An Empirical Study of Information Retrieval and Machine Reading Comprehension Algorithms for an Online Education Platform

Eduardo F. Montesuma[1], Lucas C. Carneiro[1], Adson R. P. Damasceno[2], João Victor F. T. de Sampaio[1], Romulo F. Férrer Filho[1], Paulo Henrique M. Maia[2], Francisco C. M. B. Oliveira[2]

[1]Federal University of Ceará
Fortaleza – CE – Brazil

[2]State University of Ceará
Fortaleza – CE – Brazil

{firstname.lastname@dellead.com}

Abstract. *This paper provides an empirical study of various techniques for information retrieval and machine reading comprehension in the context of an online education platform. More specifically, our application deals with answering conceptual students questions on technology courses. To that end we explore a pipeline consisting of a document retriever and a document reader. We find that using TF-IDF document representations for retrieving documents and RoBERTa deep learning model for reading documents and answering questions yields the best performance with respect to F-Score. In overall, without a fine-tuning step, deep learning models have a significant performance gap with comparison to previously reported F-scores on other datasets.*

1. Introduction

In distance learning courses, tutors play a crucial role due to performing several activities, such as pedagogical support, student performance, interaction monitoring, dropout detection, prevention, and reduction [Barker 2002, Denis et al. 2004], thus helping students to finish their course successfully [Simpson and Sharma 2002, Lentell 2004]. However, when a large number of students attend courses, human tutors can be overloaded, which may have a negative impact on their work. Bernath and Rubin (2001) report how the sheer volume of online activities can be too much for the teacher and the student and why the workload on online teachers is often reported to be significantly greater than what it is in a face-to-face teaching context.

In this realm, Damasceno *et al.* (2020) previously proposed a chat-bot called STUART for easing the burden in distance learning courses. It uses Natural Language Processing (NLP), machine learning techniques, and interaction with Dell Accessible Learning (DAL) [1] learning tools for responding to student's pedagogical, technical, and content demands. [Damasceno et al. 2020] does so by sending proactive pedagogical recommendations according to the student's profile and for ensuring the reduction of activities that require pedagogical resources and human tutoring. In this work we focus on content responses using the Question Answering (QA) technology.

[1]http://leadfortaleza.com.br/dal/

As follows, QA technology is the possible solution to mitigate that problem [Wen et al. 2012]. A QA system aims to automatically answer some of the students questions, thus reducing the workload on teachers. Modern QA systems are composed by various sub tasks. In this work we focus on two of then: document retrieval and document reading, under the perspective of NLP. Concerning the document retrieval task, it consists of finding the document that is most similar to a given question. This is done either by calculating the similarity between documents, through the usage of document representations as Term Frequency-Inverse Document Frequency (TF-IDF) [Jones 1972] or by word representations [Pennington et al. 2014].

After finding the appropriate document, one needs to extract the answer. A possible approach task that gained attention in the literature is called Machine Reading Comprehension (MRC) [Hermann et al. 2015], which aims at teaching machines to answer questions after comprehending given passages or contexts. The state-of-the art in MRC are deep language models based on attention mechanism, such as BERT [Devlin et al. 2019], ELECTRA [Clark et al. 2020] and ALBERT [Lan et al. 2020]. Despite the many advances in the field, the usage of these tools remains data intensive. For instance, when applying BERT for a domain-specific QA, such as biology, a fine-tuning step is necessary since the word distribution can drastically change across domains [Lee et al. 2020]. This is an important drawback, specially for small to medium-sized distance learning platforms that do not have a reasonable corpus for fine-tuning the model.

Due to the wide variety of strategies for performing retrieval and reading tasks, choosing the appropriate model for each of them is cumbersome. In this sense, the majority of work evaluate either the retrieval task, or the reading performances alone. Nevertheless, some surveys provide a broader view on the subject. For instance, in [Abbasiantaeb and Momtazi 2021], the authors provide a comprehensive review of the complete QA pipeline, covering various deep learning approaches. Moreover, Fu et al. (2020) covers traditional methods, such those that rely on predefined rules or templates for answering questions.

