IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 9, Septmber 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

NLP and its Components: A Detailed Discussion

Prof. Zohaib Hasan[1], Prof Zeba Vishwakarma[2], Prof. Nidhi Pateriya[3]

Department of Computer Science Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, M.P, India[1,2,3]

ABSTRACT: Natural Language Processing (NLP) encompasses computational techniques for processing and analyzing human language, primarily through Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU focuses on interpreting language by analyzing phonology (sounds), morphology (word structures), syntax (sentence structures), semantics (meaning), and pragmatics (context). These processes enable machines to comprehend the nuances of human language for accurate interpretation and response generation.

NLG, in contrast, involves producing human-like text from structured data. It includes content determination (identifying relevant information), text planning (organizing information), sentence planning (constructing grammatically correct sentences), and surface realization (generating the final text). NLG is crucial for applications like automated report generation and chat bots, where coherent and contextually appropriate responses are essential.

The synergy between NLU and NLG underpins many NLP applications. In machine translation, NLU interprets the source text, while NLG generates the translated text. In question-answering systems, NLU processes the query, and NLG formulates the response. Deep learning advancements have significantly enhanced both NLU and NLG, enabling more sophisticated and human-like interactions.

This paper explores the components and processes of NLU and NLG, offering a comprehensive understanding of their mechanisms and the advancements driving modern NLP systems.

KEYWORDS: Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG), Deep Learning, Machine Translation, Question-Answering Systems

I. INTRODUCTION

Natural Language Processing (NLP) is a rapidly growing area of research that links computer science, artificial intelligence, and linguistics. The primary goal is to provide computers with the ability and motivation to comprehend, interpret, and produce human language in a meaningful and practical manner. The increasing prevalence of digital technology has made it more crucial to use natural language when dealing with machines. This has resulted in significant progress within NLP, making it a fundamental technology for numerous applications [1].

The inception of NLP can be traced back to the 1950s, when computational linguistics was first introduced. The first efforts were mostly based on rules, which meant that they relied on hand-written linguistic guidelines to process text. These systems were constrained in their scope and frequently encountered difficulties due to the inherent complexity and variability of human language. Nonetheless, the field has undergone significant changes, particularly with the advent of machine learning and deep learning. The paradigm has been transformed from rule-based to data-driven approaches by these approaches, which have made it possible to construct more advanced and precise NLP models [2].

One of the key difficulties in using NLP is the inherent ambiguity and variability that human language presents. A single concept can be conveyed in various manners depending on the context, and a variety of meanings may exist within words and sentences. The problem is addressed by the use of NLP at different levels of language analysis, such as phonology, morphology (typological and aesthetic properties of speech), syntax/language processing, semantics, and pragmatics. Phonology is concerned with the auditory characteristics of words, while morphology deals with word structure, syntax with sentence organization, semantics with meaning, and pragmatics to language in context. Understanding and constructing human language is complicated, with each level playing a part [1].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

There are numerous applications of the broadest form of NLP. The use of machine translation systems like Google Translate has facilitated the translation of text and speech between languages, significantly increasing global accessibility to cross-lingual translations. The use of email spam detection systems can safeguard users from unwanted and potentially harmful messages. Various tools can be used to extract information from large text collections, with the ability to automatically identify and classify important data points for use in decision-making processes. The use of summaries techniques enables the convolution of lengthy documents into brief, understandable sums, and medical diagnostics applications employ NLP to interpret clinical notes and research articles, contributing to improved healthcare delivery. NLP can be applied to various fields, including question answering systems like Siri and Alexa that offer users immediate responses to their inquiries [3].
There is a great potential for NLP in the future. With the advancement of models, they will become more proficient at comprehending and constructing natural language, potentially reaching levels of fluency and comprehension comparable to those observed in humans. Nevertheless, the advancements also pose difficulties in addressing ethical dilemmas like discrimination in language models and safeguarding privacy and security in NLP applications. Through the advancement of this field, researchers and practitioners strive to design systems that are more stable, impartial, and dependable, capable of being integrated into daily life, altering how we interact with technology [4].

II. LITERATURE REVIEW

Natural Language Processing (NLP) has seen significant evolution from its inception to the present day, driven by advancements in computational methods and increased understanding of linguistic structures. One of the earliest works that laid the groundwork for NLP was Turing's seminal paper on artificial intelligence, which posed the fundamental question of whether machines could think and paved the way for subsequent research in machine translation and computational linguistics [2].

The transition from rule-based systems to statistical and machine learning methods marked a significant shift in the field. Early rule-based approaches, while innovative, were often limited by their inability to handle the complexity and variability of human language effectively. Manning, Raghavan, and Schütze's work on information retrieval provided a comprehensive overview of how statistical methods could be applied to textual data, showcasing the potential for more flexible and scalable NLP systems [3]. Jurafsky and Martin's book further expanded on this by introducing various probabilistic models and algorithms used for processing natural language, highlighting the benefits of data-driven approaches [4].

One critical aspect of NLP is disambiguation, a challenge that Navigli's survey addressed comprehensively. Word sense disambiguation, the process of determining the meaning of a word based on context, is fundamental to many NLP applications. Navigli's work provided a detailed taxonomy of disambiguation methods, ranging from supervised learning approaches to knowledge-based techniques [5].

