New State-of-the-Art for Question Answering on Portuguese SQuAD v1.1

E. H. M. da Silva[1], J. Laterza[2], T. P. Faleiros[3]

Universidade de Brasília, Brazil
[1]eric.hans@gmail.com, [2]laterza22@hotmail.com, [3]thiagodepaulo@unb.br

Abstract. In the Natural Language Processing field (NLP), Machine Reading Comprehension (MRC), which involves teaching computers to read a text and understand its meaning, has been a major research goal over the last few decades. A natural way to evaluate whether a computer can fully understand a piece of text or, in other words, test a machine's reading comprehension, is to require it to answer questions about the text. In this sense, Question Answering (QA) has received increasing attention among NLP tasks. For this study, we fine-tuned BERT Portuguese language models (BERTimbau Base and BERTimbau Large) on SQuAD-BR - the SQuAD v.1.1 dataset translated to Portuguese by the Deep Learning Brazil group - for Extractive QA task, in order to achieve better performance than other existing models trained on the dataset. As a result, we accomplished our objective, establishing the new state-of-the-art on SQuAD-BR dataset using BERTimbau Large fine-tuned model.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: bert, extractive question answering, fine-tune, language model, squad v1.1 portuguese, transfer learning

1. INTRODUCTION

Natural Language Processing (NLP) studies the capacity and limitations of a machine to comprehend human language. In this field, the Machine Reading Comprehension (MRC), which involves teaching computers to read a text and understand its meaning, has been a major research goal over the last few decades [Patel et al. 2020; Wadhwa et al. 2018; Zeng et al. 2020]. The MRC is a challenging task, since it requires both understanding of natural language and knowledge about the world [Patel et al. 2020; Rajpurkar et al. 2016; Zeng et al. 2020]. Nevertheless, with the advances in Deep Learning (DL) techniques and increasing accessibility of large-scale datasets, it is now possible to train a model to read and understand a language and perform specific NLP tasks, such as text classification, machine translation, named entity recognition and question answering [Devlin et al. 2019; Pranesh et al. 2020; Wadhwa et al. 2018; Yamada et al. 2020; Zeng et al. 2020].

A natural way to evaluate whether a computer can fully understand a piece of text or, in other words, test a machine's reading comprehension, is to require it to answer questions about the text. In this sense, Question Answering (QA) has received increasing attention among NLP tasks [Mayeesha et al. 2021; Wadhwa et al. 2018; Zeng et al. 2020]. QA is concerned with building systems that automatically respond to questions posed by humans in a natural language [Mayeesha et al. 2021; Patel et al. 2020; Rajpurkar et al. 2016; Zeng et al. 2020]. The Extractive QA task, specifically, shares the same goal of the typical machine reading comprehension task, which can be formulated as the following supervised learning problem: given a passage of text and a question as input, select a contiguous span of text in the passage as the answer. Consequently, training models to perform Extractive QA task requires labeled datasets, most commonly human-constructed (e.g., via crowdsourcing) [Cambazoglu et al. 2021; Joshi et al. 2020; Liu et al. 2021; Rajpurkar et al. 2016; Yamada et al. 2020; Zeng et al. 2020].

The driver behind progress of question answering research has been the availability of high-quality large datasets and release of models performing well on these datasets [Cambazoglu et al. 2021; Liu et al. 2021; Pranesh et al. 2020; Rajpurkar et al. 2016; Zeng et al. 2020]. An example is the Stanford Question Answering Dataset (SQuAD), the most widely used MRC dataset, that contains more than 100k questions generated by crowd-workers, contributing to the emergence of state-of-the-art models like ELMo, BERT and XLNet [Rajpurkar et al. 2016; Zeng et al. 2020]. Nonetheless, the abundance of benchmark datasets available in the QA field is limited to few languages, with a clear predominance of English and Chinese, as shown in [Cambazoglu et al. 2021]. Thus, due to lack of natural high-quality reading comprehension datasets, similar progress has not been achieved in other languages, such as Portuguese [Cambazoglu et al. 2021; Mayeesha et al. 2021].

