Support Cybersecurity Risk Public Awareness with AI Machine Comprehension

Gongqi Lin[*], Yuan Miao[*], Yidan Hu[+], and Zhiqi Shen[+ ^]

[*]Victoria University, Melbourne, Australia
[+]Nanyang Technological University, Singapore

lingongqi2009@gmail.com

Abstract

Public awareness or public education plays an important role in cybersecurity. However, public education on cybersecurity not only is expensive but also faces the already alarming shortage of cybersecurity professionals. In addition, the fast evolution of information technology led to extraordinary update rate of cyber risks. Many people do not know whom to ask when they have questions concerning their system or account security. Inspired by the recent success of machine reading comprehension (MRC) on documents that were only 'comprehensible' to human users, this paper explored the potential of turning technical documents into a large source of knowledge that can be used to provide answers to the public. We call this Cybersecurity Document Reading Comprehension (CDRC). To the best of our knowledge, this is the first attempt applying machine reading comprehension on cybersecurity documents to provide affordable and easy to access advice to the public or certain groups. In this work, we built an CDRC dataset and an AI comprehension algorithm. The algorithm was trained on the CDRC dataset and can reach the Reading Comprehension Ability Test (CAT) level 2. At this level, the algorithm is able to comprehend and identify facts presented in cybersecurity articles. Therefore, for each type of risks or prevention mechanisms, as long as there are technical articles available, the contained knowledge can be comprehended by the algorithm, based on which to provide answers and advice to the public. Our algorithm is based on the

^ Corresponding author

recent breakthrough of the language model BERT. We improved the basic BERT model by integrating it with common knowledge model and trained it with CDRC dataset. Experimental results demonstrated that basic BERT has performance similar to human readers if the questions are asked with proper terms used in the technical articles. However, BERT often fails on handling inquiry with different terms or non-technical language while our BERT + Common Knowledge model performs well.

Keyword: Cybersecurity, Public Awareness, Machine Reading Comprehension, Reading Comprehension Ability Test, BERT, Common Knowledge.

I. Introduction

Lacking knowledge or awareness of cybersecurity has become one of the highest risks in organisations. Many major leakages or breaches start from human users neglection or lack of knowledge about the risks. Both the recent Equifax breach (145 million US consumers records) and the SA Master Deeds Leak (30 million South Africans ID numbers) [18] were caused by human error. If public users (potential victims) are able to easily acquire crucial information about cybersecurity risks and best practices, it can help them protect from attacks. Governments and commercial organizations around the globe make extensive use of Information and Communications Technologies (ICT), and as a result, their security is of utmost importance. To achieve this, they deploy technical security measures, and develop security policies to regulate the behaviour of employees, consumers and citizens. Unfortunately, many individuals do not comply with the policies or expected behaviours [1]. There are many possible reasons for this, but two of the most compelling are that people are not aware of (or do not perceive) the risks or, they do not know (or fully understand) the policy.

Security awareness is defined in NIST Special Publication 800-16 [2] as follows: “The purpose of awareness presentations is simply to focus attention on security. Awareness presentations are intended to allow individuals to recognize IT security concerns and respond accordingly”. It identifies the fact that people need not only to be aware of possible cyber risks but also, behave accordingly. In recent

years and likely in the near future, major cyber security attacks have frequently occurred and will continue to occur [3]. A likely reason for this could be the fact that attackers are becoming more skilled. At the same time, the security knowledge is often too difficult or too technical for the public. However, the training and education of the public is not only costly but also takes time. If AI algorithms can 'read' vast number of technical documents of cybersecurity and answer related questions for the public, it will be very helpful for people to 'consult' the virtual experts or AI readers, clear their doubts， gain awareness and act accordingly to reduce their risks .

Reading Comprehension Ability Test, or reading CAT, is proposed by Yuan and et. al [17] to determine whether an AI algorithm has gained similar comprehension ability as human beings. The test setting is similar to Turing Test, but the human questioner does not only issue a series of questions, but a series of articles and the corresponding question sets, as shown in Figure 1. Each reading CAT has a number of test units. Each test unit consists of an article and a set of questions. By examining the answers, if it cannot tell whether the answers were produced by a human being or an AI algorithm, the algorithm has passed the test. The reading CAT is classified into four levels [17]. The first level is able to locate fact as is and answer the corresponding questions. BERT [5] based algorithms have made breakthrough at this level (level 1) and a number of algorithms at the leader board have beat humans. However, some simple replacement of the terms, such as 'a host' and 'a server' often being exchanged in human communication, can totally confuse these algorithms. Similarly, although there are clear technical differences between a malware and a virus, many users do not differentiate them. In that case although the article carries the fact, AI algorithms can fail to answer. This requires the next level of reading CAT (level 2).