In this context, our work differentiates itself from these approaches since it provides an empirical study of deep learning-based QA in the context of online distance courses, predominantly with technology subjects. Therefore we present an empirical study of retrieval and reading tasks working in consonance. Our study is centered around an existing educational platform, the DAL platform. We investigate the following questions: (i) Which algorithms better fit the pipeline used for QA? (ii) Can pre-trained deep learning models perform as well even when fine-tuning is not possible? (iii) Can deep learning models be used for answering conceptual questions of the DAL platform students?

The remainder of this paper is organized as follows. Section 2 gives an overview of the proposed methodology with Document Retriever and Document Reader. Section 3 presents the results and inferences obtained from them. Section 4 draws the conclusions we can draw from our work.

2. Methodology

In this section we explore the methodology for QA. In particular, we propose using a pipeline similar to that of [Abbasiantaeb and Momtazi 2021], with slight modifications. It is composed by three elements: (i) a corpus of contexts, (ii) a document retriever module, and (iii) a document reader module. An overview on how these components work in consonance is shown in Figure 1.

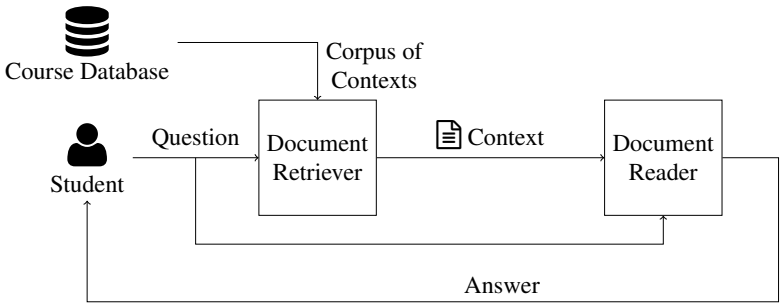

Figure 1. Pipeline for the MRC module.

2.1. Corpus of Contexts

We create a usable knowledge base for answering the students questions by gathering textual content from courses offered in the DAL platform. The DAL platform provides a variety of multidisciplinary online courses, mostly about technology or management. The lectures of the courses are displayed as web pages or videos.

The data collection process was made through an automatic script that downloads all the content from DAL's platform courses. For online lectures, we parsed the HTML files to extract the textual content of web lectures. For video lectures, we extracted the textual content from subtitles. In total, we gathered 674 documents, from 189 lectures of 18 courses as illustrated by Figure 2.

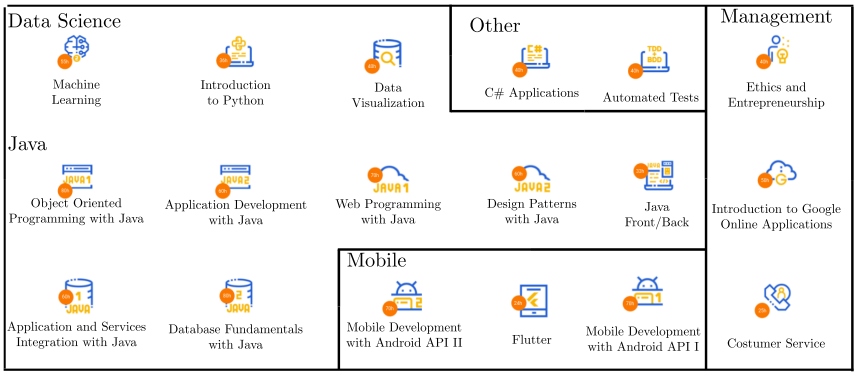

Figure 2. An overview of the 18 courses available in the DAL's platform.

After obtaining the texts, we split the documents into segments consisting of a maximum of $T = 128, 256,$ and 512 tokens, resulting in three knowledge bases with the same content, but divided in distinct number of segments. In addition, we executed prepossessing steps in the texts, which include normalization in lower case, removing stop words, punctuation, numbers and accentuation, and lemmatisation.

