NukeBERT: A Pre-trained language model for Low Resource Nuclear Domain

Ayush Jain*, Dr. N.M. Meenachi†, Dr. B. Venkatraman†
*BITS PILANI, Pilani, India, †IGCAR, Kalpakkam, India
*ayushjain1144@gmail.com †meenachi@igcar.gov.in, †bvenkat@igcar.gov.in

***Abstract*—Significant advances have been made in recent years on Natural Language Processing with machines surpassing human performance in many tasks, including but not limited to Question Answering. The majority of deep learning methods for Question Answering targets domains with large datasets and highly matured literature. The area of Nuclear and Atomic energy has largely remained unexplored in exploiting available unannotated data for driving industry viable applications. To tackle this issue of lack of quality dataset, this paper intriduces two datasets: NText, a eight million words dataset extracted and preprocessed from nuclear research papers and thesis; and NQuAD, a Nuclear Question Answering Dataset, which contains 700+ nuclear Question Answer pairs developed and verified by expert nuclear researchers. This paper further propose a data efficient technique based on BERT, which improves performance significantly as compared to original BERT baseline on above datasets. Both the datasets, code and pretrained weights will be made publically available, which would hopefully attract more research attraction towards the nuclear domain.**

***Index Terms*—Natural Language Processing, Question Answering, Bidirectional Representational Transformers, SQuAD, Nuclear, Pretraining, Fine-tuning.**

I. Introduction

The nuclear industry presents a viable solution to end energy crisis. Advancements in machine and deep learning is playing a significant role in enhancing the progress in the medical domain ([1], [2], [3]) and a lot of groundwork is already done in generation of various datasets and pretrained models, thus reducing the barrier for future research. Similar advancements needs to happen in the nuclear field. A survey indicated that limited effort has gone into the domains of power plants and atomic energy[4]. This research aims to develop quality datasets and explore the applicability of BERT (Bidirectional Encoder Representations from Transformers)[5] in the nuclear field, which lacks high quality data.

On this ground, a model is proposed in this research which will be referred to as Nuclear Bidirectional Encoder Representations from Transformers for language understanding (NukeBERT) (Figure - 1). NukeBERT is a contextualized nuclear word embedding, based on BERT, which can be fine-tuned further on numerous downstream tasks including Question Answering, Named-entity recognition and Sentence segmentation. BERT is trained on wikipedia corpus, which is strikingly different and lacking in nuclear terminologies. Further, BERT requires a huge corpus for generating word embeddings, which is difficult to build in a low resource nuclear domain. To incorporate the nuclear jargons, we developed a custom dataset, "NText", for training BERT. The model is then evaluated for its performance on Question Answering task. Due to lack of question-answering dataset in the nuclear domain, we developed a new dataset, called NQuAD (Nuclear Question Answering Dataset), prepared with the help of nuclear domain experts at Indira Gandhi Center for Atomic Research (IGCAR) as part of this research.

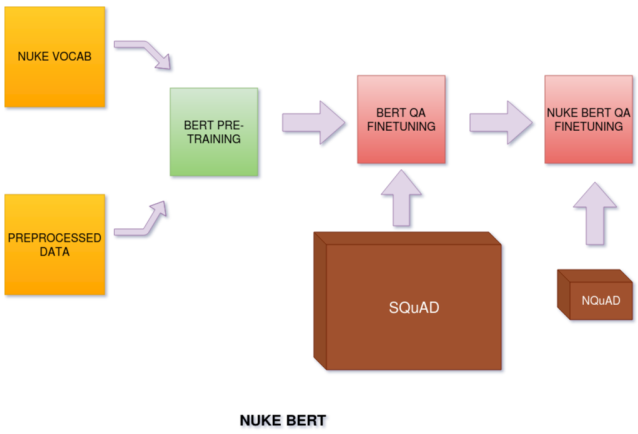


Fig. 1: **NukeBERT Pipeline**: Using a vocabulary having nuclear jargons, the preprocessed data is pre-trained on a corpus of nuclear text, which is then further finetuned on SQuAD and NQuAD.