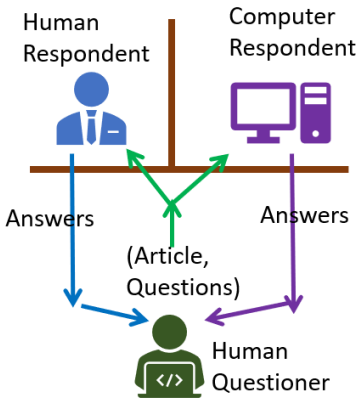

Figure 1. CAT - Comprehension Ability Test

In this article we introduce the Cybersecurity Documents Reading Comprehension (CDRC) for common knowledge question answering in the style of SQuAD 2.0. We propose a novel method for question generation, in which human annotators works at the CAT level 2, and are asked to submit questions using common knowledge with provided cybersecurity articles. The main contributions of our works are as followings:

- This is the first attempt applying reading comprehension on cybersecurity documents with common knowledge.
- We provided case studies on cybersecurity question and answering.
- A new method for generating commonsense questions at scale fro SQuAD 2.0.
- We verified the popular language model, BERT, on our cybersecurity dataset using the way as SQuAD 2.0, which achieves high performance.
- We experimented common knowledge generation procedure produced a dataset, found a large gap in accuracy between algorithms and human experts, including state-of-the-art pre-trained language models.

II. Related Works

Given the context that cybersecurity is gaining comprehensive concerns and the shortage of human resources, there is an increasing interest to address cybersecurity challenges using natural language processing (NLP) techniques, even though it is still at a very initial stage. Lim et al. [6] has introduced

a dataset of annotated malware reports for facilitating future NLP work in cybersecurity. Rieck et al. [7] and Alazab et al. [8] proposed models using machine learning techniques for detecting and classifying malware through system calls.

2.1 Reading Comprehension

Machine common sense to reason about an open-ended world has long been acknowledged as a critical component for natural language understanding. Early works focus on an environment in natural language [19] or leverage a world-model for deeper language understanding [20]. Many commonsense representations and inference procedures have been explored [21-23]. However, evaluating the degree of common sense possessed by a machine remains difficult.

Researchers have also investigated question answering that utilizes common sense. Science questions often require common sense, and have recently received attention [24]. In this article we present a work applying machine comprehension to help with cybersecurity awareness and public education. Our model is based on a deep neural network model called BERT. BERT has the best performance for many natural language process tasks. Its reading comprehension ability to identify ground truth facts in articles has outperformed human users. However, BERT is a statistical pattern-based model which matches words with the highest similarity in the corresponding sentences. It works very well on a number of widely used datasets but could fail badly in providing answers to public inquiries. When the public phrase the question with different terms, we human beings can easily understand. But for BERT, if the term is not part of the article, there will be no answers. Our experiments have confirmed this hypothesis.

SQuAD [9] and the Google Natural Questions datasets [10] are two large datasets which are widely applied. SWAG dataset situated common-sense inference, with 113k multiple-choice questions [11]. BERT performs very well in all of the three datasets. Commonsense Dataset Adversarially-authored by Humans (CODAH) [12] is a dataset which contains 2.8k questions. It strengthened the requirement on inference and reduced the possibility of matching statistically (as compared to the SWAG dataset). However, none of the datasets have a focus on cybersecurity.

2.2 CAT Level 2 Common Knowledge

To enable machine comprehension to read cybersecurity articles, provide answers or education to the public, the algorithm needs to be able to handle diverse terms even though they may refer to the same meaning. This ability is at level 2, according to reading CAT, while BERT is essentially a level 1 algorithm. We have modelled the corresponding common-sense knowledge, created a dataset which have questions requiring the application of common-sense knowledge, experimented with BERT. BERT had a very poor performance. We enhanced the solution to address the problem. A 'equivalence' type, which common knowledge of the dataset is essentially, has three specific types: synonym, definition and attribute.

