Natural Language Processing Applications in Cyber Security

RAVINDER SAINI[1], SUNIL KUMAR SINGH [2]

[1]Research Scholar, Chandigarh College of Engineering and Technology, Chandigarh India (e-mail: (ravinder_cse@ccet.ac.in)

[2]Professor, Chandigarh College of Engineering and Technology, Chandigarh India (e-mail: sksingh@ccet.ac.in)

ABSTRACT Natural Language Processing, a subfield of artificial intelligence, is dedicated towards understanding and generating a natural language with the help of computers. Basically, it can make machines communicate with humans just like humans do. Natural Language Processing is a combination of computer science, linguistics and artificial intelligence. As human languages are quite complex, there are certain rules that should be known to machines in order to correctly interpret the language. Moreover each language has its own set of rules known as grammar. So it is not practically possible for a machine to rely on structured rules only to interpret the language as this may give rise to ambiguities. NLP enables machines to contextualize the text with intelligence rather than relying on rigid rules and codes. This enables NLP to learn and adapt to various dialects, new commands or questions which might go unanticipated by the programmer. Amazon Alexa, Google Voice Assistant and Apple Siri are the tools that make use of NLP to interpret and respond to human speech. Along with these applications, customer service chatbots, automated recommended replies to the emails, text auto-complete in search engines, spell checking and machine translation are some of the other applications which might be familiar to the reader. With the growing standards of industry, the usage of NLP is also growing

KEYWORDS Natural Language Processing; Cyber Security

I. INTRODUCTION

Initially being used just for human-computer interaction with the help of high performance computing and parallel architectures [1], [2], NLP these days is finding its applications in interpreting non-human languages as well. As the involvement of technology in enterprises grows [3], NLP is also matching the pace by making its way through this technology parade.

A. WORKING OF NLP

Somebody might wonder that how is it possible for a computer to interpret and understand human language. In order to make computers understand the natural language and perform tasks related to it, the natural language should be represented in a form recognized by computers. This is a process that would require a lot processing power with minimum overhead which might be achieved by the parallel and distributed architectures [4]. The computer has to be taught to recognize various perspectives of usage of a single word with respect to context dependent exchanges and tenses.

B. LEVELS OF NLP

To process the natural language and get an expected outcome, the text has to undergo various phases, which help in understanding the language and remove the ambiguities existing in the text. These levels may be understood in the following chronology:

- Phonological Analysis: This is an optional level and is applied if the origin of the text is a speech. In this phase, the speech sounds are interpreted to get an idea about the meaning of text.
- Morphological Analysis: In this phase, the words are dealt with in relation to the smallest unit of meaning. This unit may be called a morpheme. For example, “incompleteness” can be broken down into the three morphemes i.e. in (prefix), complete (stem), and ness (suffix); the prefix un- refers to “not being”, while the suffix -ness refers to “a state of being”. The stem complete is considered as a free morpheme since it is a “word” itself. The prefixes and suffixes are bound variables and would require a free morp heme to be attached to.
- Lexical Analysis: this phase deals with the identification and analysis of the structure of the words. A lexicon refers to the collection the words and phrases in a language. To analyze the data lexically, the text is divided into chunks (paragraphs, sentences, words). Stemming and Lemmatization (Lexicon normalization processes)

are generally performed in this phase. In stemming the suffixes are generally removed from word leaving behind the stem. Lemmatization is a more complex process of obtaining the root form of the word step by step in an organized way. Lemmatization makes use of grammatical rules, word structure and vocabulary to find out the root word.

- Syntactic Analysis: This phase deals with the grammatical structure of the sentences. In the sentences, combinations of words are analyzed for their validity according to the rules of grammar. Sometimes a sentence might seem ok but it is not grammatically correct. For example, "Visiting relatives can be boring". This sentence may seem ok but the combination does not convey whether the visiting is to be taken as an adjective for the word relative or visiting is a verb. The help of Part of Speech tagging is taken here to understand the applicability of the word for its grammatical relevance.
- Semantic Analysis: This phase deals with the word level meanings and the interaction among those words. Some sentences may seem grammatically alright but practically they do not convey any meaning. For example, "The colorless red sky was walked upon by the boy". The grammar of this sentence is justified but it does not convey any meaning practically as the combination of "colorless and red" is rejected in this phase.
- Discourse Integration: This phase focuses on the text as a whole by making the connections between various parts of the text. It generally takes the context into consideration and tries to convey the meaning of the sentences accordingly. In simpler words, we can say that it is the phase which integrates all the above mentioned phases to analyze the text in the most meaningful way.
- Pragmatic Analysis: It is one of the most challenging phases for AI as it goes beyond the theoretical limits to analyze the text. It takes a more practical approach than any of the phases. For example, for a sentence like "He was denied a license", license could be a driving license or an arms license or a business license. Pragmatics would help to deal with such conditions.