With the same configuration as used by original BERT model for fine-tuning, NukeBERT is fine-tuned on Stanford Question Answering Dataset (SQuAD)[6] and then on NQuAD train set. On testing, the NukeBERT outperformed the original BERT model significantly.

Our contributions are as follows: 1) We introduce a novel nuclear natural language datasel called NText, which can be further used for numerous language modelling tasks. 2) We further introduce the first nuclear question answering dataset, which has expert curated 700+ question answer pairs which can be used to train and test various NLP models for nuclear domain. 3) We further propose a novel state-of-the-art technique which is able to achieve superior performance while using orders of magnitude less data.

For fostering future research and industrial applications, we will be open-sourcing NText, NQuAD and pretrained weights.

II. Related Work

Word2Vec [7] pioneered the semi-supervised approach to read massive corpora and generate meaningful word embed-

dings. However, vanilla Word2Vec suffers massively from polysemy, i.e. words with different meaning in different contexts and inability to effectively deal with out of vocabulary words ([8], [9])

The advent of ELMO (Embeddings from Language Models) [10], GPT (Generative Pre-Training Model) [11] and BERT demonstrated the benefit of unsupervised language modelling pre-training on large corpus for various sophisticated downstream tasks. While GPT and ELMO required task specific architectures, BERT showed excellent results on various NLP tasks and benchmarks with minimum need of architectural modifications. BERT was trained on 3.3 Billion words dataset to generate BERT embeddings, which were then fine-tuned to various downstream tasks. With minimal architecture modification, BERT was able to achieve a state of the art in 11 natural language processing tasks.

SciBERT, a BERT based approach proposed for scientific data, was trained on 1.14 Million papers from semantic scholar having 3.17 Billion tokens. It was able to improve the model accuracy on several scientific and biomedical datasets than BERT [12]. BioBERT, proposed for biological domain, is trained on 18 Billion words dataset of medical corpus and 3 Billion words BERT dataset. BioBERT could significantly improve the state-of-the-art performance on various medical datasets. [13].

In contrast to BERT, SciBERT and BioBERT, nuclear domain suffers from lack of large, high quality datasets. After collecting data from nuclear research papers. a 8 Million words nuclear dataset was generated, which is two magnitudes lower than dataset by the existing BERT based approaches. Hence a novel approach to optimize the formation of Nuclear Vocab (NukeVOCAB) and pre-training to achieve better results than BERT is developed. Unlike the medical domain, the nuclear domain do not have any gold training dataset to gauge the performance of this newly developed model. Hence a Nuclear Question Answering Dataset (NQuAD) is developed with the effort of Nuclear scientists. The domain experts have evaluated the system for its performance. The NukeBERT methodology is explained in the following section.

III. Methodology

A. NText Preparation

This research work proposes NText, which is a Nuclear Textual dataset containing textual data related to nuclear domain. For the dataset preparation, 7000 internal reports, thesis and research papers in the PDF format were taken from the Indira Gandhi Centre for Atomic Research (IGCAR). The sizes of the reports ranged from a couple of pages to a few thousand pages. A substantial portion of nuclear corpus consisted of very old reports, some of which were stored as scanned copies. The reports primarily dealt with the nuclear domain, many of them explicitly dealing with Fast Breeder Reactor(FBR). The raw corpus needed extensive cleaning and preprocessing to convert it into pre-training corpus suitable for BERT. The next section will discuss the preprocessing pipeline in detail.

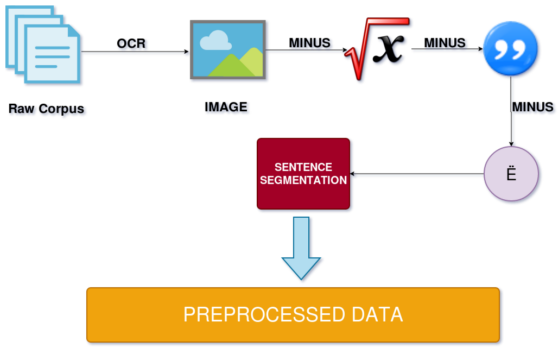


Fig. 2: Flowchart of steps for preparation of preprocessed data from raw corpus of 7000 research papers.