On the other hand, advances in language representation using DL have made it viable to transfer the learned internal states of large pre-trained Language Models (LM). In other words, it is now possible to adapt models pre-trained on large linguistic corpora to downstream tasks like QA [Devlin et al. 2019; Pranesh et al. 2020; Souza et al. 2020]. This transfer learning approach is highly beneficial when labeled data is scarce, making pre-trained LMs valuable resources specially for languages with few annotated datasets [Mayeesha et al. 2021; Souza et al. 2020]. Regarding Portuguese, [Souza et al. 2020] trained BERT models for Brazilian Portuguese, resulting in BERTimbau, a LM that achieves state-of-the-art performances on three NLP tasks for two Portuguese benchmark datasets [Guillou 2021a; 2021b; Souza et al. 2020].

Shortly after, the Deep Learning Brazil group translated the SQuAD v.1.1 dataset to Portuguese using the Google Translate and some crowdsource corrections [DeepLearningBrasil 2021]. Throughout this work, this dataset is referred to as "SQuAD-BR". With a powerful new LM and a large MRC dataset available for Portuguese, it became possible, through transfer learning approach, to build models with better performance for QA tasks in that language [Mayeesha et al. 2021]. So far, little research has been conducted on SQuAD-BR dataset and, to the best of our knowledge, the state-of-the-art Extractive QA model trained on this dataset emerged in [Guillou 2021b].

Given this opportunity to advance in MRC tasks for the Portuguese language, in this research we improved the state-of-the-art on Extractive QA on the SQuAD-BR dataset by improving the fine-tuning of the models proposed in [Guillou 2021b]. In Section 2, we briefly introduce the main related works that supported our research. Then, Section 3 describes our models, and Section 4 presents the experiments carried out, involving the dataset used, the pre-processing steps, the training hyperparameters and the results achieved. Finally, Section 5 concludes our research and proposes future work to improve the model's performance.

2. RELATED WORK

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) refers to a self-supervised approach for pre-training transformer layers, before fine-tuning it for a specific NLP task. BERT uses two unsupervised tasks called Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). The MLM randomly masks some of the tokens from the input in order to predict the original vocabulary based on its context. In this way, it enables the representation to fuse the left and the right context, allowing the pre-training of a deep bidirectional transformer. The NSP objective is to predict whether a given sentence B is the actual continuation of a sentence A or whether it is a random sentence. By doing so, the pre-trained model can capture sentence level relationships, which is very beneficial to important downstream tasks such as QA [Devlin et al. 2019].

BERT was pre-trained in two model sizes: BERT Base (L=12, H=768, A=12, Total Parameters=110M) and BERT Large (L=24, H=1024, A=16, Total Parameters=340M), where L is the number of layers (i.e., Transformer blocks), H is the hidden size and A is the number of self-attention

heads. In addition, a multilingual BERT model (mBERT) was trained on 104 languages using the BERT Base architecture [Souza et al. 2020]. BERT models has demonstrated to have a good performance in QA tasks, significantly improving results on the SQuAD dataset. Beyond that, the recent rise in powerful LMs like BERT and its variants has made it possible for NLP tasks to make great progress even in low resource languages [Cambazoglu et al. 2021; Devlin et al. 2019; Mayeesha et al. 2021; Zeng et al. 2020].

2.2 BERTimbau

In recent years, much effort has been devoted on pretraining BERT-derived models on other languages, such as French, Dutch, Spanish, Italian, and others. Following this trend, [Souza et al. 2020] trained BERT models for Brazilian Portuguese using data from Brazilian Web as Corpus (BrWaC) [Wagner Filho et al. 2018], giving rise to the BERTimbau LM.

BrWac dataset is a huge Portuguese corpus which contains 2.68 billion tokens from 3.53 million documents. A new vocabulary of around 30,000 subword units was generated for the model based on this dataset and random sentences from Portuguese Wikipedia articles. The resulting vocabulary was then converted to WordPiece format, the input representation for BERT [Souza et al. 2020].

BERTimbau was evaluated on three NLP downstream tasks (sentence textual similarity, recognizing textual entailment, and named entity recognition) for the ASSIN2 and the First HAREM/MiniHAREM Portuguese datasets. As a result, the model improved the state-of-the-art on these tasks over multilingual BERT and previous monolingual approaches [Souza et al. 2020]. Like original BERT, BERTimbau is available on two model sizes: BERTimbau Base and BERTimbau Large [Souza et al. 2020].

