Applied Intelligence (2023) 53:27560–27570
https://doi.org/10.1007/s10489-023-04751-w

Enhancing yes/no question answering with weak supervision via extractive question answering

Dimitris Dimitriadis[1] · **Grigorios Tsoumakas**[1]

Accepted: 30 May 2023 / Published online: 14 September 2023

Abstract
The effectiveness of natural language processing models relies on various factors, including the architecture, number of parameters, data used during training, and the tasks they were trained on. Recent studies indicate that models pre-trained on large corpora and fine-tuned on task-specific datasets, covering multiple tasks, can generate remarkable results across various benchmarks. We propose a new approach based on a straightforward hypothesis: improving model performance on a target task by considering other artificial tasks defined on the same training dataset. By doing so, the model can gain further insights into the training dataset and attain a greater understanding, improving efficiency on the target task. This approach differs from others that consider multiple pre-existing tasks on different datasets. We validate this hypothesis by focusing on the problem of answering yes/no questions and introducing a multi-task model that outputs a span of the reference text, serving as evidence for answering the question. The task of span extraction is an artificial one, designed to benefit the performance of the model answering yes/no questions. We acquire weak supervision for these spans, by using a pre-trained extractive question answering model, dispensing the need for costly human annotation. Our experiments, using modern transformer-based language models, demonstrate that this method outperforms the standard approach of training models to answer yes/no questions. Although the primary objective was to enhance the performance of the model in answering yes/no questions, it was discovered that span texts are a significant source of information. These spans, derived from the question reference texts, provided valuable insights for the users to better comprehend the answers to the questions. The model's improved accuracy in answering yes/no questions, coupled with the supplementary information provided by the span texts, led to a more comprehensive and informative user experience.

Keywords Question answering · Yes/no question answering · Extractive question answering · Transformers

1 Introduction

Transformer-based models have achieved astonishing results in several natural language processing (NLP) tasks. For example, the T5 model [1] significantly outperformed previous state-of-the-art models in several benchmarks, including GLUE (General Language Understanding Evaluation), a collection of resources for training, evaluating and analyzing natural language understanding systems [2], and SQuAD (Stanford Question Answering Dataset), a dataset for training and evaluating extractive question answering (QA) systems [3, 4]. Such models comprise millions of parameters (11 billion in T5), which are optimized in a self-supervised fashion using huge corpora during a pre-training phase [5]. They are then typically fine-tuned to particular downstream NLP tasks.

Besides increasing model parameters and input data, which has adverse effects on the environment due to the increased computational costs [6, 7], another important avenue towards improving such models is employing multi-task learning, during either the pre-training or the fine-tuning process. In BERT [8] for example, pre-training the model on both masked language modeling and next sentence prediction achieves better results than pre-training it on masked language modeling alone. In Multi-hop QA, multi-task learning has been used for both predicting the answer of the question and for extracting evidence [9]. In the context of open

✉ Dimitris Dimitriadis
dndimitri@csd.auth.gr

Grigorios Tsoumakas
greg@csd.auth.gr

[1] School of Informatics, Aristotle University of Thessaloniki, Panepistimioupoli, Thessaloniki 54124, Greece

book QA, a model built on top of RoBERTa [10] jointly ranks passages and their sentences using a complex training objective that incorporates consistency and similarity constraints [11], managing to improve the results on the task of selecting question-relevant information from a large corpus.

With this in mind, we propose a simple, yet effective, method to improve the fine-tuning of transformers in answering yes/no questions. On top of the standard supervision, which is the correct answer of a yes/no question, we add a span of the reference text that serves as evidence for the correct answer. We extract weak supervision for such a type of span in an unsupervised manner, without any involvement of human experts, using an extractive QA model.

In summary, the main contributions of this paper are:

1. A new perspective on dealing with the yes/no QA task. Instead of focusing entirely on the binary supervision concerning the answer, we propose a multi-task learning approach for extracting simultaneously the span of the reference text that can be considered as evidence for the correct answer.
2. An approach for automatically constructing yes/no QA datasets enriched with answer related reference spans, by weakly annotating them via an extractive QA model.
3. An empirical study showing that the multi-task approach gives performance improvements on yes/no QA, along with corresponding supporting evidence for each particular answer.
4. The learning models and dataset have been made available for public use, granting individuals the opportunity to utilize them for various purposes. [1].

The rest of this article is organized as follows. Section 2 reviews related work in yes/no QA . Section 3 presents our method. Section 4 describes the experimental setup and presents the results and the qualitative analysis. Finally, Section 5 concludes this work and proposes future research directions.

2 Related work

Our approach is closely aligned with recent studies on yes/no QA, specifically those that leverage transformer-based models and multi-task learning. In this section, we present an overview of the existing approaches for answering yes/no questions, with a specific focus on methodologies utilizing the BoolQ dataset [12]. This dataset is unique as it consists solely of yes/no questions and several of the current approaches have been extensively tested on it. Moreover, we emphasize the effectiveness of Large Language Models (LLMs) [13] in tackling this specific task. Lastly, we outline the distinguishing characteristics that set our approach apart from the other methods referenced below.