B. PREPROCESSING

The detailed pipeline for preprocessing is shown in Figure - 2. Since the dataset had a lot of scanned copies which are unreadable for PDF parsing libraries, it was decided to use Optical Character Recongnition (OCR) for extracting the text. Another benefit of using OCR is that it avoids random unicode errors which typically occur while parsing PDFs and would have been unavoidable considering the diversity of the PDFs in the corpus used. Tesseract-Optical Character Recognition (OCR) [14] is used to read the text from the images. The OCR is not accurate and generates specific errors like reading 'o' as '0', 'I' as '1'. However, according to Namsyl et al. [15], these errors are infact a useful data augmentation technique. Further, it is not a significant issue since BERT tokenizer will be tokenizing the errorneous word by splitting it at the wrongly comprehended alphabet and hence would still be able to understand a meaningful representation of the word.

OCR is computationally expensive, and its accuracy and processing time highly depends upon the quality of the image. Our corpus has varied quality PDFs, with some being very old, handwritten and misaligned. It took approximately four days of continuous running on a single Google Cloud K80 Tesla GPU. The average time taken for OCR and further preprocessing (described below) was around 20 seconds/page. In general, there are many mathematical formulas in the research papers and thesis. We devised several regular expression patterns to remove them from the corpus. Further, there were a lot of references, both in-text and at the references at the end of the main text of the papers. For removing them, we designed regex patterns for common styles of in-text citations used in the literature. We noticed that last 2-3 pages generally contain references, and hence we removed last two pages from each corpus as well.

Since the original BERT model is trained on general English corpus, having non-english text would not be suitable for training purposes. The raw corpus had some research papers and thesis in German and French. There are some libraries like NLTK which can approximately identify whether a word is a non-english word. However, in our experiments, we found that it was not suitable as some of the nuclear jargons were

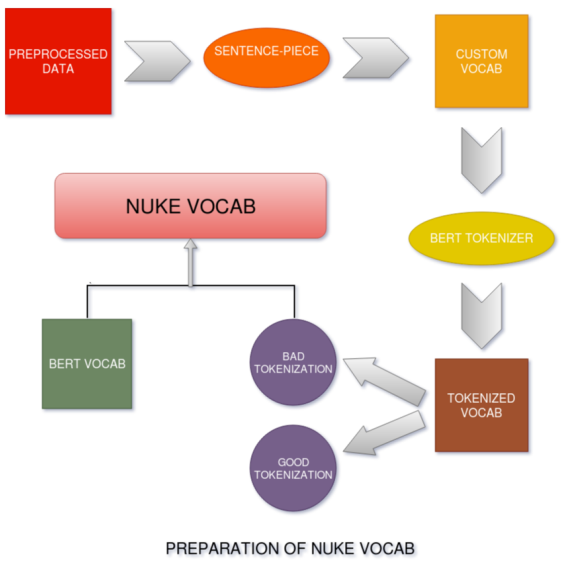


Fig. 3: Flowchart for preparation of NukeVocab. Notice we only concatenate "Bad Tokenization" with BERT Vocabulary.

Custom Vocabulary	BERT Tokenization
irradiation	'ir', '#rad', '##iation'
reactivity	'react', '#ivity'
weld	'w', #eld
plenum	'pl', '##en', '##um'
electrochemical	'electro', '##chemical'
carbide	'car', '##bid', '##e'
boron	'bo', '##ron'
pellets	'pe', '##llet', '##s'
oxides	'oxide', '##s'
annealing	'anne', '##aling'

TABLE I: Table showing tokenization by BERT tokenizer of the BERT Custom Vocabulary - the vocabulary prepared by running sentence piece algorithm on NText. Note that BERT tokenizer often performs poorly on these samples.

also falsely marked as non-english word. We noticed that non-english texts in the raw corpus typically contained characters like ë, è, é. Since ASCII representation only contains numbers and English alphabets, we ignored those lines which included these non ascii characters and thus removed about 20000 non-english lines from the corpus. Further, we also manually removed the non-english texts from the raw corpus.

The input for BERT pre-training is a text file with one sentence per line. An empty line delimits the documents. We used the gensim library for segmenting the text into sentences, which finally makes the input text file ready for pre-training and other tasks. After all the cleaning and preprocessing, we obtained NText, a Nuclear textual corpus consisting of about 8 Million words.

