Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 1 (2025) 104 - 110 e-ISSN: 2580-0760

Question Answering through Transfer Learning on Closed-Domain Educational Websites

Matiin Laugiwa Prawira Putra[1*], Evi Yulianti[2]
[1,2] Computer Science, Faculty of Computer Science, University of Indonesia, Depok, Indonesia.
[1]matiin.laugiwa@ui.ac.id, [2]evi.y@cs.ui.ac.id

Abstract

Navigating complex educational websites poses challenges for users looking for specific information. This research discusses the problem of efficient information search on closed-domain educational platforms, focusing on the Universitas Indonesia website. Leveraging Natural Language Processing (NLP), we explore the effectiveness of transfer learning models in Closed Domain Question Answering (QA). The performance of three BERT-based models, including IndoBERT, RoBERTa, and XLM-RoBERTa, are compared in transfer and non-transfer learning scenarios. Our result reveals that transfer learning significantly improves QA model performance. The models using a transfer learning scenario showed up to a 4.91\% improvement in the F-1 score against those using a non-transfer learning scenario. XLM-RoBERTa base outperforms all other models, achieving the F-1 score of 61.72\%. This study provides valuable insights into Indonesian-language NLP tasks, emphasizing the efficacy of transfer learning in improving closed-domain QA on educational websites. This research advances our understanding of effective information retrieval strategies, with implications for improving user experience and efficiency in accessing information from educational websites.

Keywords: NLP; Question Answering; Transfer Learning, Closed Domain; XLM-RoBERTa

How to Cite: M. Laugiwa and E. Yulianti, "Question Answering Question Answering through Transfer Learning on Closed-Domain Educational Websites", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 1, pp. 104 - 110, Feb. 2025.
DOI: https://doi.org/10.29207/resti.v9i1.6163

1. Introduction

Information on educational websites can be so complex and huge. A lot of information is displayed on the websites when users click on particular menus or fields. Some prospective students or the public in general who have a specific question to be answered by the information contained in the educational website can be confused and overwhelmed by the huge amount of information displayed there. This may end up with an inefficient process of information finding.

To tackle the issue of finding direct answers to users' questions, a specific task in the Natural Language Processing (NLP) field, such as Question Answering (QA), can be performed. In recent years, Natural language Processing (NLP) has become the center of attention in the world of artificial intelligence. Many researchers and practitioners want to improve the performance of NLP tasks [1], [2], [3], [4]. NLP is a multidisciplinary science between computer science, information engineering, artificial intelligence, and linguistics. NLP implementations are usually categorized into three categories, namely: text analytics, conversational systems, and machine translation [5], [6], and Covid-19 [7].

Research about QA is currently developing rapidly, most recent work on QA has explored the use of deep learning methods [8], [9], [10]. A big attention has been paid to transformer-based models, such as the bidirectional encoder representations from transformers (BERT) [11]. BERT was proposed by Devlin et al. to learn complex patterns in text from both sides (left and right) without requiring labels using some Transformer encoder layers. This way, once the BERT model is learned from a huge text, the pre-trained model can be easily fine-tuned to perform a variety of NLP tasks, such as QA. The fine-tuning process is performed by adding an output layer to the pre-trained model and re-training it using our specific dataset, so it does not require major changes to the main system's architecture.

The size of the closed domain QA dataset on education websites that are built in this study is relatively small, but that does not mean it has a small size from a data point of view, it can be calculated that the total information provided covers a large part of the target

Received: 03-11-2024 | Accepted: 13-01-2025 | Published Online: 08-02-2025

domain. Therefore, the use of this data to directly fine-tune the BERT-based models may be less effective. To improve the effectiveness of the fine-tuned QA models, we propose using a transfer-learning strategy to utilize the rich knowledge obtained from a high-resource QA dataset, such as SQUAD [12]. Three BERT-based models, such as IndoBERT [13], RoBERTa [14], and XLM-RoBERTa [15], are investigated to build our closed-domain QA system on educational websites.

To sum up, the main focus of this study is on the comparison between transfer-learning and non-transfer-learning models, and also on the comparison of three BERT-based variants, such as IndoBERT, RoBERTa, and XLM-RoBERTA in performing the closed domain QA task in educational websites. The applicability of this study is to help users find their required information on Indonesian educational websites easily, without having to manually browse the entire web on the Indonesian language educational web, more specifically, the Universitas Indonesia website https://ui.ac .id/.