2.2.1 Synonym

Synonym is the first type of "equivalence" common knowledge. We replaced the five types of words, namely verb, noun, adjective, adverb and preposition, with their synonyms in the dataset. We changed more than one word in one question to avoid BERT hitting the right answer for the wrong reason. When only one word was replaced in the questions, NLP models like BERT have a good chance to locate the answer by matching the similarity of the other words in question and paragraph. BERT was not heavily affected in long questions when we only changed one word. This replacement is a kind of disturbance to verify (and improve) the robustness of the NLP models like BERT and adapt BERT to different situations. For example, given the question "What the information technology is used to protect the threat in the computer world?", we changed "protect" to "prevent" (a synonym) and "damage" to "threat" (a synonym). The new question is "What the information technology is used to prevent the threat in the computer world?". Experiment shows that BERT is no longer able to answer. We solved the problem by augmenting BERT with common knowledge.

In this article, we use the WordNet [25] as our reference library. WordNet is a lexical database of English, where words are organized into synsets according to their senses. A synset is a set of words expressing the same sense so that a word having multiple senses belongs to multiple synsets, with each synset corresponding to a sense. Synsets are further related to each other through semantic relations.

According to the WordNet interface provided by NLTK [26], there are totally sixteen types of semantic relations (e.g. hypernyms, hyponyms, holonyms, meronyms, attributes, etc.). Based on synset and semantic relation, we define a new concept: semantic relation chain. A semantic relation chain is a concatenated sequence of semantic relations, which links a synset to another synset. For example, the synset "keratin.n.01" is related to the synset "feather.n.01" through the semantic relation "substance holonym", the synset "feather.n.01" is related to the synset "bird.n.01" through the semantic relation "part holonym", and the synset "bird.n.01" is related to the synset "parrot.n.01" through the semantic relation "hyponym", thus "substance holonym → part holonym → hyponym" is a semantic relation chain, which links the synset "keratin.n.01" to the synset "parrot.n.01". We name each semantic relation in a semantic relation chain as a hop, therefore the above semantic relation chain is a 3-hop chain. By the way, each single semantic relation is equivalent to a 1-hop chain.

2.2.2 Definition

Definition is the second highest frequently used common knowledge [13]. To add this type of common knowledge to the dataset, we replace some words with their definitions. In this case, we only need to change one or two words in one question. Because the definition is normally much long, the length of questions clearly changed. We start with this level of changing to test the performance of BERT, which confirm ours thought. With type of this changes, human users can easily to watch and find the answer. When users cannot remember the term and need to search the related questions, this type of change will be effective to test algorithm comprehension ability. For example, given the question "What causes a machine or network resource unavailable to its intended users?", we changed "unavailable" to "out of service" (a definition). The new question is "What causes a machine or network resource out of service to its intended users?" Bert normally fails but human users had no extra difficulties.

2.2.3 Attribute

A word can be many related semantic words features and properties. Attribute is a widely existent relation. We will take "snow" as an example to explain attribute. We all know that the color of snow

is white. In this sentence, “color” is as an attribute of snow. When we see the word “snow” in a paragraph, we can associate snow with the color of white if the content is about color. It is one way we combine the current reading with our common-sense knowledge. With the help of common-sense knowledge, we can answer the question and understand the paragraph more easily. However, being able to apply common knowledge is challenging to NLP models. We added common knowledge of attribute in our dataset and also try to train a more intelligence model. For example, given the question “Which way makes victims account to be locked?”, we changed “account to be locked” to “cannot log in to the system” (an attribute). “account” has an attribute of “login system”, and generated the new question, “Which way makes victims cannot log in to the system?”

III. Machine Comprehension of Cybersecurity Articles

In this section, we will explain our machine comprehension of cybersecurity articles through case studies.

3.1 Machine Comprehension - BERT

BERT is one of the most notable progresses in contextualized representation learning. When it was proposed recently, the performance in comprehension task over SQuAD scored 80+, which is close to human users’ performance. In 2019, a few variations of BERT have exceeded human users’ performance, reached a score of 90+. That is to say, over 90% of the article questions in the dataset can be answered correctly by BERT. Its reading comprehension ability has passed CAT level 1.

The idea behind BERT is that even though the word embedding layer (in a typical neural network for NLP) is trained from a large-scale corpus, training a wide variety of neural architectures that encode contextual representations only from the limited supervised data on end tasks is insufficient. BERT adopts a fine-tuning approach for it to suits different tasks. The merit of the model is that BERT requires almost no specific architecture for question answering tasks. This is desirable because intelligent agents should minimize the use of prior human knowledge in the model design. Instead, it should learn such knowledge from data. Therefore, the model can be applied to a wide range of reading comprehension tasks. We do not need to design different models for each specific task. Figure 2 shows the structure of BERT.

