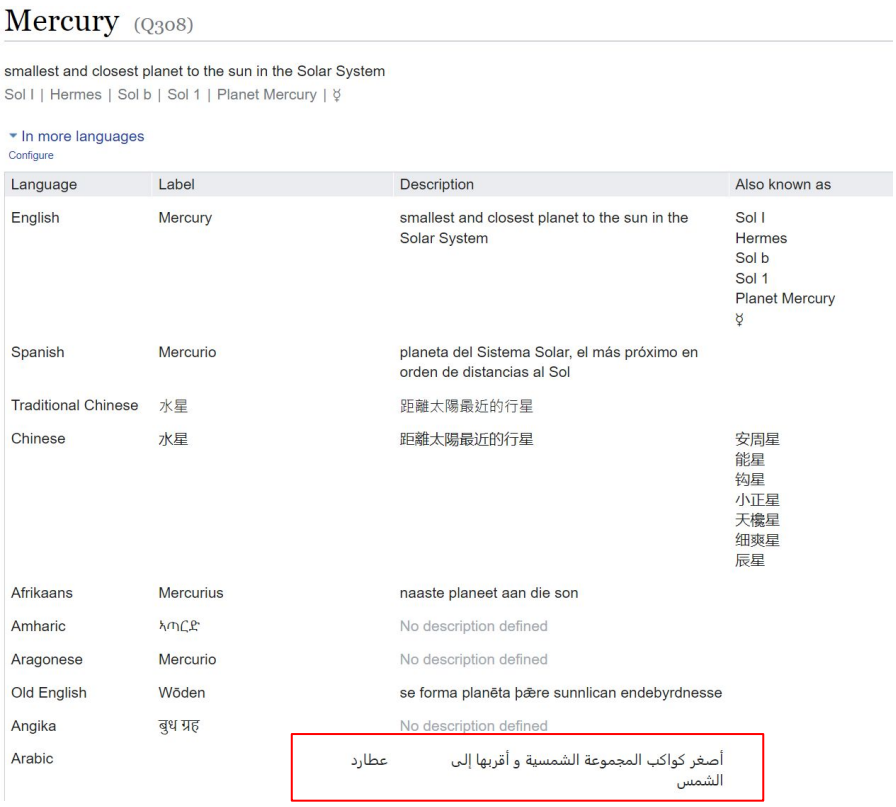
السؤال : من هو الحاصل على جائزة نوبل للسلام لعمله كرنيس لمكتب السلام الدولي؟
القطعة النصية : لوني لافونتين (1854-1949) كانت نسويه بلجيكيه وداعيه بارزه للسلم. ناشطه في الكفاح النسوي الدولي، كانت عضوا في الرابطه البلجيكيه لحقوق النساء، المجلس الوطني البلجيكي للمراه والرابطه النسائيه الدوليه للسلام والحريه. كان شقيقها هنري لافونتين، محام بلجيكي عالمي ورئيس مكتب السلام العالمي الذي حصل علي جائزه نوبل للسلام في عام 1913، وكان ايضا مدافعا قديما عن حقوق النساء وحق الاقتراع، واسس عام 1890 الرابطه البلجيكيه لحقوق النساء. انشات المكتب المركزي لتوثيق النساء عام 1909 بالقرب من مشروع موندانيوم، الذي انشاه بول اوتليت وشقيقه هنري لافونتين ولمفهوم التوثيق، وانشات في منزلها الخاص مكتبه للاتحاد البلجيكي لحقوق النساء، لتساعد النساء في خياراتهم المهنيه. توفيت
الجواب الصحيح : هنري لافونتين
الجواب المقترح : لوني لافونتين