Early transformer-based approaches played a pivotal role in advancing the field of yes/no QA, particularly in the context of the BoolQ dataset [12]. These approaches predominantly relied on transformer models such as BERT, RoBERTa, and ALBERT [14], which represented a significant breakthrough at the time. The dataset's creators utilized BERT and conducted various experiments with similar QA tasks to enhance the accuracy of the yes/no QA model. The findings revealed that the transferred knowledge from Multi-Genre Natural Language Inference (MultiNLI) [15], along with the unsupervised pre-training in BERT, had the most significant impact. Similarly, the SuperGLUE team [16] utilized BERT and BERT++, a BERT variation that adopts the STILTs style [17] of transfer learning, to experiment with the dataset. RoBERTa, a highly optimized version of BERT, achieved an 87.1% accuracy on the dataset when fine-tuned solely on it without incorporating other tasks. The DeBERTa model [18], which employs a disentangled attention mechanism and an enhanced mask decoder, achieved significantly better results (90.4% acc.) compared to other approaches. Additionally, the ALBERT XXLarge model, with 223M parameters, also attained high performance (84.8% accuracy) solely through fine-tuning on the task itself, while being pre-trained on masked language modeling and sentence ordering prediction tasks.

Our approach stands apart from the previously mentioned methods in terms of how we train the learning models to address the yes/no QA problem. While the BERT-based methods employ transfer learning and multi-task learning, utilizing various pre-existing tasks and datasets, they either rely on different datasets or solely on the BoolQ dataset. Similarly, we also employ transfer learning by utilizing pre-trained language models. However, for the BoolQ dataset, we introduce a unique artificial task to enhance the performance of yes/no QA. This approach sets us apart from pre-training methods like ALBERT, where the model is trained on both tasks without explicitly aiming to improve the performance of one task over the other. The training objective of ALBERT is to create a model that can be adapted to multiple downstream tasks, which differs from our specific objective.

The BoolQ dataset has been subjected to testing with various LLM models, yielding a wide range of outcomes in terms of performance accuracy. These models are designed to offer general-purpose solutions, rather than being specifically tailored to the task itself, with the intention of addressing a broad spectrum of NLP tasks. Many of these models have placed emphasis on reformulating input examples. One notable approach is Pattern-exploiting training (PET) [19], which utilizes patterns and rephrases input examples as cloze-

[1] https://github.com/dndimitri/EnhancingYesNoQuestionAnsweringwithWeakSupervisionviaExtractiveQuestionAnswering

style phrases. PET has undergone thorough evaluation on the BoolQ dataset, demonstrating promising results. When combined with the ALBERT base model, PET achieved an accuracy rate of 81.2%. Additionally, the iterative variant of PET attained an accuracy of 79.1%. Another variant called ADAPET [20] focused on few-shot learning, without relying on unlabeled data, and achieved an accuracy of 80% on the same dataset.

In the quest for advancements in text-to-text learning, a unified framework was proposed by Google [1], demonstrating state-of-the-art results across various tasks, including yes/no QA. In the BoolQ dataset, this framework achieved an impressive accuracy of 91.2%. Another model called FLAN [21], with 137B parameters and trained using instruction tune on 60 NLP datasets, achieved an accuracy of 82.9% specifically on the BoolQ dataset. Furthermore, the EFL [22] model reformulated NLP tasks into entailment ones, resulting in an accuracy of 86% when considering all tasks and 73.9% accuracy in a few-shot setting, focusing on eight specific tasks within the BoolQ dataset.

Our approach distinguishes itself from the mentioned methods in several key aspects, with a particular emphasis on computational costs. While many existing large language models consist of billions of parameters and are trained on extensive datasets for multiple tasks, our approach demonstrates superior performance on the BoolQ dataset compared to several models. For example, our approach outperforms BloombergGPT [23] with 50B parameters (74.59% accuracy), which is considered a state-of-the-art model for the finance domain. It also surpasses GPT-NeoX [24] with 20B parameters trained on Pile [25] (46.36% accuracy), Hyena [26] with a subquadratic drop-in replacement for attentions (51.8% zero-shot learning and 56% few-shot learning), various variations of the OPT model proposed by Meta AI, N-Grammer [27] which augments n-grams constructed from a discrete latent representation, NEO [28] which applies an "ASK ME ANYTHING PROMPTING" strategy, AlexaTM [29] with 20B parameters by Amazon utilizing a multi-lingual seq2seq model (69.44% accuracy), T5-small by Google (76.4% accuracy), and several variations of LLaMA [30], including those with 7B and 13B parameters. Moreover, our approach adopts a unique perspective by refraining from emphasizing the reformulation of input examples during the fine-tuning process, and it also does not rely on a large number of tasks.

In conclusion, none of the aforementioned approaches aim to improve the performance of a yes/no QA model by leveraging a task specifically designed for this purpose and constructing a new task using the same dataset instead of relying on pre-existing tasks. Furthermore, approaches that utilize reformulation of input examples aim to provide a general solution to NLP problems without specifically focusing on enhancing the performance of a task by leveraging other tasks.

3 Our approach

This section presents our approach, starting with the way we build a weakly labeled dataset enriched with evidence texts by leveraging an extractive QA model to answer yes/no questions. Next, we mention the architecture of our multi-task model and discuss the training and inference processes.

3.1 Obtaining weak supervision for evidence spans

Our multi-task model expects a dataset with questions and reference texts, accompanied by answers and evidence spans. In typical yes/no QA datasets, questions and reference texts are accompanied by answers only. Acquiring evidence spans would require the involvement of human annotators. To avoid this cost, we propose employing a pre-trained extractive QA model instead, in order to obtain weak supervision of evidence spans. Given a WH-question and a reference text, an extractive QA model will output a span of the reference text that it considers as the answer to the question. We assume that when applied to a yes/no question, such a model will identify a span that could serve as evidence for answering the question.

To our knowledge, no studies have yet tested the effectiveness of extractive QA in extracting evidence texts for yes/no questions. Nevertheless, this model remains the most appropriate tool for the task at hand, given its training for a similar task (extracting the answers themselves). A notable advantage of this model is its capacity for automated annotation of the training dataset, thereby eliminating the need for human annotators and the associated time and cost expenses. This benefit is particularly significant because it not only minimizes the involvement of human experts but also enables scaling of the process to larger datasets that would otherwise require extensive manual annotation.