Furthermore, we also created a corpus of conceptual questions for evaluating our pipeline. For the 10 most popular courses, DAL's tutors created questions based on their past experience of what students may ask. For each test question, the following labels are available: the course and lecture related to the question, the context that answers the question, and the desired answer within the context. This allows us to evaluate both the retrieving and reading tasks.

2.2. Document Retriever

The Document Retriever component executes an Information Retrieval task [Kolomiyets and Moens 2011] by retrieving from a corpus a document that is expected to contain the answer for a question, i.e. the context of the question. The criteria to find this context is to compute a similarity measure between the question and the documents in a given vector representation, in which the document with a higher similarity is chosen. The underlying hypothesis is that the greater the similarity between the question and a document, the more likely that document contains the answer to the question. In a distance learning application, each document refers to an excerpt of the textual content of a lecture from an online course. Besides the text of the candidate document, the name of the course and the lecture identifier is also returned.

There are various approaches for evaluating the similarity between documents and questions. In this work we focus on those techniques involving the numerical representation of documents. As follows, one tries to represent either the entire document, or its words by a vector $\mathbf{x} \in \mathbb{R}^n$. Examples of the former approach are the so-called Bag of Words (BoW) and TF-IDF [Jones 1972], which roughly rely on the word frequency in each document. Moreover, Global Vectors (GloVe) word embeddings [Pennington et al. 2014], which are trained on large corpus of texts in an unsupervised fashion is an example of the latter approach.

Once one has a representation for the document, it is still necessary to choose a notion of distance between the vectors. A common choice in the literature is using the cosine similarity. On the other hand, when using word embeddings, novel distances such as the Word Mover Distance (WMD) [Kusner et al. 2015] can be used. The WMD is particularly interesting since it considers documents as empirical probability measures over the word space.

2.3. Document Reader

The Document Reader component works as a MRC module. It implements a Deep Learning model to predict possible answers to questions made by students using the DAL platform. The Document Retriever module provides a context from which the Document Reader extracts an answer to those questions. The prediction itself is made by finding the beginning and the end of the answer in the context provided by the Document Retriever. It uses ELECTRA as its deep learning model, but BERT, RoBERTa, and ALBERT were also used for comparison purposes, and are briefly described below.

Bidirectional Encoder Representation from Transformers (BERT) [Devlin et al. 2019] is a Masked Language Model (MLM) [Taylor 1953] that selects a small subset of the unlabeled input data and masks its tokens identity (15% of tokens are masked, replaced, or left unchanged at random.) The network is then

trained to predict the original input. It achieved state-of-the-art results in various NLP tasks, such as QA and Sentiment Analysis. One of BERT's major innovations is applying bidirectional training to language representations as opposed to single direction (usually left-to-right modeling or a combination of both left-to-right and right-to-left modeling) training. It allows the fusion of both contexts, the one to the left of the masked token and the one to the right, to predict the masked word found in the original input. BERT also uses a next sentence prediction task that can capture the relationship between sentences.

The Robustly optimized BERT approach (RoBERTa) [Liu et al. 2019] improves on BERT by optimizing its method of pretraining. The authors evaluated how hyperparameter tuning and training set size impact the performance of BERT-like models. The model is trained longer, with larger batches and learning rates, on more data. The model excludes the next sequence prediction task and it is trained on longer sequences. Also, the model has a dynamic masking strategy applied to the training data, where a masking pattern is generated every time a sequence is fed to the model. The model itself is a reimplementation of BERT with the aforementioned modifications and improved on BERT's results on General Language Understanding Evaluation (GLUE) [Wang et al. 2018] and Stanford Question Answering Dataset (SQuAD) [Rondeau and Hazen 2018] benchmarks.