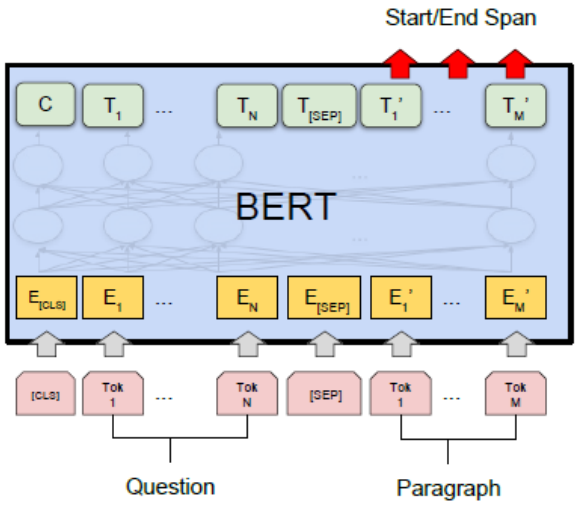

Figure 2. BERT on SQuAD [5]

BERT has two parameter intensive settings:

$BERT_{BASE}$: has 12 layers, 768 hidden dimensions and 12 attention heads (in transformer) with the total number of parameters of 110M;

$BERT_{LARGE}$: has 24 layers, 1024 hidden dimensions and 16 attention heads (in transformer) with the total number of parameters of 340M.

3.2 Case study on Cybersecurity Question Answering

We have developed and trained a BERT based comprehension model with a set of articles in cybersecurity. It is to verify and demonstrate the approach for machine comprehension to support public awareness of cybersecurity risks. In this section we use two paragraphs as case studies to illustrate the method.

Case Study 1: Distributed Denial of Service (DDoS) is one of the most powerful (damaging) weapon on the Internet. In a typical DDoS attack, the assailant begins by exploiting vulnerability and breaks in one computer system, makes it the DDoS master. The DDoS master identifies other vulnerable systems and gains control over them by either infecting the systems with malware or through bypassing the authentication controls, *i.e.*, guessing the default password on a widely used system or device. For example, a recent study found that 23 million users who were attacked by using a password '123456'.

> **Paragraph 1**: Denial of service attacks are designed to make a machine or network resource unavailable to its intended users. Attackers can deny service to individual victims, such as by deliberately entering a wrong password enough consecutive times to cause the victim account to be locked, or they may overload the capabilities of a machine or network and block all users at once. While a network attack from a single IP address can be blocked by adding a new firewall rule, many forms of Distributed denial of service (DDoS) attacks are possible, where the attack comes from a large number of points \u2013 and defending is much more difficult. Such attacks can originate from the zombie computers of a botnet, but a range of other techniques are possible including reflection and amplification attacks, where innocent systems are fooled into sending traffic to the victim.

However, many people have little knowledge about it, 'Why is it so difficult?', 'How could my security camera become an attacker?' and etc. As shown above, Paragraph 1 is about DDoS extracted from SQuAD.

There are many technical articles cover the knowledge about DDoS. However, not many public people have the time or background knowledge to read these articles. Instead, when they have an inquiry, they hope a cybersecurity professional could answer it. However, in practice such an inquiry is often left unanswered because the significant shortage of the cybersecurity skills. Now our algorithm can read the above article, as well as people's questions. The manpower also can answer with its comprehension, which is at the same level of human readers. For example, if it is questioned with

- 'What makes defending in DDoS attacks much more difficult?'

The answer will be

- 'Because the attack comes from a large number of points.'

Although we did not include the related paragraph/sentences as the further explanation, there is no essential difficulties for the algorithm to provide the elaboration accordingly.

In the above question, we can see that the terms used in the question are consistent with what used in the article. This is often not true when public people making inquiries. For example, the question can be:

- What makes defending harder in DDoS attacks?

Although it appears to be the same questions to human readers, it is very different to BERT algorithms. Our experiment shows that BERT failed badly with similar questions. 'Hader' and 'difficult' are synonyms, which is a common knowledge people have. BERT does not have common knowledge and thus is not able to answer. BERT is only at the reading CAT level 1. We applied the following integration rule to our model.

