Arabic Question Answering Using AraBert and Information Retrieval

Raghad Jamhour
Department of Computer and Electrical Engineering
Birzeit University
raghadjamhour@gmail.com

Heba Fialah
Department of Computer and Electrical Engineering
Birzeit University
fualahheba@gmail.com

Dana Hafithah
Department of Computer and Electrical Engineering
Birzeit University
Dana.hafitha@gmail.com

***Abstract* - This project develops an extractive Arabic question-answering system based on AraBERT, fine-tuned using the Arabic-SQuAD dataset. The dataset is divided into training and testing subsets to evaluate model performance with standard metrics such as Exact Match and F1 score. To improve usability, a retrieval component using TF-IDF automatically selects the most relevant context paragraphs. This enables the model to accurately extract answers directly from Arabic text.**

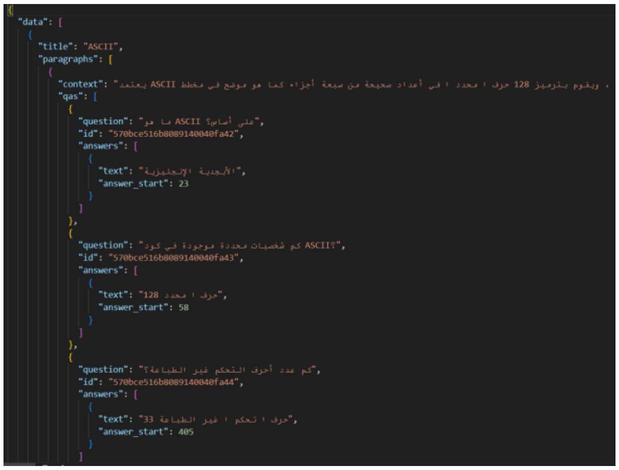
I. Introduction

In this project, we develop an extractive Arabic QA system by fine-tuning AraBERT on the Arabic-SQuAD dataset, a benchmark dataset specifically designed for extractive question answering in Arabic. The dataset is split into training and testing sets to ensure accurate evaluation of the model's performance using standard metrics such as Exact Match and F1 score.

To enhance the system's usability and accuracy, a TF-IDF was incorporated, a based retrieval component that automatically identifies the most relevant context paragraphs corresponding to a user's question. This allows the model to focus on a narrowed context, avoiding the need for users to provide context explicitly. The combination of TF-IDF retrieval with AraBERT's powerful language understanding enables precise extraction of answers directly from Arabic text.

This approach balances efficient retrieval with advanced deep learning techniques, demonstrating the effectiveness of modern Arabic language models for building practical and accurate QA applications.

II. Dataset



```
{
  "data": [
    {
      "title": "ASCII",
      "paragraphs": [
        {
          "context": ". ويقوم بترميز 128 حرف ا محدد ا في أعداد صحيحة من سبعة أجزاء كما هو موضح في مخطط ASCII يعتمد
          "qas": [
            {
              "question": "على أساس؟ ASCII ما هو",
              "id": "570bce516b8089140040fa42",
              "answers": [
                {
                  "text": "الأبجدية الإنجليزية",
                  "answer_start": 23
                }
              ]
            },
            {
              "question": "كم شخصيات محددة موجودة في كود ASCII؟",
              "id": "570bce516b8089140040fa43",
              "answers": [
                {
                  "text": "حرف ا محدد 128",
                  "answer_start": 58
                }
              ]
            },
            {
              "question": "كم عدد أحرف التحكم غير الطباعة؟",
              "id": "570bce516b8089140040fa44",
              "answers": [
                {
                  "text": "حرف ا تحكم ا غير الطباعة 33",
                  "answer_start": 405
                }
              ]
```

Fig. 1. Data Set format

The Arabic-SQuAD dataset serves as a foundational benchmark for extractive general question answering (QA) in Arabic. Its primary goal is to facilitate the development and evaluation of systems capable of reading an Arabic text passage and precisely identifying the answer to a given question as a contiguous span within that text.