C. NukeVOCAB

BERT makes use of a technique referred as WordPiece to build a thirty two thousand word vocabulary on their general words corpus. Since the nuclear domain contains a lot of jargons which are underrepresented in general text corpus. Hence, it is essential to have a vocabulary tailored to the nuclear domain. In this direction, we built a Nuclear domain specific vocabulary of thirty thousand words which will be referred as **NukeVocab** (Nuclear Vocabulary) in further discussions. The procedure for constructing NukeVocab is summarized in Figure-3. Since WordPiece library is publically unavailable, we modified the output of sentence piece library, an open source library released by tensor2tensor [16], to make it similar to BERT vocabulary. For generating NukeVocab, the sentence piece library is used to generate a BERT like vocabulary of thirty thousand words from NText data. We will refer it as Custom-Vocab (Table-I), which will be further modified to obtain NukeVocab. Note that while BERT generated their vocabulary on general word corpus, we generate Custom-Vocab from text dealing primarily with nuclear domain, and hence would sufficiently represent the important and most frequent jargons in Nuclear field.

Next, we wanted to identify those words which BERT does not comprehend given its present vocabulary. For achieving this, first Hugging Face's Pytorch implementation of BERT tokenizer [17] is used to tokenize the Custom Vocab. The BERT tokenizer breaks the words into tokens, with each token belonging to BERT vocabulary. If a complete word is part of BERT vocabulary, it does not get broken into tokens. Otherwise, the word is divided into tokens with each subword, except the first subword, getting prefixed by ##. A sample from the tokenization of Custom-Vocab is shown by Table - I. Using the original BERT vocabulary, the BERT tokenizer is run on the custom-vocab produced in the previous step. In the output, all those words which were retained as complete words were removed while the rest were stored in a CSV file with two columns: first column being the original word in Custom Vocab and second column being the tokenized form generated by BERT tokenizer. Around 17 thousand words were broken down into more than one tokens by BERT indicating that about 53% words in custom-vocab did not overlap with BERT vocabulary. As expected, these words comprised primarily of nuclear jargons and hence were not present in BERT off the shelf vocabulary. On a high level, this process segregates domain specific words which need special attention, as these are the set of words which are considered frequently occuring and important by Sentence Piece Library and are not part of BERT Vocabulary meaning that BERT could perform poorly on text containing a lot of them.

With the help of domain experts, all the 17 thousand words outputted in the previous step were segregated into two groups: "good" and "bad". Words in "Good" category meant that the tokenization is either successfully extracting the root word or is able to break the word into tokens which can jointly

"GOOD" Words	BERT Tokenization Split
electrochemical	'electro', '##chemical'
conductivity	'conduct', '##ivity'
neutrons	'neutron', '##s'
ultrasonic	'ultra', '##sonic'
shutdown	'shut', '##done'
exchanger	'exchange', '##r'
polarisation	'polar', '##isation'
radiography	'radio', '##graphy'
warranty	'warrant', '##y'

TABLE II: Table showing tokenization by BERT tokenizer of the "GOOD" Words. Note that BERT tokenizer performs reasonably well on these words, and thus shouldn't be added to NukeVocab.

"BAD" Words	BERT Tokenization Split
lethargy	'let', '##har', '##gy'
lubricant	'lu', '##bri', '##can', '##t'
lubricated	'lu', '##bri', '##cated'
lubrication	'lu', '##bri', '##cation'
luminescence	'lu', '##mine', '##sc', '##ence'
machining	'mach', '##ining'
ferritic	'fe', '##rri', '##tic'
vapour	'va', '##pour'
flange	'fl', '##ange'

TABLE III: Table showing tokenization by BERT tokenizer of the "BAD" Words. Note that BERT tokenizer performs very poorly on these words, and thus should be added to NukeVocab.

represent the word's actual meaning (Table-II). This further means that those words would most likely converge to their proper embedding after pretraining on domain specific text. On the other hand, words in "Bad" category meant that BERT tokenizations generates near meaningless tokens which would most likely fail to represent the meaning of the whole word. (Table-IV)