2. Research Methods

2.1 Close Domain Question Answering

QA system is a system designed to answer questions given by users to obtain information. The input from the QA system itself is a list of information documents. The QA system will search for and find the start and end positions related to the document's answers. QA is divided into two, namely QA in closed domains and open domains. Closed-domain QA focuses on a particular problem or a specific domain.

One of the studies that discusses closed-domain QA is research by Devi et al. [8]. This research proposes a closed-domain question-answering system that utilizes semantics and ontology to improve answer selection. It includes components for domain resources, corpus, and MPS (Most Probable Sentence) components. In open-domain QA, researchers often use up-to-date sources such as Wikipedia as a source for searching for answers. One of the research related to open domain QA is QA research with a special topic, namely the COVID-19 virus, which can be called Dr.Qa, which has a focus on addressing questions about the COVID-19 virus [16].

2.2 Machine Reading Task for Question Answering

The task of Machine Reading for QA has been on the rise in recent years. This technique focuses on how the system can read and learn from documents that have been previously provided to the QA system. There are several datasets specifically created for tasks such as QA, such as SQuAD [12], CoQA [17], WikiQA [16], etc.

One of the research that has attracted public attention during the post-COVID-19 period is research initiated by [16] regarding Dr.QA. This research has been built based on open-domain QA, which uses a retriever-reader architecture. This architecture uses bigrams and TF-IDF to search for relevant documents from Wikipedia. The use of bidirectional RNN and bigram TF-IDF is a speciality of this research.

Several recent studies have been busy using the BERT algorithm to carry out NLP tasks, such as research by Archaya et al. [8] This research proposes a question-answering system that utilizes Named Entity Recognition (NER) and Bidirectional Encoder Representations from Transformers (BERT) for communication purposes. The system creates a customized dataset based on user answers, updates it using NER, and then employs BERT and CDQA to fetch the best answer for a given question. The system has potential applications in assisting people with speech impairments, integrating with communication devices for the disabled, and in academic fields. Spacy is used for NER, and CDQA architecture is used for question-answering.

Research related to machine reading with BERT is not only applied to international languages but there are several studies applied to the mother languages in each country, such as Vietnamese [18] Kazakh [19], etc. both of which use the development of the BERT method which is aligned with their respective mother languages.

In this research, we use a model architecture, namely a retriever-reader, in a machine reading task, which is similar to research by Rachmawati et al. [7], Alzubi et al. [10], and Yang et al. [20] However, this research uses a different algorithm and transfer learning approach.

2.3 Transfer Learning

The transfer learning approach is a method with the aim of improvising performance on tasks with limited resources. The transfer learning approach describes a learning scheme for a task with the target of better performance [21], [22].

Research on transfer learning is starting to become a new idea that is quite often used by current research. including research on QA by Kadam et al. [23] by utilizing context retrieval and a deep neural network with attention-based ranking to order documents based on their relevance. Transformer-based pre-trained models and transfer learning are employed to enhance accuracy. The next research is research by Alshammari et al. [24] where with transfer learning, the proposed TAQS model surpassed the performance of the state-of-the-art BiLSTM with SkipGram by a gain of 43.19% in accuracy. This demonstrates that transfer learning of BERT with BiLSTM improved the model's performance significantly. Another research that uses transfer learning in its model is QA research by Pudasaini et al. [23], where transfer learning has been shown to improve the evaluation results of QA systems in the context of biomedical research. Therefore, transfer learning has demonstrated its effectiveness in improving the evaluation results of QA systems in the biomedical domain. There is still much more research using transfer learning, and it can enhance the evaluation value of deep learning models in question-answering tasks.

2.4 Architecture

The architecture used in this research refers to previous research [7], [10], [20], [11] However, this research did not use a retriever in the process. Figure 1 shows an illustration of the architecture and reader with the dataset that we have collected. The reader we use is implemented using the BERT model and fine-tuned to the SQuAD dataset. After pre-training, the model was then trained again, but this time on the SQuAD dataset. In this phase, the model learns how to apply its general understanding of language. Fine-tuning involves improving the accuracy of a model by adjusting how it's trained [25], [26], [27].