A. Core Structure: Question-Context-Answer Triples

The Arabic-SQuAD dataset is organized into "question-context-answer triples." Each triple comprises:
A Question: A natural language query posed in Arabic.
A Context (Passage): An Arabic paragraph or longer text segment that is guaranteed to contain the answer to the accompanying question.
An Answer: The exact segment of text, or "span," extracted directly from the provided context that directly answers the question. This characteristic defines it as an extractive QA dataset, meaning answers are not generated summaries but direct excerpts.

B. Dataset Division for Training and Evaluation

For effective model development and accurate performance evaluation, the dataset is partitioned. The larger portion, approximately 80%, forms the training subset, used to teach the QA model how to map questions to answers within contexts. The remaining 20%, serves as the testing subset, crucial for evaluating the trained model's generalization capabilities on previously unseen data.

C. JSON Format Breakdown

Figure 1 shows the structured JSON format of the Arabic-SQuAD dataset. Key elements include:
"data": This is the top-level array, containing multiple distinct articles or documents.
Each document entry then includes a "title" (identifying the source article) and a "paragraphs" array.
Within the "paragraphs" array, each object contains a "context" field (the actual Arabic text passage) and a "qas" array (listing question-answer pairs relevant to that specific context). Each qas entry is composed of "question" (the Arabic query), unique "id",
and "answers" array, which can hold multiple valid answer spans for a single question.
Finally, each individual answer within the "answers" array specifies its "text" (the exact answer span) and its "answer_start" (the precise character-level starting index of that answer text within the context string).

This comprehensive and precise structure is fundamental for enabling the effective training and accurate evaluation of Arabic extractive QA models using AraBert. [1]

III. AraBERT Fine-Tuning for Extractive QA

This project leverages AraBERT, a pretrained BERT model specifically designed for Arabic NLP tasks, to perform extractive question answering. The model is fine-tuned on the Arabic-SQuAD training set, which consists of question-context-answer triples where the answer is a continuous span within the retrieved context. Fine-tuning involves adapting the pretrained weights to the task by training the model to predict the start and end positions of the answer span in the given context paragraph. This approach allows AraBERT to accurately extract answers from relevant texts.

The fine-tuning process utilizes Hugging Face's Transformers library and Trainer API, which simplifies model training and evaluation. The model is trained using the data with appropriate hyperparameters to optimize performance while preventing overfitting. After fine-tuning, the model's ability to pinpoint correct answer spans is tested against the reserved test set, ensuring robust evaluation of its extractive QA capabilities.

IV. TF-IDF Retrieval for Context Selection

The retrieval component was implemented as a crucial part of the Question Answering (QA) system. A module was developed to retrieve the most relevant paragraph from a corpus of paragraphs based on a question provided by the user. To accomplish this, the TF-IDF (Term Frequency-Inverse Document Frequency) model was used to transform each paragraph into a numerical vector that captures the importance of its words. These vectors were then compared with the question vector using cosine similarity, allowing the paragraph with the highest semantic relevance to be identified. Although the exact answer is not extracted at this stage, the relevant context is successfully narrowed down and passed to the QA model (AraBERT) for final answer extraction. This retrieval-first strategy is crucial; it makes the process easier for users, as they don't need to provide context [2][3]

V. Evaluation Metrics: Exact Match and F1 Score

The system's overall performance in extractive question answering was assessed by comparing its predicted answers against the gold-standard (actual) answers from the Arabic-SQuAD test set. Two widely-used metrics were employed: Exact Match (EM) and F1 score.