Question : Who won the Nobel Peace Prize for his work as head of the International Peace Bureau?
Passage : Lonnie La Fontaine (1854-1949) was a Belgian feminist and prominent pacifist. Active in the international feminist struggle, she was a member of the Belgian League for Women's Rights, the Belgian National Council of Women and the Women's International League for Peace and Freedom. Her brother, Henri La Fontaine, was an international Belgian lawyer and head of the Universal Peace Bureau who won the Nobel Peace Prize in 1913. He was also a long-time advocate of women's rights and suffrage, and in 1890 founded the Belgian League for Women's Rights. She established the Central Office for Documentation of Women in 1909 near the Mondanium project, which was established by Paul Otlet and his brother Henri La Fontaine and for the concept of documentation, and she established in her private home an office for the Belgian Federation for Women's Rights, to help women in their professional choices. She died
Actual Answer : Henry La Fontaine
Prediction : Lonnie LaFontaine

السؤال : ماهو العدد الذري لعنصر النيتروجين؟
القطعة النصية : العظيم مباشره. اما باقي العناصر ذات كتل ذريه اثقل من الهيدروجين والهيليوم فقد نشات «طبخت» في قلب النجوم حيث الحراره العاليه التي تفوق 14 مليون درجه مئويه واحيانا تصل الي مليار درجه مئويه بحسب كتله النجم. في النجوم تتكون العناصر الاثقل من الهيدروجين والهيليوم عن طريق اندماجها النووي وتتكون العناصر منها الليثيوم (العدد الذري 3) والكربون (العدد الذري 6) والنيتروجين (وعدده الذري 7) والاكسجين (عدده الذري 8) والصوديوم (العدد الذري 11) وهكذا حتي الحديد وعدده الذري 26. اما العناصر الاثقل من ذلك فهي تتكون خلال انفجار النجوم فيما يسمي مستعرات عظمي. عندما تقترب نهايه عمر نجم كبير تنفجر وتبعثر كميات هائله
الجواب الصحيح : 7
الجواب المقترح : 26

Question : What is the atomic number of nitrogen?
Passage : Great directly. As for the rest of the elements with atomic masses heavier than hydrogen and helium, they originated and were "cooked" in the core of stars, where the temperature is high, exceeding 14 million degrees Celsius and sometimes reaching a billion degrees Celsius, depending on the mass of the star. In stars, the heavier elements are formed from hydrogen and helium through nuclear fusion. The elements are composed of lithium (atomic number 3), carbon (atomic number 6), nitrogen (atomic number 7), oxygen (atomic number 8), sodium (atomic number 11), and so on, even iron (atomic number 11). 26. As for the elements heavier than that, they are formed during the explosion of stars in what are called supernovae. When a large star approaches the end of its life, it explodes and scatters huge amounts of energytheir professional choices. She died
Actual Answer : 7
Prediction : 26

Figure 5: Examples of TyDi-based AraELECTRA's incorrect predictions on the ArTrivia development dataset.



Mercury (Q308)

smallest and closest planet to the sun in the Solar System

Sol I | Hermes | Sol b | Sol 1 | Planet Mercury | ☿

▾ In more languages

Configure

Language	Label	Description	Also known as
English	Mercury	smallest and closest planet to the sun in the Solar System	Sol I Hermes Sol b Sol 1 Planet Mercury ☿
Spanish	Mercurio	planeta del Sistema Solar, el más próximo en orden de distancias al Sol	
Traditional Chinese	水星	距離太陽最近的行星	
Chinese	水星	距離太陽最近的行星	安周星 能星 钩星 小正星 天機星 细爽星 辰星
Afrikaans	Mercurius	naaste planeet aan die son	
Amharic	ኦጣርድ	No description defined	
Aragonese	Mercurio	No description defined	
Old English	Wōden	se forma planēta þǽre sunnlican endebyrdnesse	
Angika	बुध ग्रह	No description defined	
Arabic	عطارد	أصغر كواكب المجموعة الشمسية و أقربها إلى الشمس	

Figure 6: The entity (Article Title) description in Wikipedia. This entity description can be accessed manually for any page in Arabic Wikipedia by clicking on tools (أدوات), then page information (معلومات الصفحة).