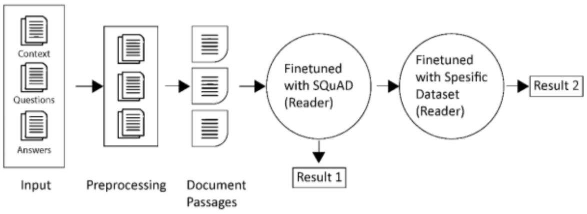


Figure 1. A flow that runs on the architectural system in this research.

The system operates through the following steps: input is received and segmented into passages, initial fine-tuning occurs using the SQuAD dataset to yield the initial outcome and evaluation, followed by a secondary fine-tuning process tailored to the specific task to achieve the subsequent outcome.

2.5 Reader System

The reader system receives passages from top-n documents received from the retriever system. Three models are used as readers in this research, IndoBERT-base, RoBERTa-base, and XLM RoBERTa-base, which is an enhancement of the initial model, namely BERT. BERT in research conducted by [11] is a model that is state-of-the-art to date. BERT is in two phases, namely pretraining and fine-tuning, as in Figure 2.

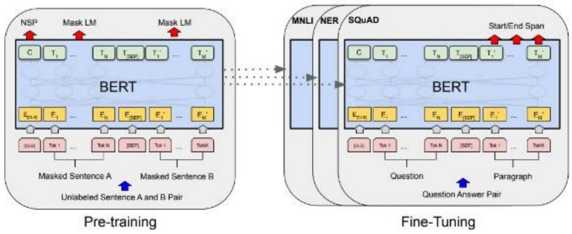


Figure 2. An original BERT flow that runs on the architectural system.

In the pretraining phase, BERT is asked to understand "What is Language? What is Context?" BERT learns from supervised tasks simultaneously, including Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). This is what makes BERT understand the context across different sentences and get a good language understanding [11] Finetuning allows BERT to create a future train for a Specific task by replacing the Fully Connected Output Layer with a fresh set of output layers that become the desired output answers [28].

This one is BERT, which was adapted into Indonesian whose research was started by Willie et al.[13] where the case used is uncased, which means there is no difference between lowercase and uppercase in the document. IndoBERT itself is a model that is an enhancement of the BERT model by pretraining the Indo4B dataset https://dumps.wikimedia.org/backup-index.html.2 where the data is utilized using bytes pair encoding (BPE) as a vocab generator. IndoBERT itself was pre-trained using TPUv3-8 in two phases: the first phase was trained at a maximum sequence length of 128, and the second phase was trained at a maximum sequence length of 512. Both of which used a batch size of 256 and a learning rate of 2e -5.

RoBERTa is a transformers-based artificial neural network model that utilizes a Robustly Optimized BERT Pretraining Approach. RoBERTa uses a span-based training method to predict the start and end logits of the answer indexes, resulting in improved performance on non-description questions with short answers [14]. RoBERTa itself, research by Liu et al. carried out various modifications and optimizations in terms of dynamic masking, FULL-SENTENCES without NSP loss, large mini-batches, and larger byte-level BPE. Apart from several factors that are optimized, RoBERTa also carries out enhancements by adding pre-trained data to larger data and longer training times compared to the BERT model to achieve the best value.

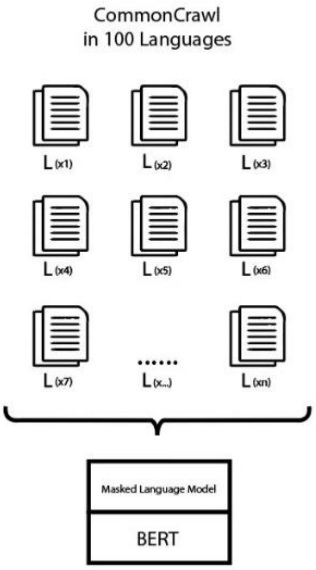


Figure 3. Here’s how the multilingual language process works.