In Fig. 1, we illustrate this concept using an example from the BoolQ dataset, involving the question: "*Is escape from gringotts ride a roller coaster?*" and the reference text "*Harry Potter and the Escape from Gringotts is an indoor steel roller coaster at Universal Studios Florida, a theme park located within the Universal Orlando Resort. Similar to dark rides, the roller coaster utilizes special effects in a controlled-lighting environment and also employs motion-based 3-D projection of both animation and live-action sequences to enhance the experience. The ride, which is themed to the Gringotts Wizarding Bank, became the flagship attraction for the expanded Wizarding World of Harry Potter when it opened on July 8, 2014.*". The answer to this question is

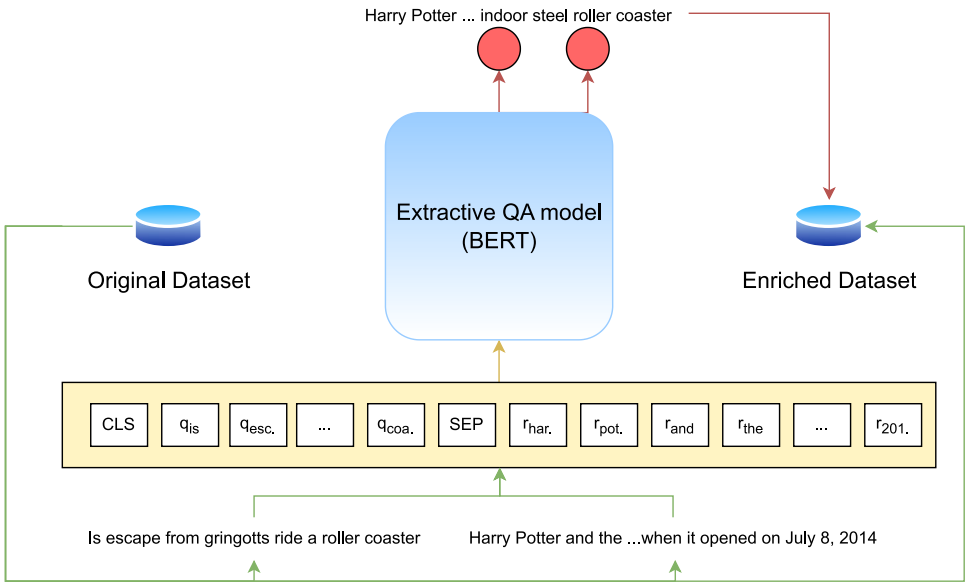

Fig. 1 Constructing the enriched dataset leveraging an Extractive QA model based on BERT

apparently yes. When we give this question and reference text to an extractive QA model, it outputs the reference text span "*Harry Potter and the Escape from Gringotts is an indoor steel roller coaster.*". This span contains evidence for correctly answering the question. We append the evidence span to the question, reference text and answer to create an enriched data instance. Repeating this process for all question and reference text pairs of a typical yes/no QA dataset, we construct an enriched dataset that can be subsequently used by our multi-task model.

We employ a typical extractive QA model, where a standard pre-trained language model is extended with two special vectors, a span-start embedding S and a span-end embedding E, which will be learned during fine-tuning [31]. Given question q and reference text r, we obtain a span-start probability $P_s(i \mid q, r)$ for each token i by computing the dot product between S and the output representation of i, followed by a softmax over all tokens in r. The same process is followed for estimating the span-end probabilities $P_e(i \mid q, r)$. The model outputs the text span maximizing the product of the probabilities of the start and end positions.

3.2 Multi-task model

We define a multi-task learning problem, where a model is responsible for predicting both the answer to a yes/no question and a span of the reference text that can be considered as evidence for the answer of the question. We hypothesize that a yes/no QA model equipped with such knowledge can infer the correct answer easier.

We extend the architecture of the extractive QA model discussed in the previous section, by adding a linear layer for predicting the answer to the question (yes/no). Since we are working with transformer-based models, the linear layer gets as input the multi-dimensional vector of the special token indicating that the input data will be used in text classification (e.g. [CLS] in BERT).

During training, we utilize an enriched dataset, synthesized as described in Section 3.1. The questions and reference texts pass to the model described in this section, while the answers to the questions and the evidence spans are considered as targets. Figure 2 shows the training process considering one instance.

We define a training objective that considers both tasks. The negative log-likelihood of the correct answer for each input is used for the yes/no QA task ($L_{yes/no}$), while the average of the negative sum of log-likelihoods of the correct start and end positions for each input for the evidence extraction task ($L_{extractive}$). The final loss we are using (L) is defined as the sum of these two losses:

$$L = L_{yes/no} + L_{extractive}$$

We defined $L_{extractive}$ as the average, instead of the sum, of its constituents to avoid the bias of the second task in the total loss. Through this objective, the model will also consider the weakly labeled evidence text in the process of learning the correct answers to the questions.

During inference, the model is fed with a question and a reference text and predicts the answer and evidence span. The latter can be ignored, since the main reason of its existence is to help the model learn the correct answer. However, in Section 4.3, we present examples, where the evidence span serves indeed as a valid explanation for the corresponding answer.