A. Exact Match (EM)

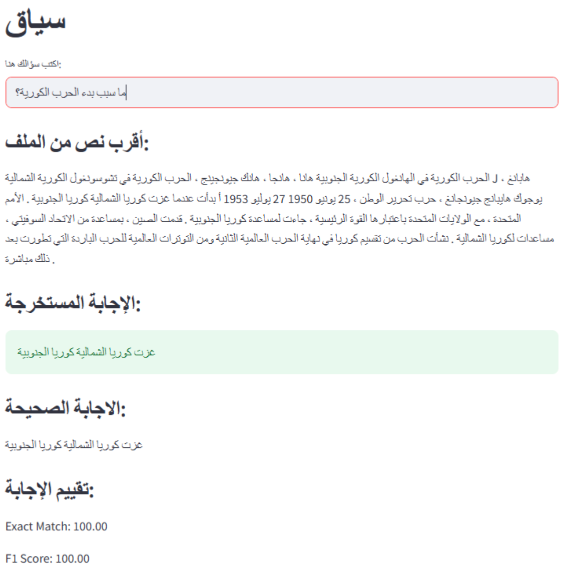
The Exact Match (EM) metric provides a stringent measure of accuracy, calculating the percentage of predictions that achieve a perfect, character-for-character match with the ground truth answers. A prediction scores 1 only if it's identical to a reference answer; otherwise, it's 0. While valuable for showing precise extraction, EM's strictness can understate performance if a semantically correct answer has minor lexical differences.

B. F1 Score

The F1 Score offers a more forgiving and representative evaluation, particularly useful when slight variations in answer spans are common. This metric measures the token-level overlap between predicted and true answers, balancing precision and recall, where Precision quantifies how many predicted tokens are correct and relevant and Recall measures how many true answer tokens were successfully captured by the prediction. The F1 score is the harmonic mean of precision and recall. A higher F1 score indicates effective identification of relevant tokens while minimizing irrelevant ones, providing a robust understanding of the model's true extraction capabilities, even when exact wording differs.

By employing both Exact Match and F1 Score, the evaluation framework provides a reliable measure of how accurately the model extracts correct information, covering both strict identity and semantic overlap. [4][5]

VI. Results and Analysis



سياق

اكتب سؤالك هنا:

ما سبب بدء الحرب الكورية؟

أقرب نص من الملف:

هابانغ ، ل الحرب الكورية في الهانغول الكورية الجنوبية هانا ، هانجا ، هانك جيونجينج ، الحرب الكورية في تشوسونغول الكورية الشمالية
يوجوك هايبانج جيونجانغ ، حرب تحرير الوطن ، 25 يونيو 1950 27 يوليو 1953 أ بدأت عندما عزت كوريا الشمالية كوريا الجنوبية . الأمم
المتحدة ، مع الولايات المتحدة باعتبارها القوة الرئيسية ، جاءت لمساعدة كوريا الجنوبية . قدمت الصين ، بمساعدة من الاتحاد السوفيتي ،
مساعدات لكوريا الشمالية . نشأت الحرب من تقسيم كوريا في نهاية الحرب العالمية الثانية ومن التوترات العالمية للحرب الباردة التي تطورت بعد
ذلك مباشرة .

الإجابة المستخرجة:

عزت كوريا الشمالية كوريا الجنوبية

الاجابة الصحيحة:

عزت كوريا الشمالية كوريا الجنوبية

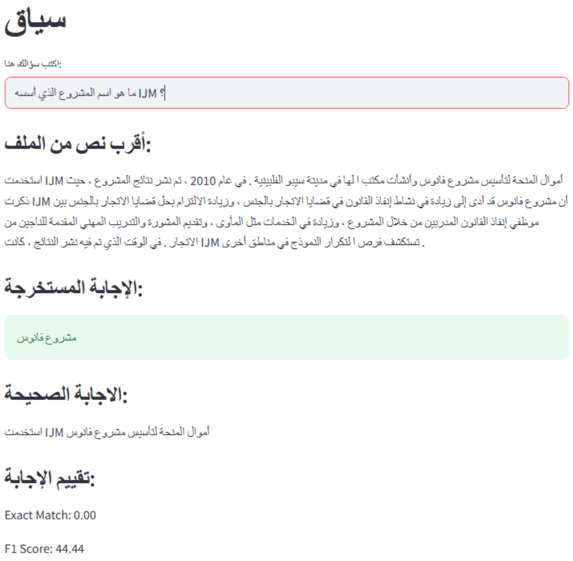
تقييم الإجابة:

Exact Match: 100.00

F1 Score: 100.00

Fig. 2. QA output with correct retrieval and answer extraction 1

Figure 2 shows the result of a question from the test set. The question is **"ما سبب بدء الحرب الكورية؟"** (What caused the start of the Korean War?). The system extracted the answer: "غزت كوريا الشمالية كوريا الجنوبية" (North Korea invaded South Korea), which matches the reference answer provided in the test data. As a result, the system achieved a perfect score 100% for both Exact Match and F1 Score, demonstrating that the predicted answer was fully correct according to the test set evaluation.



سياق

اكتب سؤالك هنا:

ما هو اسم المشروع الذي أسسه IJM ؟

أقرب نص من الملف:

أموال المنحة لتأسيس مشروع فانوس وأنشأت مكتب ا لها في مدينة سيبو الفلبينية . في عام 2010 ، تم نشر نتائج المشروع ، حيث IJM استخدمت
أن مشروع فانوس قد أدى إلى زيادة في نشاط إنفاذ القانون في قضايا الاتجار بالجنس ، وزيادة الالتزام بحل قضايا الاتجار بالجنس بين IJM ذكرت
موظفي إنفاذ القانون المدربين من خلال المشروع ، وزيادة في الخدمات مثل المأوى ، وتقديم المشورة والتدريب المهني المقدمة للناجين من
. تستكشف فرص ا لتكرار النموذج في مناطق أخرى IJM الاتجار . في الوقت الذي تم فيه نشر النتائج ، كانت

الإجابة المستخرجة:

مشروع فانوس

الاجابة الصحيحة:

أموال المنحة لتأسيس مشروع فانوس IJM استخدمت

تقييم الإجابة:

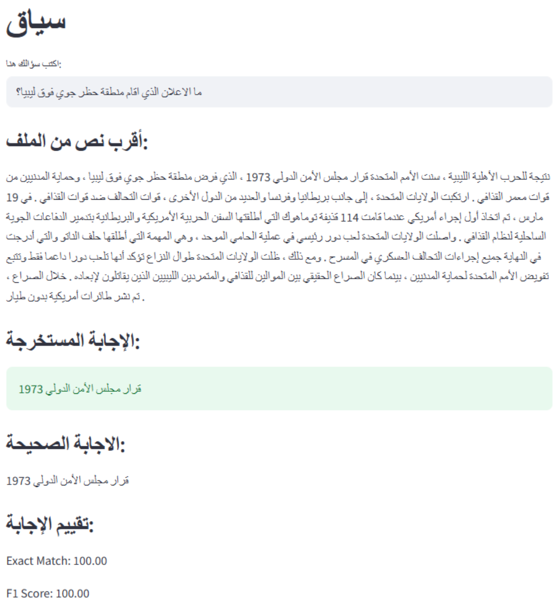
Exact Match: 0.00

F1 Score: 44.44

Fig. 3. QA output with correct retrieval and answer extraction 2

Figure 3 shows another result for a question from the test set. The question "ما هو اسم المشروع الذي أسسه IJM ؟" (What is the name of the project that IJM created?), required the system to identify a specific project name associated with "IJM". The system retrieved a relevant passage (context) from the documents that mentions the IJM project, then it extracted the answer, **(**Project Fanous) "مشروع فانوس"**.**

The answer is indeed the correct and sufficient answer to the question. However, because the reference answer in the test set is phrased differently " أموال المنحة لتأسيس مشروع فانوس IJM استخدمت" there is no exact match for the extracted answer. As a result, the system received a low evaluation: 0.00 for Exact Match and 44.44 for F1 Score, despite the predicted answer being accurate. This highlights a limitation in automatic evaluation metrics when semantically correct answers differ in wording from the reference.