2.3 Portuguese BERT QA models

To the best of our knowledge, the first Extractive QA model trained on SQuAD-BR emerged in [Guillou 2021a], using transfer learning approach from the BERTimbau Base LM. The research was facilitated by several Artificial Intelligence (AI) institutions, such as Hugging Face, Neuralmind and the Deep Learning Brazil group, which provide online tools (datasets, LMs, scripts and GPU platforms). The model performed well on the SQuAD-BR dataset. However, the author pointed out that English models perform even better, and, to achieve the same levels, it would be necessary to use a more efficient LM and a QA dataset in Portuguese with more examples.

The extent of the research was released a few months later. [Guillou 2021b] trained a new model on SQuAD-BR dataset using BERTimbau Large, a bigger LM than the previous one, giving rise to the new state-of-the-art, as far as we can tell. The author, however, still states that further advances are possible in order to obtain similar performance when compared to English models.

3. MODELS

This section describes the details from the models fine-tuned for the task of Extractive QA on SQuAD-BR dataset [DeepLearningBrasil 2021]. Extractive QA systems seek to find the answer for a given input question from the input context as a contiguous sequence. The model predicts the starting index and the ending index of the tokens in the context that correspond to the answer span [Cambazoglu et al. 2021; Joshi et al. 2020].

We used transfer learning from BERTimbau to improve the QA model when compared to a model trained from scratch [Malte and Ratadiya 2019]. In the next subsections we show the rationale for this approach.

3.1 Pre-trained Models

As already mentioned, the original BERT Base and Large models were pre-trained jointly on two tasks: MLM and NSP. Pre-training a model not only on MLM but also on NSP improves the performance of QA tasks significantly when applied to the original english SQuAD v1.1 [Devlin et al. 2019].

BERTimbau uses the same tasks as the original BERT for pre-training, only changing the dataset: the Wikipedia is replaced by the Brazilian Web as Corpus (BrWaC) dataset. As a result, BERTimbau outperforms BERT Multilingual on three downstream NLP tasks for two Portuguese benchmark datasets [Souza et al. 2020]. Based on BERTimbau state-of-the-art results on Portuguese datasets, we used pre-trained BERTimbau models (Base and Large) to fine-tune for a new task: Extractive QA.

3.2 Fine-tune BERTimbau on Extractive QA

This step is leveraged by both MLM and NSP tasks, mainly by the latter, due to the understanding of the relationship between two sentences [Devlin et al. 2019]. It needs less data to give good results when compared to building a model with weights randomly initialized and no transfer learning [Malte and Ratadiya 2019].

One important limitation is that the maximum input size for BERT models is 512 tokens. To circumvent this, we split each input into multiple inputs that fit the limit size at the cost of losing some context of the entire text. This is shown in Section 4.2.

Some important advantages of using BERTimbau pre-trained models for fine-tune are:

(1) Use of embeddings learned from a huge corpus to represent words and sub-words from various contexts and possible meanings, thus improving the overall understanding of the input.
(2) Use of BERT [Devlin et al. 2019] architecture that allows for significantly more parallelization [Vaswani et al. 2017].

We fine-tuned two pre-trained Masked LMs [Ahn et al. 2016]: BERTimbau Base and BERTimbau Large. In both models we add a linear layer with two outputs at the end of the pre-trained model hidden layer and use the cross entropy as our loss function for backpropagation. These two outputs predict the start index and the end index in the input context that contains the answer for the input question.

4. EXPERIMENTS

The experiments were conducted by comparing the two models described above using the pre-trained models and adding a linear layer with two outputs at the end of the pre-trained model. Every layer is trainable (i.e., unfrozen) for fine-tune. The code used in this research is available on github repository[1] with the dataset used to train and validate the models.

4.1 Dataset

As mentioned before, the SQuAD dataset is one of the first large MRC datasets, with a collection of more than 100k crowdsourced question/answer pairs posed on several Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. The data is split into training set (85%) and development set (15%). Since it was released in 2016, SQuAD v1.1 quickly became the most widely used dataset for machine reading comprehension tasks [Devlin et al. 2019; Rajpurkar et al. 2016; Zeng et al. 2020].