After segregating the words into "Good" and "Bad" category, lemmatization on the "Bad" category words is performed to prevent duplication of words with same roots. For example: 'lubricant,' 'lubrication,' lubricated' were replaced by a single word 'lubric.' This is especially important as it effectively increases the number of times the word in dictionary would be referred while training, thus enabling a more robust representation of the word in embedding space. For instance, the vocabulary word 'lubric' will be referenced everytime whenever either of the words 'lubricant', 'lubrication' or 'lubricated' appears in the text, and hence would have greater probability to converge to a more meaningful representation. Following the above procedure, 429 words were filtered from the "bad" category. Bert Vocabulary contains about 1000 "UNUSED" tokens whose word embeddings are not trained. We replace 429 "UNUSED" tokens with the selected 429 words from "bad" category, hence completing the formation of NukeVOCAB.

IV. NQUAD: Nuclear Question Answering Dataset

We decided to evaluate the performance of the newly generated vocabulary on question answering task, which is considered as one of the most difficult and practical task in Natural Language Processing research community. But unlike medical and general domain, the nuclear domain does not have any open source Question Answering Dataset. Such a dataset could be extremely useful to nuclear research community as a benchmark to evaluate their models. Moreover, it could be useful for numerous practical applications in nuclear engineering, such as identifying the right tools and chemical compounds needed to perform an experiment using just the text manual feeded in an artificially intelligent system. For these reasons, we propose a novel, high-quality Question Answering dataset as part of this research work.

One way to generate such a question answering dataset could be using the grammatical structure of the language domain to parse the text and generate questions from the parsed output. A lot of research has been done in this regard ([18], [19], [20]), which takes a paragraph as input and generates general questions. But this approach is infeasible for complex fields like nuclear domain because of two reasons. First, the existing parsers failed to parse the complex, jargon-laided nuclear texts into meaningful structure thus leading to generation of garbage questions. Even the few meaningful questions generated by above techniques were too simple to serve as a comprehensive evaluation benchmark. Even more important reason not to use synthetic question answering corpus is that solving those tasks usually require just the knowledge of sentence syntax rather than the semantic of the sentence. Although human engineered datasets are substantially smaller in size as compared to synthetic ones, they are typically more meaningful, useful and representative of the learning of the system. Hence, it was decided to generate an expert-driven question answering dataset, which will be released for fostering further research.

The question-answering dataset is built with a format similar to SQuAD dataset. For that, research papers were randomly selected out of the 7000 research papers corpus. From these research papers, around 200 paragraphs were randomly selected to form the questions on. Around 50 paragraphs were distributed to each domain expert, asking them to create questions on the paragraphs. They were encouraged to develop questions in their own words, the only restriction being that the answer must exactly lie within the paragraph as followed by SQuAD. The advised answer length was 8-10 words, though, in the final dataset, some accepted answers were longer than that also for generalization purposes. Experts found some paragraphs not suitable for forming questions, and hence,

Context Paragraph	Questions	Answers
Advanced Heavy Water Reactor (AHWR) fuel consists of (Th, Pu)O2 and (Th, 233U)O2. Thermochemistry of the compounds formed by the interaction of mixed oxide fuel and fission products is important in predicting the behaviour of this fuel in the reactor. Studies on the thermochemical properties of compounds in rare earth tellurium–oxygen system are being carried out in our laboratory in order to predict the chemical behavior of the fission products. Cerium and tellurium are amongst the fission products which are formed during the burn up of the nuclear fuel with significant yield. Tellurium is corrosive in nature and can exist in various chemical states in irradiated fuel. It can cause embrittlement of clad.	What is the fuel used for Advanced Heavy Water Reactor (AHWR)? What works are carried out in the laboratory in order to predict the chemical behavior of the fission products? What are formed during the burn up of the nuclear fuel with significant yield?	fuel consists of (Th, Pu)O2 and (Th, 233U) Studies on the thermochemical properties of compounds in rare earth tellurium–oxygen system are being carried out in the laboratory. Cerium and tellurium

TABLE IV: Table showing a sample paragraph and question answer pairs from NQuAD.

those were discarded. Finally, we were able to generate 612 questions on 181 paragraphs (Figure - **??**).