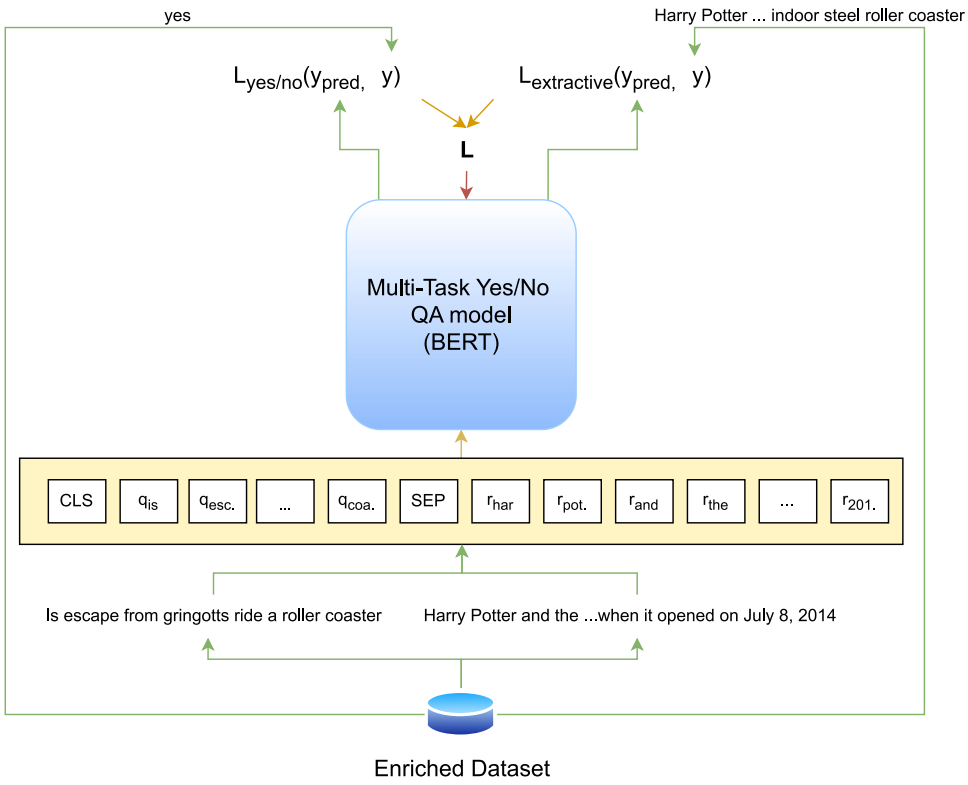

Fig. 2 Training the Multi-task yes/no QA model based on BERT considering an enriched dataset

4 Experimental design and findings

This section commences with a description of the experimental setup utilized to assess the effectiveness of our approach. Subsequently, we present the outcomes of our method in comparison to robust baselines and alternative methods. Ultimately, we provide a qualitative analysis that illustrates the actual impact of our approach through real-world examples.

4.1 Experimental setup

Our work relies mainly on torch version 1.11 (provided by the PyTorch team [32]) and transformers version 4.17 (provided by the Hugging Face team [33]), two libraries that are used for building neural network models with strong GPU acceleration and for leveraging pre-built state-of-the-art neural network models respectively.

For the extractive QA task we leveraged BERT$_{large}$ model pre-trained on uncased English texts and fine-tuned on SQuAD 1.0[2] and RoBERTa$_{base}$ fine-tuned on SQuAD 2.0[3]. SQuAD 1.0 [3] contains more than 100K questions posed by crowdworkers on a set of Wikipedia articles, where the answer is a segment of text from the corresponding reading passage. SQuAD 2.0 [4] enriches the collection of SQuAD 1.0 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. We selected those two models from a collection of several others, because they include details about how they have been built, enhancing the replicability of this study.

For the multi-task model, we build on top of these models the yes/no QA task and evidence text extraction task respectively. We set the maximum sequence length to be 256 and truncate all tokens beyond the maximum context size of the model. We tuned the learning rate (LR) ($1e{-}5$, $2e{-}5$, $3e{-}5$) and batch size (B) (4, 8, 16, 24) for 5 different seeds and 10 epochs with the AdamW optimizer.

We used the BoolQ dataset [12] for evaluating our approach. BoolQ comprises a collection of yes/no questions gathered from anonymized, aggregated queries to the Google search engine, selecting only questions that can be answered by a Wikipedia page. Human annotators select the most relevant passage from the corresponding page and specify whether the answer is yes or no. Each instance of the dataset is thus a triple consisting of a question, a passage, and a yes/no answer. The dataset has been split into train, development, and test sets with 9,427, 3,270, and 3,245 instances, respectively.

To estimate the performance of our approach in the yes/no QA task, we present results of the hyper-parameter tuning process in the BoolQ development set. As our method has been implemented on top of the BERT and RoBERTa models, we compare it with the same language models as baselines.