The model A Lite BERT (ALBERT) [Lan et al. 2020] is a low memory consumption model similar to BERT. It has two parameter reduction techniques that lower its memory consumption and increase its training speed. The first technique consists in decomposing the embedding matrix into two smaller ones, so it is easier to grow the model hidden size without increasing the number of parameters. The second one is sharing parameters across all layers to prevent the number of parameters from growing with network depth. Both techniques improve parameter efficiency without reducing model performance.

Efficiently Learning an Encoder that Classifies Token Replacement Accurately (ELECTRA) was first proposed in [Clark et al. 2020]. The model detects replaced tokens instead of recovering the original input. The input data is corrupted by replacing tokens with some others generated by a small masked language model. The network is then pretrained as a discriminator that predicts if every token is in the original input or not and can be fine-tuned on downstream tasks. ELECTRA's greatest advantage over MLM is that it learns from all input tokens instead of only from a small subset of the original data, which makes it computationally more efficient.

3. Results

In this section we describe three experiments, and discuss our results. Section 3.1 explores the space generated by the TF-IDF representation. Section 3.2 presents a comparative study of information retrieval and deep learning-based MRC algorithms. Finally, Section 3.3 provides a broad discussion on the results we obtained.

3.1. Corpus Visualization

Based on the pre-processed text from the corpus, we built the TF-IDF representation consisting on a matrix $\mathbb{R}^{N \times M}$, for N, the number of segments, and $M = 18,233$, the vocabulary size. To further explore this representation of the corpus, we perform two dimensionality reduction steps: (i) we apply Principal Component Analysis (PCA) to the

data, reducing it to a reasonable dimension (100), (ii) from the reduced vector, we apply t-distributed Stochastic Neighbor Embeddings (t-SNE) [Van der Maaten and Hinton 2008] to visualize the data on $\mathbb{R}^2$, as shown in Figure 3.

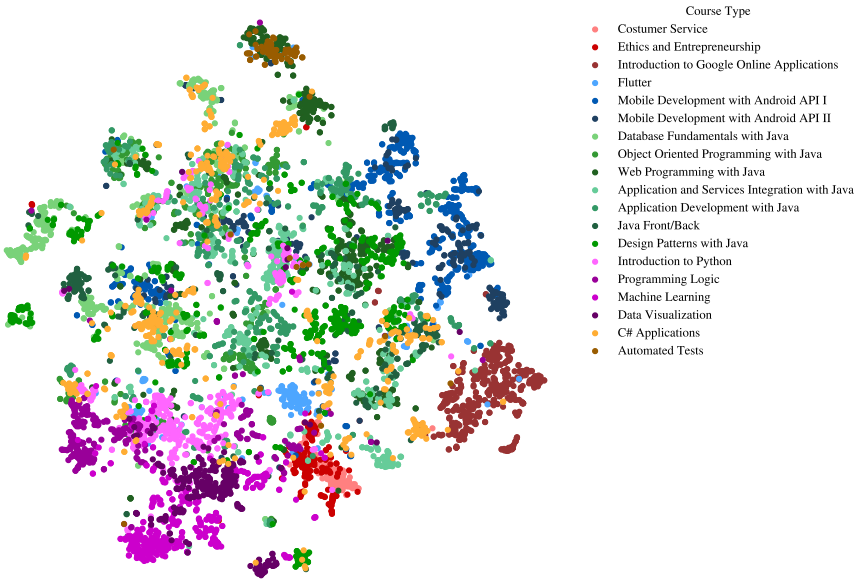

Figure 3. t-SNE embedding of the TF-IDF representation of documents. The colours represent each the 18 courses in the DAL platform.

3.2. Pipeline Evaluation

The evaluation of the pipeline shown in Figure 1 is carried out in two steps. First, the information retrieval task is evaluated, according to both course and lecture accuracy. It is important to note that a lecture prediction is only correct if the document retriever also predicts the course correctly. At this step, we also compare different choices for document segmentation, as we evaluate each candidate retrieval algorithm for a maximum of 128, 256 and 512 tokens.