سياق

:اكتب سؤالك هنا

ما الاعلان الذي اقام منطقة حظر جوي فوق ليبيا؟

أقرب نص من الملف:

نتيجة للحرب الأهلية الليبية ، سنت الأمم المتحدة قرار مجلس الأمن الدولي 1973 ، الذي فرض منطقة حظر جوي فوق ليبيا ، وحماية المدنيين من قوات معمر القذافي . ارتكبت الولايات المتحدة ، إلى جانب بريطانيا وفرنسا والعديد من الدول الأخرى ، قوات التحالف ضد قوات القذافي . في 19 مارس ، تم اتخاذ أول إجراء أمريكي عندما قامت 114 قذيفة توماهوك التي أطلقتها السفن الحربية الأمريكية والبريطانية بتدمير الدفاعات الجوية الساحلية لنظام القذافي . واصلت الولايات المتحدة لعب دور رئيسي في عملية الحامي الموحد ، وهي المهمة التي أطلقها حلف الناتو والتي أدرجت في النهاية جميع إجراءات التحالف العسكري في المسرح . ومع ذلك ، ظلت الولايات المتحدة طوال النزاع تؤكد أنها تلعب دورا داعما فقط وتتبع تفويض الأمم المتحدة لحماية المدنيين ، بينما كان الصراع الحقيقي بين الموالين للقذافي والمتمردين الليبيين الذين يقاتلون لإبعاده . خلال الصراع ، تم نشر طائرات أمريكية بدون طيار .

الإجابة المستخرجة:

قرار مجلس الأمن الدولي 1973

الاجابة الصحيحة:

قرار مجلس الأمن الدولي 1973

تقييم الإجابة:

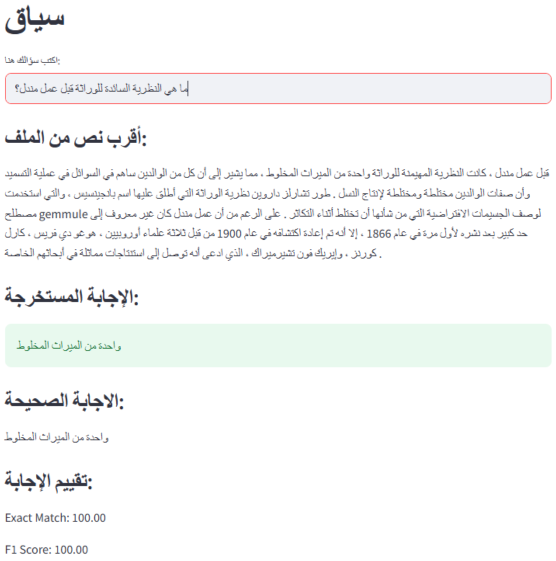
Exact Match: 100.00

F1 Score: 100.00

Fig. 4. QA output with correct retrieval and answer extraction 3

Figure 4 displays the result of another question from the test set. The question, " ما الاعلان الذي أقام منطقة حظر جوي فوق ليبيا؟ " (What announcement imposed a no-fly zone over Libya?), required the system to identify a specific political announcement from a historical context. The system successfully retrieved a relevant passage that includes detailed context about international military intervention in Libya and the associated UN resolution. From this retrieved context, the system correctly extracted the answer: "1973 قرار مجلس الأمن الدولي " (UN Security Council Resolution 1973).

This answer exactly matches the reference answer in the test set, resulting in perfect evaluation scores: 100% for both Exact Match and F1 Score. This example demonstrates the effectiveness of both the retrieval component and the answer extraction in identifying and presenting the correct response.