[2] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

[3] https://huggingface.co/deepset/roberta-base-squad2

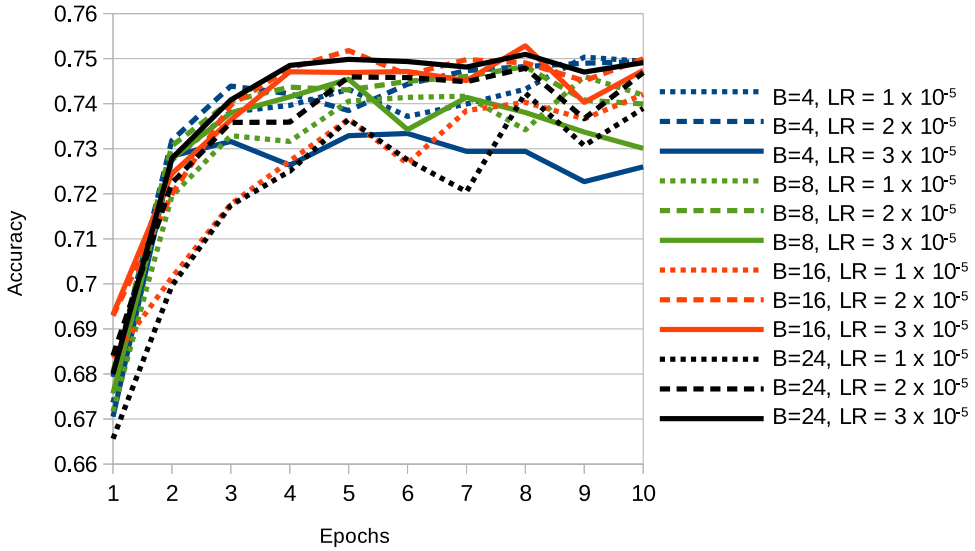

Fig. 3 Evaluation on BoolQ validation set with different batch sizes (different colors) and learning rates (different line styles) for 10 epochs using the BERT base model

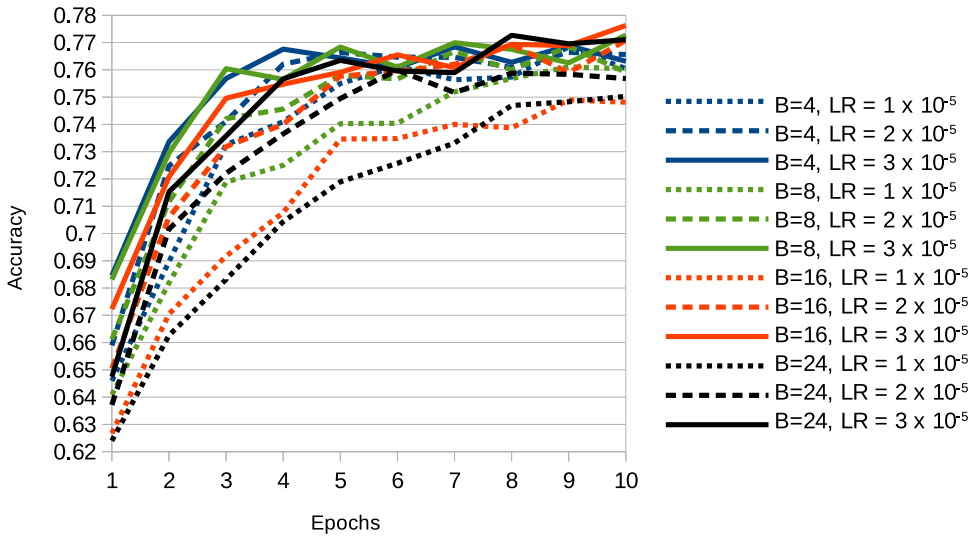

Fig. 4 Evaluation on BoolQ validation set with different batch sizes and learning rates for 10 epochs using the BERT base model with the proposed method

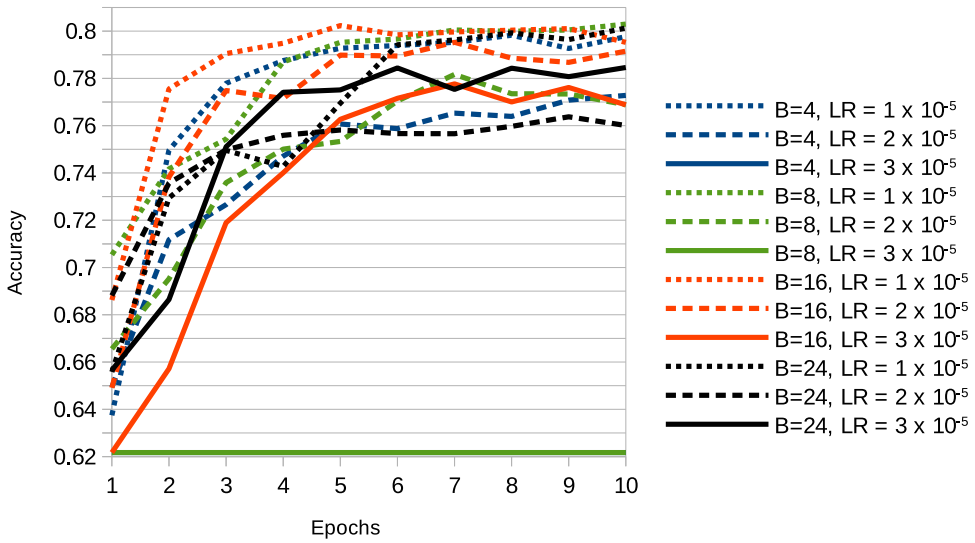

Fig. 5 Evaluation on BoolQ validation set with different batch sizes and learning rates for 10 epochs using the RoBERTa base model