Thus, we compare 4 different approaches for retrieving documents: (i) the TF-IDF representation [Jones 1972], (ii) the usage of BoW, both using cosine similarity (iii) Portuguese GloVE Word Embeddings [Hartmann et al. 2017] using the so-called WMD and (iv) using Word Centroid Distance (WCD), both as proposed by [Kusner et al. 2015], which reported that WCD outperformed both BoW and TF-IDF in tasks of Information Retrieval by significant margins in various datasets. The comparison is shown in Table 1.

Furthermore, for the comparison of MRC methods presented in Section 2.3, we use the F-Score, a metric previously used in [Rajpurkar et al. 2016] for QA. Based on this metric, we present two comparisons. First, we consider that the answer context is known *a priori*, thus yielding a similar evaluation to previous studies, but in a different context (DAL database). Second, we consider the evaluation of the pipeline as a whole. In this case, the context is predicted using the best information retrieval method, which is TF-IDF with a segmentation of 128 tokens, as presented in Table 1. The MRC results are shown in Table 2.

Feature Choice	Course Accuracy 128	256	512	Lecture Accuracy 128	256	512
GloVE Embeddings						
- WMD	38.18	33.66	37.12	26.73	23.76	22.77
- WCD	35.15	34.16	31.18	23.76	17.82	13.86
BoW	50.49	42.08	42.08	39.11	32.17	32.17
TF-IDF	72.77	65.84	63.86	62.37	55.94	52.97

Table 1. Comparison of Information Retrieval methods based on course and lecture accuracy.

Model	# Parameters	Context Known	Context Retrieved
BERT	334M	13.39	9.22
ALBERT	235M	15.42	14.16
ELECTRA	334M	15.81	16.62
RoBERTa	354M	15.88	16.83

Table 2. Comparison of MRC algorithms based on F-Score, in two scenarios: (i) context is known *a priori*, and (ii) context is retrieved by the document retriever.

3.3. Discussion

Figure 3 evidences the difficulty of detecting the appropriate course for a given question, as the course's content can be highly multi-disciplinary. For instance, the extreme top cluster of Figure 3 is composed by documents from "Web Programming with Java" and "Automated Tests". At a first glance, these two courses are unrelated, but the documents treat the same subject (in the former course, it treats tests in Java).

Another source of confusion is related to the courses "Machine Learning", "Data Visualization" and "Introduction to Python", at the bottom of Figure 3, in shades of purple. Even though the first 2 of these 3 courses treat advanced topics, they are taught using Python, thus there is an intersection in their content, mostly in the initial lectures. This is also the case for the course "C# Applications", in yellow. The course content involves, among other topics: object-oriented programming, databases, and web applications. Consequently, the documents of this course are scattered throughout the t-SNE embedding. Hence, the Information Retrieval task achieved better results for all methods with the segmentation of 128 tokens, indicating that short texts are more easily distinguished. This may happen since these have less possibility of overlapping topics.

As a possible solution for this issue, we may leverage the DAL platform by retrieving which courses a given student is enrolled in. This allows us to narrow the search to a few courses. Even though a student may be enrolled in courses that share content, the system will likely perform better in terms of document and lecture accuracy.

Secondly, unsupervised document representation approaches, such as BoW and TF-IDF have superior performance than GloVe embeddings. This was expected, since the domain of our application is very specific. More specifically, Hartmann *et al.*

(2017) trained the embeddings on a corpus consisting mainly on Wikipedia and news pages, whereas our application is mainly composed of technology-related courses. As a consequence, training and test data follow different probability distributions, harming the test performance. A possible solution that may improve the performance of GloVe is training on a corpus that is more similar to our application, or performing fine-tuning. However, due to the limited size of the DAL's corpus, this was not feasible.