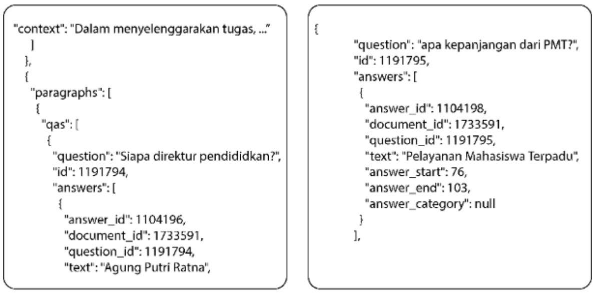
XLM-RoBERTa works by leveraging pre-training on a large amount of unlabeled corpus, finetuning with domain-specific data, and combining with traditional models to achieve high performance in legal document processing tasks [29]. XLM RoBERTa is not only trained using BERT on monolingual data but almost more than 100 languages are pre-trained or can be called Multilingual Models the illustration in Figure 3, but it doesn't stop there because the multilingual model can be enhanced again to cross-linguistic transfer learning, where training can be done in one language and transferred to another. That's why they are called Cross-Lingual Models (XLM), which are currently still state-of-the-art in most NLP tasks [15].

2.6 Dataset & Overlapping Statistics

The dataset used in this study is a dataset taken by researchers with two annotators from the University of Indonesia who worked on data annotation using an annotator application for two months. The data collected is closed domain data located on the website menu: https://ui.ac.id/. Apart from data collected separately, researchers also used SQuAD [12] to test transfer learning models with SQuAD and data on the closed domain, especially the website. Details of the dataset are presented in Table 1.

Table 1. Dataset Detail

Dataset	Dataset Detail		
	Total Document	Total Pairs	Topic
SQuAD	422	87.589	Open Domain
Edu-QA	126	2.692	Closed Domain



```
"context": "Dalam menyelenggarakan tugas, ..."
   ]
  },
  {
   "paragraphs": [
    {
     "qas": [
      {
       "question": "Siapa direktur pendididkan?",
       "id": 1191794,
       "answers": [
        {
         "answer_id": 1104196,
         "document_id": 1733591,
         "question_id": 1191794,
         "text": "Agung Putri Ratna",
```

```
{
       "question": "apa kepanjangan dari PMT?",
       "id": 1191795,
       "answers": [
        {
         "answer_id": 1104198,
         "document_id": 1733591,
         "question_id": 1191795,
         "text": "Pelayanan Mahasiswa Terpadu",
         "answer_start": 76,
         "answer_end": 103,
         "answer_category": null
        }
       ],
```

Figure 4. An example of annotation results that have been generated using annotation tools in SQuAD format.

The process of dividing the tasks of the two annotators in annotating Edu data taken from the web https://ui.ac.id/ with a total number of documents in the form of the context of 126 where each annotator creates questions and answers from each of the 63 documents with a total number of topic pairs of 2692. The two annotators also use annotation tools, namely on the web https://haystack.deepset.ai/ to create data structures in SQuAD format. An example of the results of data annotation using annotation tools is shown in Figure 4 where the results are in SQuAD format.

The researcher carried out data preprocessing overlapping data to see how two different annotators distributed data samples in answering questions in a context with the data form, as in Table 2. From the results of data processing in the format in Table 2, the researcher carried out an analysis using a distribution plot for answer data, especially those taken based on annotators.

Table 2. Content & Description Overlapping Preparations

Content	Description Overlapping Preparations
Paragraph Length (Char)	Total one paragraph per character
Paragraph Length (Word)	Total one paragraph per word
Question Length (Char)	Total One question per character
Question Length (Word)	Total one question per word
Answer Length (Char)	Total one answer per character
Answer Length (Word)	Total one answer per word

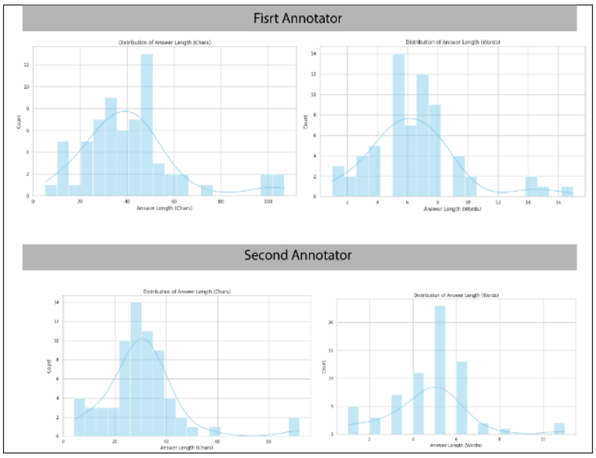


Figure 5. Distribution Plot, which shows the distribution of answers taken by each annotator.