The paragraph-questions-answers were recorded in shared google docs, responses from which were merged into an excel file. The file was randomly divided into two sections: a train set of 155 paragraphs, 536 questions, and dev set of 26 paragraphs, 76 questions.

To fine-tune the question-answering platform after training on SQuAD dataset, the custom-made Paragraph-Question-Answer dataset was converted to JSON with the structure similar to SQuAD v1. For turning it into SQuAD format, an already available platform for question generation was adapted for generating the NQuAD dataset. It is built using Angular as frontend with a spring boot backend using MongoDB database.

Since the dev set requires three answers per question, domain experts were asked to answer each of the 100 questions individually. The generated excel was again was converted into JSON format using the question generation platform described above.

V. Experiment

A. BERT PRETRAINING

Pre-Training BERT is a computationally expensive job and requires a large amount of data to pretrain. Our corpus of 8 Million words is two orders of magnitude less than the 3 Billion words corpus used by original BERT. It would be infeasible to pretrain BERT from scratch on this small corpus. Also, as mentioned earlier, there was almost 47% overlap with the BERT vocab, which have already converged to proper representation in the pretrained weights. Pretraining from scratch would mean throwing away those learned embeddings. Pre-training from scratch becomes infeasible for domains like Nuclear, where the amount of data is pretty less. To tackle this issue, we used Nuclear Vocabulary (NukeVOCAB) in place of BERT vocabulary and pretrained it on the custom preprocessed corpus starting from BERT checkpoint. Due to Out of Memory issues, we used the batch size of 128 with maximum length set at 128. We pretrained it till the training loss more or less stopped decreasing, which happened after approximately 600000 steps. It took around 54 hours on a single Google Colab TPU for pre-training BERT. We will be referring this model as NukeBERT for further discussions.

B. BERT Fine-Tuning

For the base model, BERT was fine-tuned on SQuAD using the same configurations used by original BERT paper. On testing on NQuAD dev set, it scored 83.11 on exact match score and 92.66 on F1 score. The fine-tuning took around 8 hours on Google Cloud's Tesla k80 GPU (It is the version of GPU provided for free by google colab).

With the same configuration as used by original BERT for fine-tuning, NukeBERT was fine-tuned on SQuAD dataset for two epochs. It took around 8 hours to train on Google Cloud GPU.

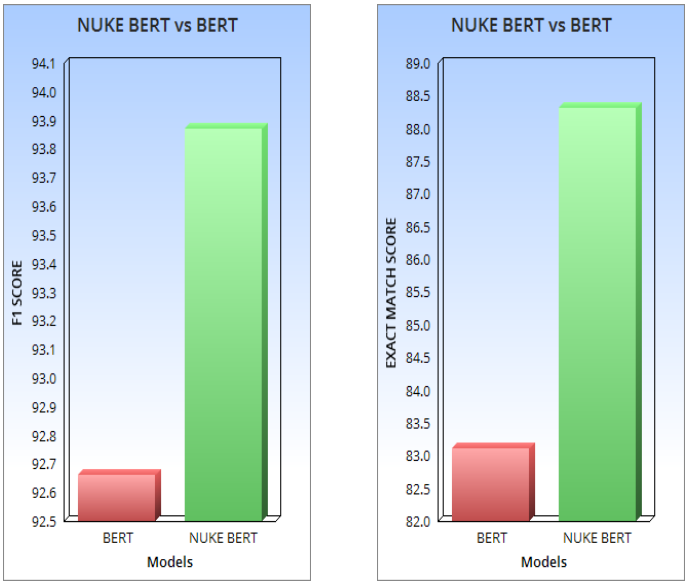
Further NukeBERT was fine-tuned on the NQuAD train set for three epochs at a learning rate of 3e-6.