Thirdly, in comparison with the results reported on the respective paper of each model, there is still a wide margin improvement. For instance, Clark *et al.* (2020) report a F-Score of 94.9 on SQuAD v1.1 dataset, whereas in the DAL test set the performance is much lower (16.83 maximum). The reason for this performance gap is, again, the distributional shift between training and test sets for these models. Possible approaches for solving this issue are: re-training, fine-tuning or even performing transfer learning [Ganin et al. 2016].

Finally, note that as reported in Table 2, RoBERTa and ELECTRA have slightly better performance when the context is retrieved through the document retriever. This is mainly due the fact that different contexts may contain the answer for a given question. This also highlights that the context indicated by the tutors may not be the most appropriate for the model to extract the answer from.

4. Conclusion

This work presented an empirical study of information retrieval and machine reading comprehension algorithms for an online education platform. We used a pipeline consisting of a corpus of contexts based on courses content, a document retriever, and a document reader. The pipeline has two challenges to overcome. First, contexts are not always informative. They might be guessed wrongly by the document retriever, or the information may be scattered across documents. Secondly, the Language Model (LM), which have been built using general language and broad scope texts, are applied to a corpus that is very content specific (technology courses, for instance). This hinders the quality of answers.

Concerning the best choices for the QA pipeline, using unsupervised document representations, such as TF-IDF, yields better performance than pre-trained word embeddings, such as GloVe. Among these, the former has the better retrieval performance. Furthermore, when comparing MRC algorithms, ELECTRA and RoBERTa give the best results, having similar performance. Moreover, the overall pipeline, that is, TF-IDF retriever and RoBERTa-based reader with a document segmentation of 128 tokens, yields slightly better answers than using the reader with contexts provided by the DAL tutors. Furthermore, without fine-tuning, there is a considerable gap in F-Score when using pre-trained deep learning-based MRC algorithms in a specific domain QA problem, such as answering conceptual questions about technology courses.

Given the shortcomings of applying pre-trained deep learning-based models in specific contexts, deep learning models can be used for answering conceptual questions. Future work involve: (i) improving the DAL's data base for allowing the fine-tuning of MRC algorithms, and (ii) employing transfer learning for efficiently using the data at our disposal for improving pre-trained models.

References

Abbasiantaeb, Z. and Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1412.

Barker, P. (2002). On being an online tutor. *Innovations in Education and Teaching International*, 39(1):3–13.

Bernath, U. and Rubin, E. (2001). Professional development in distance education – a successful experiment and future directions. *Innovations in Open & Distance Learning, Successful Development of Online and Web-Based Learning*, pages 213–223.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Damasceno, A. R., Martins, A. R., Chagas, M. L., Barros, E. M., Maia, P. H. M., and Oliveira, F. C. (2020). Stuart: an intelligent tutoring system for increasing scalability of distance education courses. In *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10.

Denis, B., Watland, P., Pirotte, S., and Verday, N. (2004). Roles and competencies of the e-tutor. In *Networked Learning 2004: A Research Based Conference on Networked learning and lifelong learning: Proceedings of the fourth international conference, Lancaster*, pages 150–157.

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

Fu, B., Qiu, Y., Tang, C., Li, Y., Yu, H., and Sun, J. (2020). A survey on complex question answering over knowledge base: Recent advances and challenges. *arXiv preprint arXiv:2007.13069*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131. SBC.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28:1693–1701.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966. PMLR.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lentell, H. (2004). The importance of the tutor in open and distance learning. In *Rethinking Learner Support in Distance Education*, pages 76–88. Routledge.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rondeau, M.-A. and Hazen, T. J. (2018). Systematic error analysis of the stanford question answering dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 12–20.

Simpson, O. and Sharma, R. C. (2002). Book review-supporting students in open and distance learning. *International Review of Research in Open and Distance Learning*, 3(3).

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Wen, D., Cuzzola, J., Brown, L., and Kinshuk, D. (2012). Instructor-aided asynchronous question answering system for online education and distance learning. *International Review of Research in Open and Distributed Learning*, 13(5):102–125.