II. NLP APPLICATIONS

For some people, NLP might only be the way to make intelligent chat bots or virtual assistants, but in reality NLP spans across almost all the fields ranging from browser searches, translations, spell checking, sentiment analysis [5] etc. Some of the applications of NLP have been listed below:

Talking about the technical scenario, with the increase in the digital devices increases the need to secure these devices. This is where cyber security comes into play. This field can exhibit itself in various forms like securing devices, securing data in storage as well as data on the go, securing software, etc. As the field of Natural Language Processing is maturing, it is holding the hands of other fields like cyber security [6]. NLP can help automate certain security tasks or algorithms for achieving better performance. In the upcoming paragraphs, we will discuss the role of NLP in cyber security in different cases.

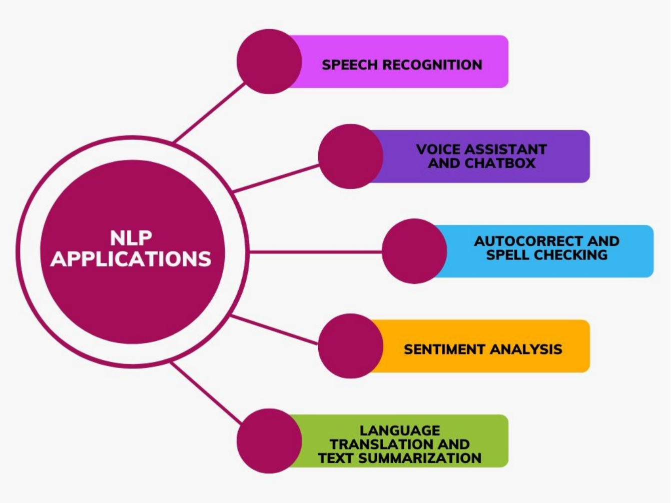
NLP APPLICATIONS
SPEECH RECOGNITION
VOICE ASSISTANT AND CHATBOX
AUTOCORRECT AND SPELL CHECKING
SENTIMENT ANALYSIS
LANGUAGE TRANSLATION AND TEXT SUMMARIZATION

FIGURE 1: Role of NLP in Cyber Security

III. HOW NLP CAN BE USED TO IDENTIFY SOFTWARE VULNERABILITIES

As we know, NLP and AI can solve many real-world problems, such as weather forecasting [2], security [7]–[10], authentication [11]–[13], healthcare [14]–[16] Till now, we have discussed NLP as a technique to understand and generate human language but little did we know that it can be used to process programming languages [17] as well. And this leads us to use NLP to make our software more secure by automating the process of finding the bugs in it. This seems hypothetical until we look at the fact that the code written by the programmers is generally repetitive because of the strict syntactic rules of the languages. Thus this reduces the perplexity of the text involved in the software code and would make it look like a prosaic text. Ultimately this helps to predict the statistical properties of seemingly complex and powerful languages.

In order to identify the bugs in the software, we can follow either of the two approaches. Going with the first technique would require cleaning up the already written code in order to enhance its readability. Once the code is clean enough, this code can be compared to the existing language model (prepared in accordance with the code conventions of the organization) and if further cleaning i.e. removal of the bugs is required, the NLP model can suggest the changes accordingly. The second approach may let the computer write the code on its own without any human intervention. Along with writing the code automatically, we can make our computer system describe the code as well on the basis of previously written codes. We can make use of various tools that use NLP for code generation. These tools can be trained using machine learning models [18]

IV. FASTER QUERIES

Generally the security experts would look at the security logs and reports manually [19]. This task can be automated with the help of NLP and AI [20]. Just as we ask the voice assistants for the general weather queries, or for nearby grocery stores; security professionals can query and govern the security management systems with voice commands. Additional tasks such as troubleshooting, or highlighting the alerts that require human intervention would also be done by NLP systems in the near future [21].

V. FASTER THREAT DETECTION

Threat detection requires the analysis of a large amount of unstructured data [22]. Although processing large amounts of data is one of the primary tasks of computers, the analysis of the data to produce some meaningful reports out of it would require the computer system to first understand what it is looking at. This is where NLP comes into play and helps in threat detection. For example, in order to detect a malicious code embedded in otherwise genuine code, the computer would break down the code and understand the relationship between various units of the code. In this process, the malicious code would be detected without the need to execute it. NLP can also be helpful in detecting Zero-Day attacks.