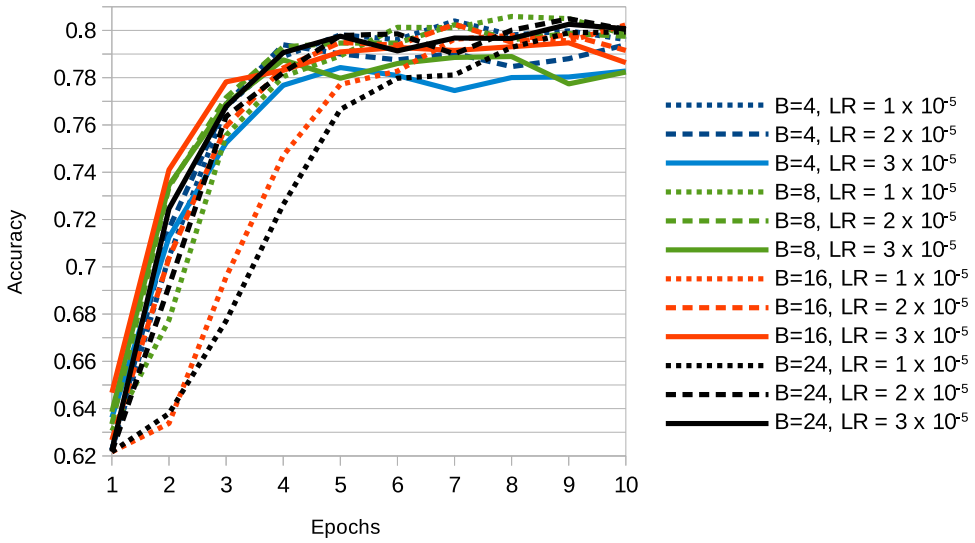

Fig. 6 Evaluation on BoolQ validation set with different batch sizes and learning rates for 10 epochs using the RoBERTa base model with the proposed method

The baseline models are trained considering only the original BoolQ dataset, while our method gets advantage of the enriched dataset constructed using the extractive QA models. Next, we present the results on the BoolQ test set and compare our models with the model of the BoolQ creators.

4.2 Results

Figure 3 shows the hyper-parameter tuning performance of BERT base without considering our method. As we can see, the baseline model does not exceed 76% accuracy in the average of five runs. The two hyper-parameters affect significantly the performance of the learning model. For example, when B is 4 and LR is $3e-5$, we observe that the highest accuracy is 73.34% in the 6_{th} epoch. However, with the same LR and $B = 24$ the model accuracy is approximately 75%. For $B = 16$ and $LR = 3e-5$, we have the highest average accuracy (75.28%). In general, higher values of LR seem to improve the performance of the BERT model for larger batches.

When applied to BERT, our method appears to significantly improve the performance for all values of the hyper-parameters (Fig. 4). Although in the first epoch most of the models have poor performance considering most of the questions to belong to one class (i.e. the model answers yes most of times), in later epochs, our method outperforms the baseline models. The accuracy is higher than 76% in most cases, while the best accuracy is 77.62% for $B = 16$ and $LR = 3e-5$. We observe that higher LR values lead to better performance. We also notice that with our method, BERT needs more time to increase the model's accuracy in most cases. This was expected since the task that it has to solve is more difficult. The model has to both answer the question and extract an evidence span, while it is also trained from noisy data produced automatically by the extractive QA model.

The RoBERTa baseline model is unstable for different hyper-parameters (Fig. 5). The model cannot learn from data when $LR = 3e-5$ and $B \in \{4, 8\}$ meaning that it predicts for each question the same answer. In contrast to BERT, RoBERTa benefits from lower LR values, while the parameter selection seems to affect much more the performance of the model. The average accuracy does not exceed 81%, while there are cases where the accuracy is lower than 77%. The highest average accuracy is 80.29% for batch size 8 and LR $1e-5$. The best RoBERTA baseline models outperform the BERT models in most cases.

In contrast to the RoBERTa baseline models, the models fine-tuned with our method are stable since different hyper-parameters do not hurt the overall performance (Fig. 6). Furthermore, our RoBERTa model is not affected by the initial random seeds. As with our BERT model, we have poor performance in the first epoch. However, in later steps, the accuracy is higher than 77% and is not significantly affected

Table 1 Final Results on BoolQ test set selecting the models with the highest accuracy during parameter tuning

Models	B	LR	Epoch	Val. Acc.	Val Diff.	Test Acc.	Test Diff
BERTBase	16	3e-05	7	0.7691		0.740	
BERTBase (ours)	4	3e-05	3	0.7887	+0.0196	0.798	+0.058
RoBERTaBase	16	1e-05	4	0.8162		0.794	
RoBERTaBase (ours)	24	3e-05	8	0.8168	+ 0.0006	0.799	+ 0.005

Table 2 Results of the average accuracy of five runs for our method, the baseline and the performance of BoolQ dataset creators (C)

Model	Val Acc.
BERT Base	0.7528
BERT Large (C)	0.7690
BERT Base (ours)	0.7763
RoBERTa Base	0.8029
RoBERTa Base (ours)	0.8059

by the selection of the hyper-parameters. The highest average accuracy is 80.59% which is +0.03% higher from the baseline model for batch size 8 and LR $1e-5$ after eight epochs.

We performed paired t-tests to assess the significance of the differences between our methods and the corresponding baselines. Each variable in the t-test represents the best average accuracy achieved across all epochs for different combinations of batch sizes and learning rates.

Our analysis revealed significant improvements in accuracy when comparing our model based on BERT to the baseline. Our model exhibited a statistically significant enhancement in accuracy ($M = 0.7629$) compared to baseline ($M = 0.7426$), with a mean difference of 0.0203 ($t = -6.333$, $p < 0.001$). These results demonstrate the noteworthy improvement achieved by the modifications implemented in our model.

Similarly, for our model based on RoBERTa, the paired t-test indicated a significant enhancement in the measured variable ($M = 0.7988$) compared to the baseline ($M = 0.7603$), with a mean difference of -0.0385 ($t = -2.214$, $p = 0.049$). These findings confirm that the adjustments made in our model led to a significant improvement in accuracy.

Besides the different outcomes from the models that adopt our method and the baselines during hyper-parameter tuning in the development set, there are also differences in the performance in the unseen test set of BoolQ. In Table 1, we summarize these results. We selected the model with the highest accuracy for each of the models parameterized above. Our approach overcomes the base models in all cases in both the validation and test sets. Furthermore, our method affects much more the BERT base model than RoBERTa. Our BERT model overcomes the baseline RoBERTa model in the test set while our RoBERTa model has the best test set accuracy overall (79.9%).