The distribution plot in Figure 5 shows that the distribution of the two annotators is relatively the same. However, even so, we can see how the activities of annotators one and two are in answering questions. The second annotator answered the question relatively to the point on the target word, while the first annotator was different in that the answer was a sentence with more words. The model used to evaluate overlapping data is F1-score, Precision, and Recall value. It can be seen from Table 3 that with a score of almost 66%, there is no significant difference in the F1-score, Precision, and Recall value of each annotator, which means that the data annotation carried out has relatively the same tendency.

Table 3. Overlapping Evaluation

Metrics	Score
F1 Score	66.10
Precision	66.70
Recall	66.01

Indirectly, annotators also see the validity of overlapping sample data with the amount of data for each context, question, and answer sample, namely 10, 66, and 66. This overlapping data will also be considered later by researchers related to annotators who annotate the data to see whether there are differences. data collection techniques are used in different ways. The division of split data in this paper refers to research conducted by Nguyen et al. [29] with a division of training, evaluation, and testing with a proportion of 70%: 20%: 10%.

3. Result and Discussion

3.1 Pre-processing Hardware and Finetuning Process

This research uses several Python functions with the help of annotation tools to carry out pre-processing on Edu-QA data, such as removing stopwords, punctuation, lowercasing, and replacing multiple spaces. We use Google Colab Pro + as the computational machine for this research. this machine uses an additional 400 compute units for a total and accesses their highest memory machines.

The next process is a finetuning process, which is carried out on the reader using a pre-trained BERT model and its development. This treatment was also finetuned to the SQuAD version 1.1 data. Hyperparameter details can be seen in Table 4. The results of the model are by extracting and ranking the answers in the closed domain dataset Edu-QA, using a transfer learning approach process that has been trained on a much larger dataset in the hope of increasing the level of evaluation.

Table 4. Hyper Parameters for Reader

Hyperparameters	Details
Training Batch Size	4
Learning Rate	3e-5
Train Epoch	10
Optimizer	AdamW
Max Seq Length	384
Max Query Length	64
Gradient Accumulations Steps	3
Docs Stride	128

In the evaluation section, the metric that will be used in this research is the F1-Score and Exact Match to determine whether the model results can approach a more general ground truth answer due to the lack of Indonesian language sources. F1-Score and Exact Match can be suitable metrics for this research.

3.2 Experiment Result

The result shows us the performance of each model in handling the closed domain QA task case without being double fine-tuned as in the research by Rachmawati and Yulianti [7]. The difference between the IndoBERT-base model and RoBERTa-base and from its development. XLM RoBERTta-base itself is an enhancement of RoBERTta-base. Basically, this comparison is obtained with the same 10 epochs. There is a possibility that the model can make subtle improvements when the epoch value is increased until the evaluation and loss model starts to stagnate.

The next result, Table 5 is a model with double finetuning as was done by Yang et al. [30] in their research, which carried out double finetuning to carry out reader enhancement. This time, each model experienced an increase with an average increase in the models before and after the double finetuning of 4.8%. The highest difference is the IndoBERT-base reader with IndoBERT-base-SQuAD, which almost reaches 7.3% this indicates that double finetuned is quite effective in increasing reader evaluation for this QA task. Even the smallest difference was experienced by the RoBERTa-base and RoBERTa-base-SQuAD readers, which only touched the largest of 3.10%. This shows that double fine-tuning can increase the evaluation value of the reader and model, but the level of effectiveness is different.