سياق

:اكتب سؤالك هنا

ما هي النظرية السائدة للوراثة قبل عمل مندل؟

أقرب نص من الملف:

قبل عمل مندل ، كانت النظرية المهيمنة للوراثة واحدة من الميراث المخلوط ، مما يشير إلى أن كل من الوالدين ساهم في السوائل في عملية التسميد وأن صفات الوالدين مختلطة ومختلطة لإنتاج النسل . طور تشارلز داروين نظرية الوراثة التي أطلق عليها اسم بانجينسيس ، والتي استخدمت لوصف الجسيمات الافتراضية التي من شأنها أن تختلط أثناء التكاثر . على الرغم من أن عمل مندل كان غير معروف إلى حد كبير بعد نشره لأول مرة في عام 1866 ، إلا أنه تم إعادة اكتشافه في عام 1900 من قبل ثلاثة علماء أوروبيين ، هوغو دي فريس ، كارل كورنز ، وإيريك فون تشيرميراك ، الذي ادعى أنه توصل إلى استنتاجات مماثلة في أبحاثهم الخاصة . gemmule مصطلح

الإجابة المستخرجة:

واحدة من الميراث المخلوط

الاجابة الصحيحة:

واحدة من الميراث المخلوط

تقييم الإجابة:

Exact Match: 100.00

F1 Score: 100.00

Fig. 5. QA output with correct retrieval and answer extraction 4

Figure 5 presents the result of a question from the test set focusing on historical theories of heredity. The posed question, "ما هي النظرية السائدة للوراثة قبل عمل مندل؟" (What was the prevailing theory of heredity before Mendel's work?), required the system to identify a specific scientific concept from a historical context.

The system successfully retrieved a highly relevant passage from the knowledge base. This passage provides a detailed explanation of pre-Mendelian theories. From this context, the system accurately extracted the answer: "واحدة من الميراث المختلط" (one of mixed inheritance).

This extracted answer precisely matches the reference answer in the test set, leading to perfect evaluation scores: 100% for both Exact Match and F1 Score. This particular example effectively demonstrates the system's robust capability in handling factual questions related to historical scientific concepts.

VII. Conclusion and Future Work

In this project, a complete Arabic Question Answering (QA) system was developed based on a general dataset. The data was converted into Squad format suitable for training. The AraBERT model was fine-tuned to answer Arabic questions, and a retrieval module using TF-IDF was built to find the most relevant paragraph for each question. The power of modern natural language processing techniques, when applied to Arabic was demonstrated by this project. Finally, both the retriever and the fine-tuned QA model were integrated into a simple but functional user interface, enabling end-to-end interaction. A user can ask a question in Arabic, the system retrieves the most relevant paragraph using TF-IDF, and the fine-tuned model extracts and displays the most likely answer along with its evaluation metrics.

This project shows the effectiveness of combining information retrieval techniques with transformer-based QA models for real-world applications in Arabic NLP. The system also establishes a strong foundation for future improvements. This includes expanding the dataset to cover more topics and scenarios, making it more versatile. We can also integrate advanced retrieval methods, particularly vector models, to better understand what questions mean and find more relevant information. Furthermore, improving our way of evaluating answers is important to properly score responses that are correct but not exact word-for-word matches.

REFERENCES

[1] i0xs0, “Arabic-SQuAD Dataset,” Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/i0xs0/Arabic-SQuAD

[2] C. Macdonald and J. Ounis, “Information Retrieval Models Tutorial,” University of Twente, Enschede, The Netherlands, Tech. Rep., 2008. [Online]. Available: https://ris.utwente.nl/ws/portalfiles/portal/5588097/IRModelsTutorial-draft.pdf

[3] S. Robertson, The Probabilistic Relevance Framework: BM25 and Beyond. Cambridge University Press, 2009. [Online]. Available: https://assets.cambridge.org/97805218/65715/frontmatter/9780521865715_frontmatter.pdf

[4] T. Tuyen, “Metrics for QA (Things to note),” Medium, Nov. 21, 2024. [Online]. Available: https://medium.com/@tuyen66tst/metrics-for-qa-things-to-note-1f9af6871ef4

[5] Cloudera Fast Forward Labs, “Evaluating QA: Metrics, Predictions, and the Null Response!” Jun. 9, 2020. [Online]. Available: https://qa.fastforwardlabs.com/Evaluating_BERT_on_SQuAD