> Integration Rule 1: Replace the key words with their synonyms in given cybersecurity related questions or articles, if the original version produced no answer.

Experiment shows that for this type of freely formal questions, our model can handle well. It has reached reading CAT level 2.

Case Study 2: Phishing is a cyber attack that uses disguised email as a weapon. The goal is to trick the email recipient into believing that the message is something they want or need and to click a link or download an attachment.

However, many people have little knowledge about phishing scam, even never heard this hacker skill. The following is a paragraph about phishing kits from a security expert's blog.

> **Paragraph 2**: Phishing is a cyber attack that uses disguised email as a weapon. The availability of phishing kits makes it easy for cyber criminals, even those with minimal technical skills, to launch phishing campaigns. A phishing kit bundles phishing website resources and tools that need only be installed on a server. Once installed, all the attacker needs to do is send out emails to potential victims. Phishing kits as well as mailing lists are available on the dark web. A couple of sites, Phishtank and OpenPhish, keep crowd-sourced lists of known phishing kits.

Phishing is one of the most damaging attacks. In the modern world, access is everything and phishing attacks give attackers access. Login credentials, account numbers, social security numbers, email addresses, phone numbers and credit card numbers are all pure gold to scammers – they want to steal any information that will give them access to victims' accounts. Therefore, all users should be aware and keep alert of the danger to ensure that potential scammers could not obtain access to their personal information. In this case, they may desirably want to ask security questions about how to avoid becoming a phishing victim. Our system provides users the knowledge of cybersecurity, potential risks, possible symptoms of attacks, best practices and other popular (beneficial) security topics. Users can request for more knowledge by asking (a sequence of) further questions. For example, if the system is questioned with

- 'What is a cyber attack using email as a media?'

By reading the above article (machine reading and comprehension), the answer will be

- ‘Phishing.’

In the above question, we can see that users may use the term ‘media’, but experts will normally use the term ‘weapon’ in their article. The equivalence of the two term in this context is a common knowledge to us.

The user may want to continue with a question as:

- ‘Which tool is quite possible for cyber criminals on phishing?’

The answer would be

- ‘Phishing kits.’

We applied the following integration rule to our model.

> Integration Rule 2: Replace the key words with their definition in the given cybersecurity related questions or articles if the original result was no answer.

Experiments show that for this type of questions, our model can handle well but BERT cannot answer. This is because answering the questions requires (the integration of) common knowledge.

3.3 Constructing Pre-training Dataset

The system uses questions from SQuAD 2.0 as an input, and feeds synsets with context to generate synonyms based on Wordnet. Synsets are interlinked by means of conceptual-semantic and lexical relations, and is a presentation with context in our paper, such as word ‘plant’ has five synsets in Wordnet, shown in Table 1. The most frequently encoded relation among synsets is the super-subordinate relation, which called hyperonymy, hyponymy or ISA relation.

Synset	Context	Example
Plant.n.01	buildings for carrying on industrial labor	They built a large plant to manufacture automobiles.

Plant.n.02	(botany) a living organism lacking the power of locomotion	There are a lot of plants in my garden.
Plant.v.01	Put or set seeds, seedings or plants into the ground	Let's plant flowers in the garden.
Plant.v.02	Fix or set securely or deeply	He planted a knee in the back of this opponent.

Table 1. Sample of Synsets

We apply three rules of CAT Level 2 Common Knowledge, shown in Section 2.2 to process questions, and we define three main rules to filter the specific phrases to enhance the accuracy. One rule is built using part-of-speech tagger [26], such as proper noun. The second rule is built for phrases, such as the upper/lower house that means the legislative context in together, and we build it by ourselves. The last rule is based on spacy token attributes [29], such as filtering number or like number words, stop words, etc. The system needs to keep proper nouns or commonly used phrases, which cannot be used to generate synonyms. Each question will be tokenized as a single word or a phrase with nltk toolkit [26].

3.4 Fine-tuning BERT for Cybersecurity Article Comprehension

We mostly follow the same architecture, optimization, and hyperparameter choices used in [5]. We differ slightly in using an additional conditional random field, which made evaluation easier by guaranteeing well-formed entities. In all settings, we apply a dropout of 0.1 and optimize cross entropy loss using Adam [27]. We finetune for 2 to 5 epochs using a batch size of 32 and a learning rate of 5e-6, 1e-5, 2e-5, or 5e-5 with a slanted triangular schedule [28], which is equivalent to the linear warmup followed by linear decay. For each dataset and BERT variant, we pick the best learning rate and number of epochs on the development set and report the corresponding test results. We found the setting that works best across most datasets and models is 2 or 4 epochs and a learning rate of 2e-5. While task-dependent, optimal hyperparameters for each task are often the same across BERT variants.