[1] `https://github.com/erichans/question-answering-squad-pt-br`

Although the SQuAD is an English dataset, there has been several works on translating it to other languages, such as Arabic, Korean, Hindi, Spanish, and Bengali [Mayeesha et al. 2021]. Recently, the Deep Learning Brazil group translated the SQuAD v.1.1 dataset to Portuguese using the Google Translate. As this machine translation engine inevitably propagates its decoding errors into the QA engine, the group also spent about two months making corrections to the translated dataset, before releasing the SQuAD-BR dataset. [DeepLearningBrasil 2021; Ravichander et al. 2021].

To evaluate model's performance on SQuAD v.1.1 dataset, [Rajpurkar et al. 2016] proposed two metrics: Exact Match (EM) and F1 score. EM is a binary metric that measures the percentage of predictions that match any one of the ground truth answers exactly (not counting punctuation and articles). F1 score, on the other hand, is a less strict metric, calculated as the harmonic mean of precision and recall, measuring the average overlap between the prediction and ground truth answer [Mayeesha et al. 2021; Rajpurkar et al. 2016; Zeng et al. 2020]. The same metrics are used in this work to evaluate our model.

4.2 Preprocessing

For SQuAD-BR dataset the same QA model architecture described in [Devlin et al. 2019] was adopted. Therefore, we used WordPiece embeddings with the vocabulary generated in [Souza et al. 2020] to tokenize the questions and contexts of the dataset. Next, since BERT input representation consists of a single packed token sequence, we converted the question $Q = (q_1, q_2, ..., q_m)$ and the context $C = (c_1, c_2, ...c_n)$ into a single sequence $S = [CLS]\ q_1q_2...q_m\ [SEP]\ c_1c_2...c_n\ [SEP]$, before passing it to the pre-trained models referenced in Section 3.2 [Devlin et al. 2019; Joshi et al. 2020; Souza et al. 2020].

In addition, to overcome BERT model's sequence length limitation of 512 tokens, we used the overlapping window technique, in which the inputs were split into overlapping spans of the maximum length using a fixed stride [Souza et al. 2020], as shown in Figure 1. Sequences longer than the maximum length were truncated, and sequences shorter than this were padded. As the answers can be a span extracted from the context passage, there may be inputs with unanswerable questions. In these cases, we simply set the answer span to be the special token [CLS] [Joshi et al. 2020; Mayeesha et al. 2021].

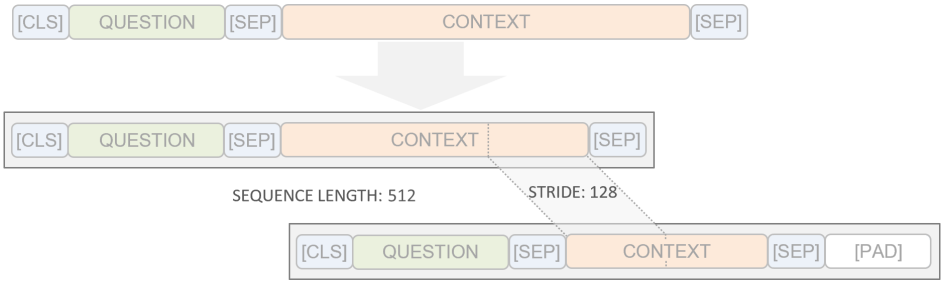

Fig. 1. Overlapping window technique with max sequence length of 512 tokens and stride of 128 tokens.

4.3 Training Hyperparameters

We had to change some hyperparameters when fine-tuning the BERTimbau Base model and the BERTimbau Large model to obtain the best results of each one.

4.3.1 *Common hyperparameters.* Both models were trained for two epochs, since larger numbers of epochs led to worse results, as verified during the training. The best models with respect to the SQuAD-BR dataset were chosen using F1 score on output. We also calculate EM to compare our models using both metrics.

We trained different max lengths for input tokens per training example, starting using 384 tokens like in [Guillou 2021a], and we got better results using 512 tokens, which is the maximum that BERT models can ingest from input. When the input example (question + context) is greater than 512 tokens, we cut the surplus tokens in the context, move them to a new example and repeat the question, repeating this process until we can have all the inputs adjusted in several examples. To compensate potential context loss, we use stride of 128 tokens from the previous example to the next one, as shown in Figure 1.