REFERENCES

[1] K. Aggarwal, S. K. Singh, M. Chopra, S. Kumar, and F. Colace, “Deep learning in robotics for strengthening industry 4.0.: Opportunities, challenges and future directions,” Robotics and AI for Cybersecurity and Critical Infrastructure in Smart Cities, pp. 1–19, 2022.

[2] N. Kumar and et al., “A novel framework for risk assessment and resilience of critical infrastructure towards climate change,” Technological Forecasting and Social Change, vol. 165, p. 120532, 2021.

[3] M. Chopra, S. K. Singh, S. Sharma, and D. Mahto, “Impact and usability of artificial intelligence in manufacturing workflow to empower industry 4.0,” 2020.

[4] S. Kumar and et al., “An efficient hardware supported and parallelization architecture for intelligent systems to overcome speculative overheads,” International Journal of Intelligent Systems, 2022.

[5] S. Balci, G. M. Demirci, H. Demirhan, and S. Sarp, “Sentiment analysis using state of the art machine learning techniques,” in Conference on Multimedia, Interaction, Design and Innovation. Springer, 2022, pp. 34–42.

[6] D. O. Ukwen and M. Karabatak, “Review of nlp-based systems in digital forensics and cybersecurity,” in 2021 9th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2021, pp. 1–9.

[7] A. Tewari and et al., “A lightweight mutual authentication protocol based on elliptic curve cryptography for iot devices,” International Journal of Advanced Intelligence Paradigms, vol. 9, no. 2-3, pp. 111–121, 2017.

[8] B. B. Gupta, K.-C. Li, V. C. Leung, K. E. Psannis, S. Yamaguchi et al., “Blockchain-assisted secure fine-grained searchable encryption for a cloud-based healthcare cyber-physical system,” IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 12, pp. 1877–1890, 2021.

[9] A. Mishra and et al., “Classification based machine learning for detection of ddos attack in cloud computing,” in 2021 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2021, pp. 1–4.

[10] A. Singh and et al., “Distributed denial-of-service (ddos) attacks and defense mechanisms in various web-enabled computing platforms: Issues, challenges, and future research directions,” International Journal on Semantic Web and Information Systems (IJSWIS), vol. 18, no. 1, pp. 1–43, 2022.

[11] A. Tewari and et al., “Secure timestamp-based mutual authentication protocol for iot devices using rfid tags,” International Journal on Semantic Web and Information Systems (IJSWIS), vol. 16, no. 3, pp. 20–34, 2020.

[12] A. Tewari and et al., “An analysis of provable security frameworks for rfid security,” in Handbook of computer networks and cyber security. Springer, 2020, pp. 635–651.

[13] J. Wang and et al., “Pcnncec: Efficient and privacy-preserving convolutional neural network inference based on cloud-edge-client collaboration,” IEEE Transactions on Network Science and Engineering, 2022.

[14] M. Kaur and et al., “Secure and energy efficient-based e-health care framework for green internet of things,” IEEE Transactions on Green Communications and Networking, vol. 5, no. 3, pp. 1223–1231, 2021.

[15] S. Gupta and et al., “Detection, avoidance, and attack pattern mechanisms in modern web application vulnerabilities: present and future challenges.”

[16] M. Hammad and et al., “Myocardial infarction detection based on deep neural network on imbalanced data,” Multimedia Systems, vol. 28, no. 4, pp. 1373–1385, 2022.

[17] S. K. Singh, Linux Yourself: Concept and Programming. Chapman and Hall/CRC, 2021.

[18] I. Singh, S. K. Singh, R. Singh, and S. Kumar, “Efficient loop unrolling factor prediction algorithm using machine learning models,” in 2022 3rd International Conference for Emerging Technology (INCET). IEEE, 2022, pp. 1–8.

[19] T.-M. Georgescu, “Natural language processing model for automatic analysis of cybersecurity-related documents,” Symmetry, vol. 12, no. 3, p. 354, 2020.

[20] P. Kaur, S. K. Singh, I. Singh, and S. Kumar, “Exploring convolutional neural network in computer vision-based image classification,” in International Conference on Smart Systems and Advanced Computing (Syscom-2021), 2021.

[21] Y. Deng, D. Lu, D. Huang, C.-J. Chung, and F. Lin, “Knowledge graph based learning guidance for cybersecurity hands-on labs,” in Proceedings of the ACM conference on global computing education, 2019, pp. 194–200.

[22] K. Hilton, A. Siami Namin, and K. S. Jones, “Metaphor identification in cybersecurity texts: a lightweight linguistic approach,” SN Applied Sciences, vol. 4, no. 2, pp. 1–22, 2022.