Table 5. Experiment Result

Models	Evaluation Metrics		
	F1-Score	*Exact Match*	*Scenario*
IndoBERT-base	29.62	2.64	Non-Transfer Learning
RoBERTa-base	53.75	3.10	
XLM RoBERTa-base	56.79	8.00	
IndoBERT-base-SQuAD	33.33	4.90	Transfer Learning
RoBERTa-base-SQuAD	56.79	6.07	
XLM RoBERTa-base-SQuAD	61.72	11.97	

Table 5 shows us the exact match score from both with and without double fine-tuning. The without double fine-tuned model has an average of 4.58%, an Exact match in all of the enhancement BERT models. The Exact match score with the double fine-tuned model has increased slightly with an average score of 7.64%. This value shows enhancement from the double fine-tuned model. and the highest increase is the XLM RoBERTa-base-SQuAD model with a 3.97% increase.

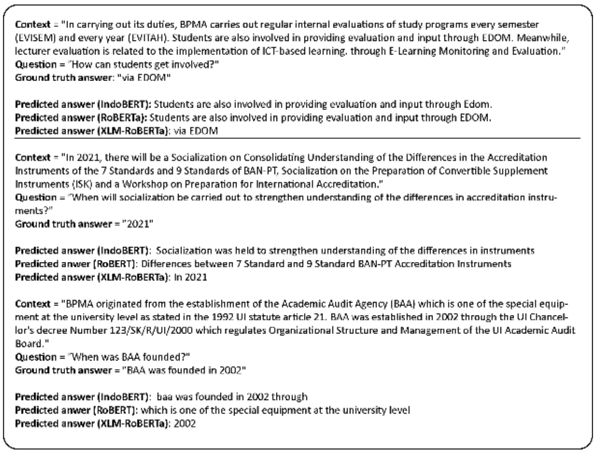
From the results of the comparison model we can see that the F1-Score value of the IndoBERT-base-SQuAD model has a final value of 33.33% with an Exact match of 4.90%, then the RoBERTa-base-SQuAD model gets an F1-Score of 56.79% and an Exact match of 6.07%, the final model is XLM-RoBERTa-base-SQuAD with an F1-Score of 61.72% and an Exact Match of 11.97%. The largest evaluation value is owned by XLM RoBERTa-base-SQuAD, with an F1-Score value obtained from 10 epochs of 61.72% and an Exact match of 11.97%. This is in line with several previous studies that state that the XLM RoBERTa model is still superior at this time in the cross-lingual generative process of Indonesian as mentioned in the research of Conneau et al. [15] . The results of this experiment are also the end of the development and verification of the model, but enhancements will continue to be carried out to get higher evaluation scores, especially for QA tasks.

3.3 Qualitative Analysis

The results of the qualitative analysis, shown in Figure 6, were carefully translated from Indonesian to English for a thorough evaluation. Initially, all models are tasked with generating answers, revealing the ability of the XLM-RoBERTa model to provide precise ground truth answers without unnecessary words. This accuracy underscores the model's skill in capturing the essence of the information in question.

However, as the scenario progresses, subtle differences emerge. Although the XLM-RoBERTa model consistently approaches the ground truth answer, there are several deviations that occur, especially in the loss of certain affixes. Although small, these differences highlight areas that need improvement in the future. In contrast, other models offer a broader range of responses, indicating differences in interpretation skills and indicating potential for improvement.

In the third scenario, all models deviate from the actual answer, but both the XLM-RoBERTa and IndoBERT models succeed in producing answers that are close to the ground truth answer. The ability to converge towards the desired results reflects a distinct understanding of the contextual nuances in the query, demonstrating the overall efficacy of this model. These insights underscore the complexity of language understanding and generation tasks. Although the XLM-RoBERTa model consistently performs well, small differences emphasize the need for continued improvement.



Context = "In carrying out its duties, BPMA carries out regular internal evaluations of study programs every semester (EVISEM) and every year (EVITAH). Students are also involved in providing evaluation and input through EDOM. Meanwhile, lecturer evaluation is related to the implementation of ICT-based learning. through E-Learning Monitoring and Evaluation."
Question = "How can students get involved?"
Ground truth answer: "via EDOM"

Predicted answer (IndoBERT): Students are also involved in providing evaluation and input through Edom.
Predicted answer (RoBERTa): Students are also involved in providing evaluation and input through EDOM.
Predicted answer (XLM-RoBERTa): via EDOM