When we used 384 tokens as our max length, we had to split 1,675 input contexts from training set and 366 input contexts from development set because they could not fit the maximum limit. It corresponds to 1.91% and 2.80% of the total contexts tokens, respectively. When using 512 tokens as max length, the split numbers dropped to 248 from training set and 93 from development set, corresponding to 0.28% and 0.71% of their total contexts tokens, respectively. Those limits are shown in Figure 2. Even though only few examples were impacted by the change from 384 to 512 token limits, this small change seemed to improved the results without the fine-tune of the other hyperparameters.

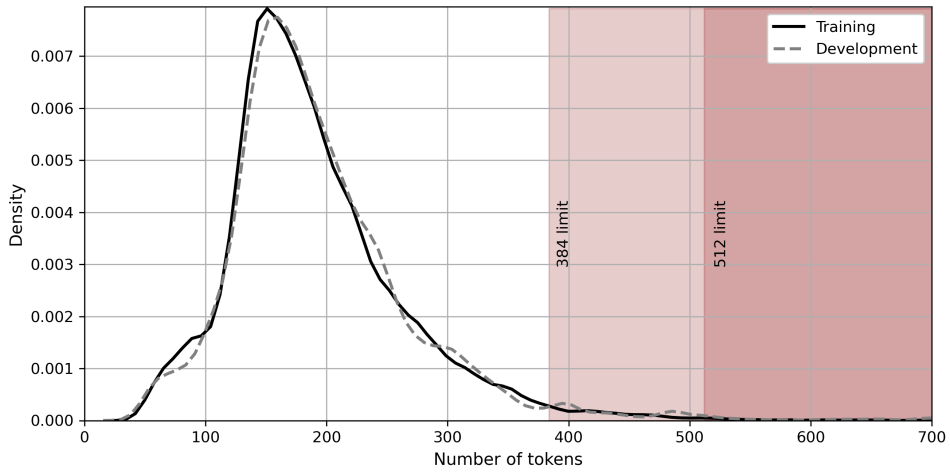

Fig. 2. Number of tokens on SQuAD-BR dataset and max sequence length limitation.

The optimizer chosen was Adam with decoupled weight decay (AdamW) [Loshchilov and Hutter 2017]. The AdamW changes the way the weight decay is treated in Adam [Kingma and Ba 2015] by decoupling weight decay and loss-based gradient updates in Adam [Loshchilov and Hutter 2017]. The weight decay applies to all layers except all bias and LayerNorm weights in AdamW optimizer [Huggingface 2021]. Despite the good results shown by [Gotmare et al. 2018] using warmup for learning rate, we had no improvements using warmup steps nor warmup ratios. This led the models to worse results in this case, so we set them to zero.

4.3.2 *BERT Base hyperparameters.* The BERT Base model was trained using a mini batch size of 16 and a learning rate of 4.25e-05. Larger or smaller batch sizes did not show any relevant improvement.

4.3.3 *BERT Large hyperparameters.* The BERT Large model was trained using a mini batch size of 8 and a learning rate of 3e-05. Even using a good GPU (Nvidia RTX 3090) with 24GB we could not fit a larger batch size.

4.4 Results

This section details the findings from our experiments using BERTimbau Base and Large for Extractive QA task on SQuAD-BR. These models had been fine-tuned to compare their performance using F1 Score and EM as shown in Table I.

Table I. Question Answering Experiment Results on portuguese SQuAD v1.1.

Architecture	**Exact Match (EM)**	**F1-Score**
$\text{BERT}_{\textbf{BASE}}$ Portuguese QA HuggingFace [Guillou 2021a]	70.49%	82.50%
$\text{BERT}_{\textbf{BASE}}$ Portuguese QA **(Ours)**	**71.09%**	**82.91%**
$\text{BERT}_{\textbf{LARGE}}$ Portuguese QA HuggingFace [Guillou 2021b]	72.68%	84.43%
$\text{BERT}_{\textbf{LARGE}}$ Portuguese QA **(Ours)**	**73.12%**	**84.74%**

Our BERTimbau Large QA model established a new state-of-the-art on SQuAD-BR and, as expected, it achieved a better result when compared to our BERTimbau Base QA model. Our BERT Base QA model also surpassed the BERT Base from [Guillou 2021a].