VI. Results

On testing, the NukeBERT achieved F1 score of 93.87 and exact match score of 88.31. Hence the model was able to

Paragraph	Questions	Answer by NukeBERT
Walters and Cockraft (1972) might have been the first ones who used finite element method in analysing the creep of weldments.Coleman et al. (1985) incorporated a three-material model and used a parametric approach to cover a wide range of weld metal:base metal creep ratios. Ivarsson and Sandstrom (1980) studied the creep stress redistribution and rupture of welded AISI 316 steel tubes by finite different method.Ivarsson (1983) studied the creep deformation of welded 12% chromium steel tubes. Eggeler et al. (1994) performed a creep stress analysis of the welded pressure vessel made of modified 9Cr-1Mo material (P91) based on Norton's creep law and specimen were studied using finite element by Segle et al. (1996) for the 1Cr-0.5Mo butt welded pipe.	Who used for the first time the finite element method in analysing the creep of weldments?	walters and cockraft
	Which method is used in analysing the creep of weldments?	finite element method
	Who used a three-material model for weld metal:base metal creep ratios ?	coleman et al.
	What does Sandstrom (1980) studied?	creep stress redistribution and rupture of welded aisi 316 steel tubes
	Who was associated with the finite different method?	ivarsson and sandstrom

TABLE V: Table showing qualitative performance of NukeBERT on unseen nuclear language passage.



(a) F1 scores inin (b) Exact Match

Fig. 4: Results of BERT and NukeBERT on NQuAD test-set

achieve 5.21 improvement on exact match criteria and 1.22 improvement on F1 score (Figure - 4).

VII. Conclusion

BERT requires large data for pre-training and generating meaningful word embeddings. However, the availability of data becomes a bottleneck for domains like nuclear whose data is profoundly different from world language. Hence, for areas like Nuclear energy, it becomes essential to use methods which we used for NukeVOCAB generation to optimize the use of meagre data. The improvement in F1 and exact match score validates the NukeBERT embeddings, which can be used for other downstream tasks too like sentence classification, named entity recognition among many NLP tasks. NukeBERT can be further generalized to many downstream tasks like named entity recognition, sentence segmentation among many, which could be highly important for nuclear industry. Moving further, the ideas used in this paper can be utilised with new transformers that have come up in recent research. Data availability and creation in nuclear field, needs attention. This paper aims at providing momentum to research in nuclear domain.

Acknowledgment

We thank all colleagues of Resource Management Group, IGCAR and BITS Pilani for their support and encouragement.

References

1 Deo, R. C., "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.
2 El-Sappagh, S., Elmogy, M., and Riad, A., "A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis," *Artificial intelligence in medicine*, vol. 65, no. 3, pp. 179–208, 2015.
3 Seebode, C., Ort, M., Álvarez, S. G., and Regenbrecht, C., "Biobank semantic information management with the health intelligence platform," *Diagnostic Pathology*, vol. 1, no. 8, 2016.

4 Meenachi, N. M. and Baba, M. S., "A survey on usage of ontology in different domains," *International Journal of Applied Information Systems*, vol. 4, no. 2, pp. 46–55, 2012.

5 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

6 Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P., "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

7 Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

8 Iacobacci, I., Pilehvar, M. T., and Navigli, R., "Sensembed: Learning sense embeddings for word and relational similarity," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 95–105.

9 Reisinger, J. and Mooney, R., "Multi-prototype vector-space models of word meaning," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 109–117.

10 Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

11 Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I., "Improving language understanding by generative pre-training," 2018.

12 Beltagy, I., Cohan, A., and Lo, K., "Scibert: Pretrained contextualized embeddings for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

13 Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J., "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

14 Smith, R., "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

15 Namysl, M. and Konya, I., "Efficient, lexicon-free ocr using deep learning," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 295–301.

16 Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J., "Tensor2tensor for neural machine translation," *CoRR*, vol. abs/1803.07416, 2018. [Online]. Available: http://arxiv.org/abs/1803.07416

17 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M., "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

18 Heilman, M., "Automatic factual question generation from text," Ph.D. dissertation, Ph. D. thesis, Carnegie Mellon University, 2011.

19 Aldabe, I., De Lacalle, M. L., Maritxalar, M., Martinez, E., and Uria, L., "Arikiturri: an automatic question generator based on corpora and nlp techniques," in *International Conference on Intelligent Tutoring Systems*. Springer, 2006, pp. 584–594.

20 Rakangor, S. and Ghodasara, Y., "Literature review of automatic question generation systems," *International journal of scientific and research publications*, vol. 5, no. 1, pp. 1–5, 2015.