We did not perform extensive hyperparameter search, but while optimal hyperparameters are going to be task-dependent, some light experimentation showed these settings work fairly well across most tasks and BERT variants

Following the BERT's success on SQuAD, we designed a question answering dataset in cybersecurity area, integrated the common-sense knowledge using the method described in Section 3.3. As a proof of concept cybersecurity QA, we plugged in the task-specific inputs and outputs into BERT and finetuned all the parameters end-to-end. At the input, sentence A and sentence B from pre-training are analogous to question-passage pairs in question answering. At the output, the token representations are fed into an output layer for question answering.

IV. Experiment

In the experiments, we designed the questions into different categories by applying different common knowledge to build dataset on cybersecurity. The idea was discussed in Section 3 with case studies. We evaluated the dataset with state-of-the-art neural question answering systems built on the BERT architecture and provided a baseline. The models and experiment setups are discussed below.

4.1 Question Categories on Cybersecurity Dataset

Our goal is to develop a machine comprehension system to support public awareness of cybersecurity risks. For that, we will need to analyze how system and human performance varies across questions in CDRC applying different types of common knowledge. We included the three types of common knowledge discussed in Section 2, *i.e.*, synonym, definition and attribute. These rules were only applied to a small portion of the questions in our dataset, but have already led to a much poor performance of the existing language models.

We manually inspected all questions in our dataset and annotated some with one or more rules reflecting the commonsense knowledge. We selected general articles about cyber security from Wikipedia, which can be easily understand by the public. We also picked technical articles from expert blogs. As shown in Table 2, general questions on Cyber Security Dataset (CSD) occupies 75.26% of

the total questions, Phishing Knowledge Dataset (PKD) takes up 7.84% as advanced technical questions, 16.9% of total questions on general Computer Security Dataset are integrated with Common Knowledge (CSD-CK). CSD-CK questions will help to verify the ability to handle questions requiring common knowledge. We will discuss them in the following sections.

Topics	No. of Questions	Percentage (%)
Cyber Security without Common-sense Knowledge	365	75.26
Phishing Knowledge, 20% involving Type 1 Common-sense Knowledge	38	7.84
Cyber Security involving Type 1, 2, and 3 Common-sense Knowledge	82	16.9

Table 2. Distribution of the Dataset on Cybersecurity

4.1.1 Cyber Security Dataset (CSD)

Knowledge sharing in information security is often achieved through training and provision of security policy procedures [14]. Knowledge sharing among employees can enhance their understanding and promote their commitment [14]. However, the complexity and technical aspects of information security knowledge are often seen as prime inhibitors among employees [15]. Our cybersecurity question and answering system will help people the needed cyber security information easily. Based on Wikipedia articles related to cyber security, we developed 365 questions in the similar way as SQuAD 2.0. We want to verify BERT's ability in handling cyber security questions at CAT level 1. This dataset includes 25 paragraphs, 248 unanswerable questions and 117 answerable questions, in total 365 questions.

4.1.2 Phishing Knowledge Dataset (PKD)

Email communications is critical to nearly every business and individual. A person's email address is the online identity for reliable and accountable communications. Most of our activity, such as banking,

healthcare and recreation are conducted online, which require us to communicate with emails. Unfortunately, email inherently possesses security flaws that malicious attacks can create fraud identity through phishing emails. Public awareness is crucial to solve this issue. Users need 'knowledge' on how to identify and respond to suspicious phishing emails. We selected phishing related topics from the Australian government website [16] to build a question and answer dataset. The dataset includes 15 paragraphs, 32 answerable questions and 6 unanswerable questions. Further, we applied synonym rules on 20% questions of the dataset. This is a type of common knowledge that people often apply in making inquiries. It is very unlikely that all users will use the standard forms or terms in the articles to ask the questions.

4.1.3 Cyber Security Question Dataset with Applying Common Knowledge (CSD-CK)

To compare with the original CSD, we selected all paragraphs from CSD, and built around 2 to 5 questions for each paragraph, which requires common knowledge in total we added 82 answerable questions. In this dataset, the common knowledge applied are Synonym, Definition, and Attribute Rules (described in Section 2). The common knowledge is applied on each question, and the distribution is shown in Table 3. However, for fair comparison and easy analysis, we only applied one or two rules on original question from CSD.