5. CONCLUSIONS AND FUTURE WORK

In this work we achieved the objective of establishing a new state-of-the-art on SQuAD-BR, the SQuAD v.1.1 dataset translated to Portuguese by the Deep Learning Brazil group, by proposing a Extractive Question Answering model that surpasses, in performance, the model proposed by [Guillou 2021b]. We progress in both metrics, Exact Match and F1 Score, using BERTimbau Base and BERTimbau Large pre-trained LMs as shown in Table I.

We plan to fine-tune BERTimbau LM on SQuAD-BR and then fine-tune the QA as proposed by [Howard and Ruder 2018]. We also plan to train a new LM model from scratch based on newer models like XLNet [Yang et al. 2019] and T5 [Raffel et al. 2020]. Another aproach that may lead to good results is use a NER classifier to identify entities in the context and add as input features to the QA model.

While our models has performed well when compared to the previous approaches, they had been trained only on SQuAD-BR. We plan to further improve the robustness of the model by training it on different datasets to reduce the effect of distribution shift, as shown in [Miller et al. 2020].

REFERENCES

Ahn, S., Choi, H., Pärnamaa, T., and Bengio, Y. A neural knowledge language model. *CoRR* vol. abs/1608.00318, 2016.

Cambazoglu, B. B., Sanderson, M., Scholer, F., and Croft, B. A review of public datasets in question answering research. *SIGIR Forum* 54 (2), Aug., 2021.

DeepLearningBrasil. Squad v1.1 automatically translated to portuguese and reviewed, 2021. https://drive.google.com/file/d/1Q0IaIlv2h2BC468MwUFmUST0EyN7gNkn/view?usp=sharing. Last accessed: 2021-09-14.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, 2019.

Gotmare, A., Keskar, N. S., Xiong, C., and Socher, R. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *CoRR* vol. abs/1810.13243, 2018.

Guillou, P. Bert base cased squad v1.1 portuguese, 2021a. https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese. Last accessed: 2021-09-13.

Guillou, P. Bert large cased squad v1.1 portuguese, 2021b. https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese. Last accessed: 2021-09-13.

Howard, J. and Ruder, S. Fine-tuned language models for text classification. *CoRR* vol. abs/1801.06146, 2018.

Huggingface. Trainer - transformers 4.7.0 docs, 2021. https://huggingface.co/transformers/main_classes/trainer.html. Last accessed: 2021-07-08.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* vol. 8, pp. 64–77, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun (Eds.), 2015.

Liu, N. F., Lee, T., Jia, R., and Liang, P. Can small and synthetic benchmarks drive modeling innovation? A retrospective study of question answering modeling approaches. *CoRR* vol. abs/2102.01065, 2021.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *CoRR* vol. abs/1711.05101, 2017.

Malte, A. and Ratadiya, P. Evolution of transfer learning in natural language processing. *CoRR* vol. abs/1910.07370, 2019.

Mayeesha, T. T., Sarwar, A. M., and Rahman, R. M. Deep learning based question answering system in bengali. *Journal of Information and Telecommunication* 5 (2): 145–178, 2021.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. *CoRR* vol. abs/2004.14444, 2020.

Patel, D., Raval, P., Parikh, R., and Shastri, Y. Comparative study of machine learning models and BERT on squad. *CoRR* vol. abs/2005.11313, 2020.

Pranesh, R. R., Shekhar, A., and Pallavi, S. Quesbelm: A bert based ensemble language model for natural questions. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)*. pp. 1–5, 2020.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21 (140): 1–67, 2020.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* vol. abs/1606.05250, 2016.

Ravichander, A., Dalmia, S., Ryskina, M., Metze, F., Hovy, E., and Black, A. W. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, pp. 2976–2992, 2021.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 403–417, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR* vol. abs/1706.03762, 2017.

Wadhwa, S., Chandu, K., and Nyberg, E. Comparative analysis of neural QA models on SQuAD. In *Proceedings of the Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Melbourne, Australia, pp. 89–97, 2018.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. LUKE: deep contextualized entity representations with entity-aware self-attention. *CoRR* vol. abs/2010.01057, 2020.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Vol. 32. Curran Associates, Inc., 2019.

Zeng, C., Li, S., Li, Q., Hu, J., and Hu, J. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences* 10 (21), 2020.