Finally, to compare our results with those of the BoolQ dataset creators, we report in Table 2 the best average accuracy of five runs based on the validation set tuning process that we followed above. Our BERT and RoBERTa models overcome the BERT large model, which has significantly much more parameters than our models.

To conclude, the results show that the proposed method is not significantly affected by hyperparameter tuning in contrast to the baselines. In the average of five runs, our BERT model overcomes the BERT large model of the BoolQ creators. Our method does not overcome their model fine-tuned on MultiNLI dataset. Our RoBERTa base model is close enough to the results of validation set (82.20% vs 81.68%) and to the results of the test set (80.43% vs 79.9%) with much less data and parameters for training.

After conducting our analysis on the validation and test set, we have successfully determined the computational time for the examined examples. Our findings indicate that the transformer-based model for question answering and evidence text extraction typically takes between 10 and 25 milliseconds to complete. These measurements were obtained using the T4 GPU within the Google Colab infrastructure.

4.3 Qualitative analysis

In this section, we present some examples from the BoolQ development dataset showing the effectiveness of our method

Table 3 Pairs of Questions (Q) and Evidences (E) from the BoolQ development dataset

Expected Outcome
Q1: does ethanol take more energy make that produces?
E1: returns from 8 to 9 units of energy for each unit expended
Q2: is pain experienced in a missing body part or paralyzed area
E2: Phantom pain sensations are described as perceptions that an individual experiences relating to a limb or an organ that is not physically part of the body
Q3: is harry potter and the escape from gringotts a roller coaster ride
E3: Harry Potter and the Escape from Gringotts is an indoor steel roller coaster
Q4: is there a word with q without u
E4: the only modern-English words that contain Q not followed by U
Q5: is there a play off for third place in the world cup
E5: A third place play-off was also played between the two losing teams of the semi-finals

Table 4 Examples of Questions (Q), Evidences (E) and Misleading Evidences ($\neg E$) from the BoolQ development dataset

Not Expected Outcome
Q1: is house tax and property tax are same
E1: Property tax or 'house tax' is a local tax on buildings
$\neg E1$: It resembles the US-type wealth tax and differs from the excise-type UK rate
Q2: is barq's root beer a pepsi product
E2: is owned by the Barq family but bottled by the Coca-Cola Company
$\neg E2$: Its brand of root beer is notable for having caffeine
Q3: is the show bloodline based on a true story
E3: -
$\neg E3$: Bloodline was announced in October 2014 as part of a partnership between Netflix and Sony Pictures Television, representing Netflix's first major deal with a major film studio for a television series

in finding the evidence text that is relevant to the given question when applied to BERT base.

In Table 3, we present 6 pairs of questions and span texts that give hints to the readers about the truthfulness of the question. For example, the first question asking for ethanol (Q1) is accompanied with a large reference text. However, the learning model extracted a very specific piece of information (E1) that indirectly answers the question. In the fifth example (Q4), the span text (E4) is incomplete, since it does not mention which are those words with q without u. However, the word "only" indicates the existence of such words even though they are not mentioned in the span text. These examples show that the model that is getting advantage of both tasks can provide grounds to the reader for the decision of the answer. Furthermore, the evidence text gives extra information to the reader, which is more valuable than the answer itself most of the times.

Next, we show some unexpected outcomes from the model (Table 4). In the first example (Q1), we hypothesize that the model finds the word "differs" and extracts that span text ($\neg E1$). However, the more relevant text is the one that mentions both two terms as local tax on buildings (E1). In the second example (Q2), the model extracts information about barq's root beer ($\neg E2$) but not the connection between it and the pepsi product (E2). Finally, in the last example (Q3), we do not expect a span text from the given reference text since there is not such information available. If we trained the extractive QA model considering as input the yes/no questions and as output the corresponding evidence texts then we may overcome such false positive evidence texts. However, building such a dataset containing the evidence texts is hard since it is a time-consuming operation and it is not always clear what part of a reference text should be considered as appropriate evidence.

5 Conclusions & future work

This paper presented a method for dealing with the yes/no QA task. In contrast to previous approaches, this method takes advantage of a pre-trained extractive QA model to guide the learning of a model to answer yes/no questions. The results are better compared to those of conventional yes/no QA models. It is also important to note that not only the accuracy has been improved by the proposed method, but also the model extracts useful parts of texts, as presented in Section 4.3. Consequently, the benefits of this method are two-fold. On one hand, the model's performance is better since it gets the advantage of multi-task learning. On the other hand, the extracted span text gives a hint to the reader to understand the output of the model. Finally, the most beneficial advantage of our method is the fact that no expert is needed. Consequently, this method can be easily scaled to larger yes/no QA datasets.

In this study, we have assumed that the evidence is a consecutive part of text contained in the reference text. This, however, is not always the case. Evidences can be scattered throughout the reference text or even found in multiple reference texts. An interesting future direction of this work would be to address such a multiple evidence scenario. Another extension of this work is dealing with situations where the evidence text is generated automatically, for example by a generative QA model, but is not a part of a reference text. It would be interesting to investigate whether we could use a multi-task model that simultaneously predicts the evidence text and answers the yes/no question in this case. Finally, it would be interesting to investigate whether we could obtain weak supervision for other types of tasks, besides extractive QA, and whether this could further boost the accuracy in yes/no QA.

Acknowledgements This work was supported by computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET S.A.) in the National HPC facility - ARIS - under project ID pa181002-NEBULA.

Funding Open access funding provided by HEAL-Link Greece.

Data Availability All data generated or analysed during this study are included in this article.

Declarations

Funding and/or Conflicts of Interests/Competing interests No Funding and no conflicts of Interests/Competing interests.