Rules	Count	Percentage (%)
Synonym	50	61
Definition	20	24.4
Attribute	12	14.6

Table 3 Distribution of required GK on CSD

4.2 Environment Setup

We did the fine-tuning on a server with 4 * Telstra v100 GPUs which has 11GB RAM. The server is not powerful enough run $BERT_{LARGE}$ but the $BERT_{Base}$ results have been good enough. We verified our datasets using $BERT_{BASE}$ based on the pre-trained weights.

4.3 Hyper-parameters

We adopted $BERT_{BASE}$ (uncased) as the basis for all experiments. The maximum length of fine-tuning is set to 320 with a batch size of 8 for each type of rules. The number of sub-batch u was set to 2, which is good enough to store each sub-batch iteration into the GPU 11G memory. We used Adam optimizer and set the learning rate to 3e-5. We trained 68,000 steps for fine-tuning model based on SQuAD 2.0 dataset.

4.4 Evaluation Metrics and Model Selection

To be consistent with the existing research on MRC, we used the same evaluation script from SQuAD 1.1 for MRC, which reports Exact Match (EM) and F1 scores. EM requires the answers to have exact string match with human annotated answer spans. F1 score is the averaged F1 scores of individual answers. F1 score is typically higher than EM and is the main metric. The F1 score is the harmonic mean of precision and recall, which is computed based on the number of questions that the algorithm correctly answered and the number of questions that human users annotated with correct answers.

4.5 Result Analysis

We started with the vanilla BERT pre-trained weights and fine-tuned it on the three different datasets discussed in sub-session 4.4. We used this as a baseline to answer questions from cybersecurity area. The running results are shown in Table 4. BERT has an outstanding performance on CSD, achieves 96.33% F1 score. However, BERT had very poor performance when common knowledge is required. The F1 score drops to 63.59% F1 score when is 20% of questions used synonyms. BERT performance became close to random (20%-25%) when all the three types of CAT Level 2 common knowledge were involved. It only achieved 38.88% in the F1 score.

Metrics	CSD (%)	PKD (%)	CSD-CK (%)
exact	95.89	48.65	33.33
F1	96.33	63.59	38.88

Table 4 Results using BERT on three different question sets

By integrating the common knowledge with BERT, we achieved similar performance on PKD and CSD-CK as what BERT achieved on CSD, which is about 96% accuracy. Given that common knowledge we examined is only an 'equivalence' of terms people use, it has little 'formal' restriction. Unlike concpetNet which tries to minimise the number of relationships, there is no limitation or even prefer to have more diversities for the conditions or context that two terms are considered as 'equivalent'. Therefore, almost everyone can contribute to it.

V. Conclusions

We proposed a new task called Cybersecurity Documents Reading Comprehension (CDRC) and investigated the possibility of turning technical documents, such as government cybersecurity reports, as a valuable resource for answering user questions to improve the public awareness of cyber risks. We adopted BERT as our base model and proposed a method to integrate common knowledge with it to improve machine comprehension ability, particularly at CAT level 2. We further explored the use of BERT in CDRC dataset. Experiment results show that the effectiveness of BERT and its limitation on questions requiring common knowledge of 'equivalence' terms, which we have addressed with the integration of common knowledge.

References

[1] N. Humaidi, V. Balakrishnan: Exploratory Factor Analysis of User's Compliance Behaviour towards Health Information System's Security. J Health Med Inform 4(2), (2013).

[2] National Institute of Standards and Technology - NIST: Building an Information Technology Security Awareness and Training Program. M. Wilson, J. Hash, Computer Security Division Information Technology Laboratory. October 2003. http://csrc.nist.gov/publications/nistpubs/800-50/NIST-SP800-50.pdf

[3] I. Kirlappos, S. Parkin, M. Sasse, A.: Learning from Shadow Security: Why understanding non-compliance provides the basis for effective security. Workshop on Usable Security, 2014.

[4] J. R. Hackman, and N. Katz, (2010). Group Behavior and Performance. New York, NY: Wiley

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.

[6] S. K. Lim, A. O. Muis, W. Lu, and C. H. Ong. 2017. MalwareTextDB: A Database for Annotated Malware Articles. In Proc. of ACL, volume 1, pages 1557–1567.