Context = "In 2021, there will be a Socialization on Consolidating Understanding of the Differences in the Accreditation Instruments of the 7 Standards and 9 Standards of BAN-PT, Socialization on the Preparation of Convertible Supplement Instruments (ISK) and a Workshop on Preparation for International Accreditation."
Question = "When will socialization be carried out to strengthen understanding of the differences in accreditation instruments?"
Ground truth answer = "2021"

Predicted answer (IndoBERT): Socialization was held to strengthen understanding of the differences in instruments
Predicted anwer (RoBERT): Differences between 7 Standard and 9 Standard BAN-PT Accreditation Instruments
Predicted answer (XLM-RoBERTa): In 2021

Context = "BPMA originated from the establishment of the Academic Audit Agency (BAA) which is one of the special equipment at the university level as stated in the 1992 UI statute article 21. BAA was established in 2002 through the UI Chancellor's decree Number 123/SK/R/UI/2000 which regulates Organizational Structure and Management of the UI Academic Audit Board."
Question = "When was BAA founded?"
Ground truth answer = "BAA was founded in 2002"

Predicted answer (IndoBERT): baa was founded in 2002 through
Predicted anwer (RoBERT): which is one of the special equipment at the university level
Predicted answer (XLM-RoBERTa): 2002

Figure 6. Show us How Qualitative Example Text

4. Conclusions

This paper explores the use of transfer learning strategies to perform closed-domain question answering (QA) on educational websites. Furthermore, the effectiveness of three BERT-based models, such as IndoBERT, RoBERTa, and XLM RoBERTa, using transfer-learning strategies was also evaluated in this study. This study also produced a new QA dataset on educational websites, consisting of 1,000 questions and answers with appropriate context. Our results show that the model using the transfer learning strategy improves the F1 Score of the non-transfer learning model by up to 4.91%. The highest performance was obtained by XLM RoBERTa which obtained an F-1 score of 61.72% and an exact match score of 11.97%. The results are also shown by qualitative analysis with three different scenarios using different transfer learning models such as IndoBERT, RoBERTa, and XLM-RoBERTa where each scenario of the XLM-RoBERTa model can produce the same answer and is close to the ground truth answer, in contrast to the other two models IndoBERT and RoBERTa which answer questions with answers that are too broad. The analysis mentions qualitatively the XLM-model RoBERTa excels. Overall, the main contribution of this research is the investigation of transfer learning to improve the effectiveness of closed-domain QA on educational websites.

Acknowledgements

I would like to thank Andar Al-afgani for his valuable contribution as an annotator in this research. Andar's dedication and expertise have played an important role in the success of our research.

5. Future Work

For future work, we are considering developing closed domain QA on educational websites such as https://ui.ac.id/ using a weighted approach and the next version of SQuAD data transfer learning. We also want to modify how extraction is done in the retriever and make the processing able to produce better evaluation values.

References

[1] T. Shao, Y. Guo, H. Chen, and Z. Hao, “Transformer-Based Neural Network for Answer Selection in Question Answering,” *IEEE Access*, vol. 7, pp. 26146–26156, 2019, doi: 10.1109/ACCESS.2019.2900753.

[2] T. Wolf *et al.*, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[3] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, “Transfer Learning in Natural Language Processing”.

[4] D. W. Otter, J. R. Medina, and J. K. Kalita, “A Survey of the Usages of Deep Learning for Natural Language Processing,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[5] B. D. Shivahare, A. K. Singh, N. Uppal, A. Rizwan, V. S. Vaathsav, and S. Suman, “Survey Paper: Study of Natural Language Processing and its Recent Applications,” in *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India: IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/CISCT55310.2022.10046440.

[6] R. Sarikaya, G. E. Hinton, and A. Deoras, “Application of Deep Belief Networks for Natural Language Understanding,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 4, pp. 778–784, Apr. 2014, doi: 10.1109/TASLP.2014.2303296.

[7] N. Rachmawati and E. Yulianti, “Transfer Learning for Closed Domain Question Answering in COVID-19,” *IJACSA*, vol. 13, no. 12, 2022, doi: 10.14569/IJACSA.2022.0131234.

[8] S. Acharya, K. Sornalakshmi, B. Paul, and A. Singh, “Question Answering System using NLP and BERT,” in *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India: IEEE, Oct. 2022, pp. 925–929. doi: 10.1109/ICOSEC54921.2022.9952050.