References

1. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21:1–67
2. Wang A, Singh A, Michael J, Hill F, Levy O (2018) Bowman SR Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461
3. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250
4. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable. arXiv:1806.03822
5. Wang H, Li J, Wu H, Hovy E, Sun Y (2022) Pre-trained language models and their applications. Eng. https://doi.org/10.1016/j.eng.2022.04.024
6. Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in nlp. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650
7. Strubell E, Ganesh A, McCallum A (2020) Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 13693–13696
8. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1(Mlm), 4171–4186. arXiv:1810.04805
9. Nishida K, Nishida K, Nagata M, Otsuka A, Saito I, Asano H, Tomita J (2019) Answering while summarizing: Multi-task learning for multihop qa with evidence extraction. arXiv:1905.08511
10. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
11. Luo M, Chen S, Baral, C (2021) A simple approach to jointly rank passages and select relevant sentences in the obqa context. arXiv:2109.10497
12. Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K (2019) BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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), pp 2924–2936. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1300. https://aclanthology.org/N19-1300
13. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, et al. (2023) A survey of large language models. arXiv:2303.18223
14. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations. https://openreview.net/pdf?id=H1eA7AEtvS
15. Paramasivam A, Nirmala SJ (2022) A survey on textual entailment based question answering. J King Saud University - Comput Inf Sci 34(10, Part B):9644–9653. https://doi.org/10.1016/j.jksuci.2021.11.017
16. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019) Superglue: A stickier benchmark for generalpurpose language understanding systems. Adv Neural Inf Process Syst 32
17. Phang, J., Févry, T., Bowman, S.R (2018) Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv:1811.01088
18. He P, Liu X, Gao J, Chen W (2021) Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations. https://openreview.net/pdf?id=XPZIaotutsD
19. Tam D, Menon RR, Bansal M, Srivastava S, Raffel C (2021) Improving and simplifying pattern exploiting training. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 4980–4991
20. Tam D, R Menon R, Bansal M, Srivastava S, Raffel C (2021) Improving and simplifying pattern exploiting training. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4980–4991. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.407. https://aclanthology.org/2021.emnlp-main.407
21. Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2022) Finetuned language models are zero-shot learners. In: International Conference on Learning Representations
22. Wang S, Fang H, Khabsa M, Mao H, Ma H (2021) Entailment as few-shot learner. arXiv:2104.14690
23. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G (2023) Bloomberggpt: A large language model for finance. arXiv:2303.17564
24. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, He H, Leahy C, McDonell K, Phang J, et al. (2022) Gpt-neox-20b: An open-source autoregressive language model. In: Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models, pp 95–136
25. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, et al. (2020) The pile: An 800gb dataset of diverse text for language modeling. arXiv:2101.00027

26. Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, Bengio Y, Ermon S, Ré C (2023) Hyena hierarchy: Towards larger convolutional language models. arXiv:2302.10866
27. Roy A, Anil R, Lai G, Lee B, Zhao J, Zhang S, Wang S, Zhang Y, Wu S, Swavely R, et al. (2022) N-grammer: Augmenting transformers with latent n-grams. arXiv:2207.06366
28. Arora S, Narayan A, Chen MF, Orr LJ, Guha N, Bhatia K, Chami I, Sala F, Ré C (2022) Ask me anything: A simple strategy for prompting language models. arXiv:2210.02441
29. Soltan S, Ananthakrishnan S, FitzGerald J, Gupta R, Hamza W, Khan H, Peris C, Rawls S, Rosenbaum A, Rumshisky A, et al. (2022) Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. arXiv:2208.01448
30. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al. (2023) Llama: Open and efficient foundation language models. arXiv:2302.13971
31. Jurafsky D, Martin JH (2022) Speech and Language Processing (3rd ed.draft). unpublished
32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019) Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32
33. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al. (2019) Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771

Grigorios Tsoumakas is an Associate Professor of Machine Learning and Knowledge Discovery at the School of Informatics of the Aristotle University of Thessaloniki (AUTH) in Greece. He received a degree in Computer Science from AUTH in 1999, an MSc in Artificial Intelligence from the University of Edinburgh, United Kingdom, in 2000 and a PhD in Computer Science from AUTH in 2005. His research expertise focuses on supervised learning techniques (ensemble methods, multi-target prediction) and natural language processing (semantic indexing, keyphrase extraction, summarization). He has published more than 150 research papers and according to Google Scholar he has more than 17k citations and an h-index of 50. Dr. Tsoumakas is a senior member of ACM and IEEE. His honors include receiving the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 10-Year Test of Time Award in 2017 and the Marco Ramoni best paper award at the 19th International Conference on Artificial Intelligence in Medicine (AIME 2021). He is an advocate of applied research that matters and has worked as a machine learning engineer, researcher and consultant in several national, international, and private sector funded R&D projects. In February 2019 he co-founded Medoid AI, a spin-off company of AUTH developing AI solutions based on machine learning technology.

Dimitris Dimitriadis holds a Ph.D. in Philosophy from the School of Informatics at the Faculty of Sciences. He obtained his Bachelor's degree in Informatics from Aristotle University of Thessaloniki (AUTH) in Greece and completed his MSc in Information Systems, also at AUTH. He successfully defended his Ph.D. in 2022. His research interests primarily lie in the fields of Machine Learning and Natural Language Processing, with a particular focus on Question Answering systems. Over the years, he has participated in the Biomedical Semantic Indexing and Question Answering Challenge, earning multiple awards for his contributions. Additionally, he has actively engaged in various European projects and collaborated with numerous researchers worldwide. In February 2017, he co-founded Contia, an IT company specializing in the development of web-based applications.