[7] K. Rieck, P. Trinius, Carsten Willems, and Thorsten Holz. 2011. Automatic analysis of malware behavior using machine learning. Journal of Computer Security, 19:639–668.

[8] M. Alazab, S. Venkataraman, and P. Watters. 2010. Towards understanding malware behaviour by the extraction of api calls. In Proc. of CTC, pages 52–59.

[9] P. Rajpurkar, R. Jia, and P. Liang. Know What You Don't Know: Unanswerable Questions for SQuAD[J]. arXiv preprint arXiv:1806.03822, 2018.

[10] T Kwiatkowski, J Palomaki, and O Redfield. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.

[11] R. Zellers, Y. Bisk, and R Schwartz. Swag: A large-scale adversarial dataset for grounded common-sense inference[J]. arXiv preprint arXiv:1808.05326, 2018.

[12] M. Chen, M. D'Arcy, and A. Liu. CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense[C]//Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP. 2019: 63-69.

[13] P. LoBue and A. Yates. Types of common-sense knowledge needed for recognizing textual entailment[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011: 329-334.

[14] S.-K. Park, S.-H. Lee, T.-Y. Kim, H.-J. Jun, and T.-S. Kim. 2017. "A Performance Evaluation of Information Security Training in Public Sector," Journal of Computer Virology and Hacking Techniques (13:4), pp. 289-296.

[15] N. S. Safa, and R. Von Solms. 2016. "An Information Security Knowledge Sharing Model in Organizations," Computers in Human Behavior (57), pp. 442-451.

[16] Australia Cyber Security Center, https://www.cyber.gov.au/index.php/threats/phishing

[17] Y. Miao, G. Lin, Y. Hu, C. Miao. Reading Comprehension Ability Test-A Turing Test for Reading Comprehension. arXiv preprint arXiv: 1909.02399, 2019.

[18] Incident, https://www.pbwcz.cz/Articles%20of%20english/incident2.html

[19] J. McCarthy. 1959. Programs with common sense. In Proceedings of the Teddington Conference on the Mechanization of Thought Processes.

[20] T. Winograd. 1972. Understanding Natural Language. Academic Press.

[21] John McCarthy and Patrick J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, Machine Intelligence 4, pages 463–502. Edinburgh University Press. Reprinted in McC90.

[22] R Kowalski and M Sergot. 1986. A logic-based calculus of events. New Gen. Comput., 4(1):67–95.

[23] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In AAAI, pages 4444–4451.

[24] Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. CoRR, abs/1803.05223.

[25] Christiane Fellbaum. 1998. WordNet. Wiley Online Library.

[26] Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, page 31. Association for Computational Linguistics.

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. ICLR.

[28] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL.

[29] https://spacy.io/api/token

Dr Gongqi Lin is a lecturer in the Information Technology Discipline in the College of Engineering and Science, Victoria University. Gongqi's research focusses on the machine reading comprehension in NLP, computer vision, and eHealth. Gongqi received his PhD from Curtin University in 2014.

Professor Yuan Miao received his PhD from Tsinghua University, Automation Department, and became an associate professor within the Department of Computer Science and Technologies. His academic career further expanded to the University of Melbourne in Australia, Nanyang Technological University in Singapore and Victoria University in Australia. He is now a professor in the College of Engineering and Science at Victoria University, and the Head of the IT Discipline. Yuan's research focusses on fuzzy cognitive modelling, knowledge oriented software engineering, edutainment and eHealth.

Hu Yidan: received the B.Eng. degree from Shandong University, Jinan, China, in 2019. She is currently pursuing a Ph.D. degree in the School of Computer Science and Engineering with the University of Nanyang Technological University, Singapore. Her research interests include the recommendation system and Natural Language Processing.

Dr Zhiqi Shen is a Senior Scientist of School of Computer Science and Engineering, Nanyang Technological University, Singapore. He obtained B.Sc. in Computer Science and Technology from Peking University, M.Eng. in Computer Engieering in Beijing University of Technology, and Ph.D. in Nanyang Technological University respectively. His research interests include Artificial Intelligence, Software Agents, Multi agent Systems (MAS); Goal Oriented Modeling, Agent Oriented Software Engineering; Semantic Web/Grid, e-Learning, Bioinformatics and Bio-manufacturing; Agent Augmented Interactive Media, Game Design, and Interactive Storytelling.