[9] H. Le, L.-M. Nguyen, J. Ni, and S. Okada, “Constructing a Closed-Domain Question Answering System with Generative Language Models,” in *2023 15th International Conference on*

Knowledge and Systems Engineering (KSE), Hanoi, Vietnam: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/KSE59128.2023.10299437.

[10] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, "COBERT: COVID-19 Question Answering System Using BERT," *Arab J Sci Eng*, vol. 48, no. 8, pp. 11003–11013, Aug. 2023, doi: 10.1007/s13369-021-05810-5.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Oct. 10, 2016, *arXiv*: arXiv:1606.05250. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1606.05250

[13] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Oct. 08, 2020, *arXiv*: arXiv:2009.05387. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/2009.05387

[14] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1907.11692

[15] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," Apr. 07, 2020, *arXiv*: arXiv:1911.02116. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1911.02116

[16] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," Apr. 27, 2017, *arXiv*: arXiv:1704.00051. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1704.00051

[17] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge," Mar. 29, 2019, *arXiv*: arXiv:1808.07042. Accessed: May 05, 2024. [Online]. Available: http://arxiv.org/abs/1808.07042

[18] N. T. M. Trang and M. Shcherbakov, "Vietnamese Question Answering System f rom Multilingual BERT Models to Monolingual BERT Model," in *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India: IEEE, Dec. 2020, pp. 201–206. doi: 10.1109/SMART50582.2020.9337155.

[19] M. Shymbayev and Y. Alimzhanov, "Extractive Question Answering for Kazakh Language," in *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan: IEEE, May 2023, pp. 401–405. doi: 10.1109/SIST58284.2023.10223508.

[20] W. Yang *et al.*, "End-to-End Open-Domain Question Answering with BERTserini," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 72–77. doi: 10.18653/v1/N19-4013.

[21] A. Akdemir, "Research on Task Discovery for Transfer Learning in Deep Neural Networks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Online: Association for Computational Linguistics, 2020, pp. 33–41. doi: 10.18653/v1/2020.acl-srw.6.

[22] A. Akdemir and T. Shibuya, "Transfer Learning for Biomedical Question Answering".

[23] N. Kadam and M. A. Kumar, "Multiple Choice Question Answering Using Attention Based Ranking and Transfer Learning," in *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India: IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/TENSYMP54529.2022.9864511.

[24] W. T. Alshammari and S. AlHumoud, "TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT With BiLSTM," *IEEE Access*, vol. 10, pp. 91509–91523, 2022, doi: 10.1109/ACCESS.2022.3198955.

[25] S. S. Lakkimsetty, S. V. Latchireddy, S. M. Lakkoju, G. R. Manukonda, and R. V. V. M. Krishna, "Fine-Tuned Transformer Models for Question Answering," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10307046.

[26] M. A. Ateeq, S. Tiun, H. Abdelhaq, and N. Rahhal, "Arabic Narrative Question Answering (QA) Using Transformer Models," *IEEE Access*, vol. 12, pp. 2760–2777, 2024, doi: 10.1109/ACCESS.2023.3348410.

[27] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J.-R. Wen, "Complex Knowledge Base Question Answering: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11196–11215, Nov. 2023, doi: 10.1109/TKDE.2022.3223858.

[28] H.-H. Hsu and N.-F. Huang, "Xiao-Shih: A Self-Enriched Question Answering Bot With Machine Learning on Chinese-Based MOOCs," *IEEE Trans. Learning Technol.*, vol. 15, no. 2, pp. 223–237, Apr. 2022, doi: 10.1109/TLT.2022.3162572.

[29] H. N. Van, D. Nguyen, P. M. Nguyen, and M. L. Nguyen, "Miko Team: Deep Learning Approach for Legal Question Answering in ALQAC 2022," Nov. 03, 2022, *arXiv*: arXiv:2211.02200. Accessed: May 06, 2024. [Online]. Available: http://arxiv.org/abs/2211.02200

[30] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering," Apr. 14, 2019, *arXiv*: arXiv:1904.06652. Accessed: May 06, 2024. [Online]. Available: http://arxiv.org/abs/1904.06652