Part-of-speech tagging, another essential NLP task, was significantly advanced by Toutanova et al., who introduced a feature-rich tagging approach using cyclic dependency networks. This method demonstrated how incorporating various linguistic features could improve the accuracy of syntactic analysis, a crucial step for many downstream NLP tasks [6].
The development of transformer models, particularly BERT by Devlin et al., revolutionized NLP by enabling the pre-training of deep bidirectional transformers for language understanding. This approach allowed models to learn contextual representations of words by considering both left and right context, resulting in substantial improvements across a range of NLP benchmarks [7].

Extracting meaningful information from text, as discussed by Mihalcea and Tarau with their Textrank algorithm, highlighted the importance of unsupervised methods for text summarization and keyword extraction. This work illustrated how graph-based ranking algorithms could effectively identify the most relevant sentences or phrases in a document [8].

Collobert et al. showcased how NLP tasks could be approached almost entirely from scratch using deep learning, without relying on extensive feature engineering. Their work emphasized the power of end-to-end learning, where models directly learn representations and perform tasks using raw text data, further simplifying the pipeline and improving performance [9].

Overall, the literature reflects the rapid advancements in NLP, driven by increasingly sophisticated computational methods and a deeper understanding of linguistic principles. These works collectively highlight the trajectory from

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

rule-based systems to modern deep learning approaches, showcasing the potential for future innovations in this dynamic field [10].

III. COMPONENTS OF NLP

The components of NLP can be split into two parts. It is Natural Langage Understanding and Natural Linguistic Generation are two methods that enhance the process of comprehending and generating textual information. NLP is categorized in a broad sense, as depicted in Figure 1. The focus of this section is on the topics of Natural Language Understanding (Linguistic) (NLU) and the Natural Law Generation (NLG).

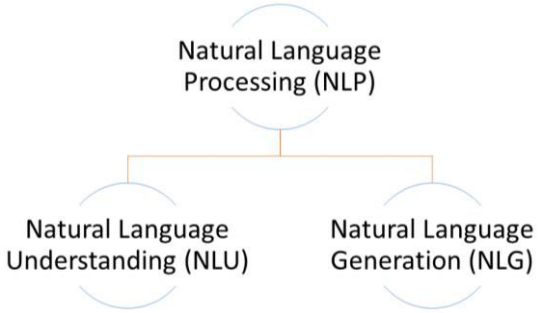

Figure 1 Broad Categorization of NLP

A. Natural Language Understanding (NLU)

The ability of machines to extract concepts, entities, emotions, keywords, and more from natural language is made possible by NLU's use in analysing it. In customer care applications, it is employed to comprehend the issues reported by customers either verbally or in writing. The science of linguistics involves investigating the meaning of language, its surroundings and the various forms that languages possess. Hence, it is essential to grasp various crucial terminologies of NLP and the different levels of neural activity. Afterward, we examine some of the most frequently employed terms in different NLP subfields

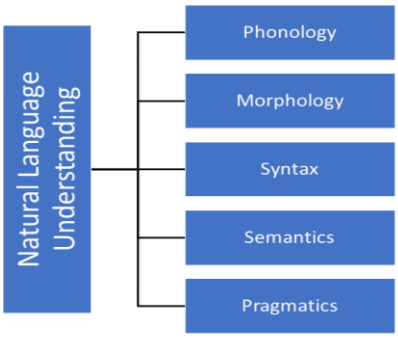

Figure 2 Components of NLU

(i) Phonology:

Linguistics encompasses the systematic structure of sound in its field, which is referred to as phonology. The term phonology originates from the Ancient Greek, where the word 'phono' which means voice or sound and has the suffix –logy to refer to speech or words. The study of sound in the language system is known as phonology, according to Nikolai Trubetzkoy in 1993. However, Lass in 1998 [13] defined philology as covering the sounds of language and investigating the sub discipline of linguistics, behaviour, and organization of sounds. In phonology, the use of sound to encode meaning in any Human language is considered semantic.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

(ii) Morphology

Morphemes are the smallest units of meaning, which are identified by their various parts in this term. The initiation of morphology, which encompasses the nature of words, is facilitated by morphemes. Morpheme is an example of how the word *precancellation* can be morphologically analyzed into three different types: first, the prefix *pre*, then and there, at right, in the root *cancella*, finally with the suffix *-tion*. The definition of morphemes remains the same across all words, as it is possible for humans to divide unidentified words into corresponding morphemes. Including the suffix *–ed* in a verb indicates that its action occurred in the past. The term used to describe words that cannot be broken down and have no meaning is referred to as Lexical morpheme (e.g. table, chair). The term table and chair are interchangeable. The words (e.g. *−ed, −ing, −est, −ly, −ful*) that are combined with the lexical morpheme are known as *Grammatical morphemes* (eg. Worked, Consulting, Smallest, Likely, Use). The combination of Grammatical morphemes, such as bound phenotypes, is known as the case. The words "-ed" and "ing" are both in English. *Bound morphemes* can be classified into two types: inflectional and derivational phenotypes. The incorporation of *Inflectional morphemes* in a word alters its various grammatical attributes, such as the tense, gender, person, mood, aspect and definiteness. When inflectional morphemes *–ed* are added, the root *park* is transformed into *parked* as an illustration. When a word is combined with *derivational morphemes*, it modifies the word's semantic meaning. By adding the bound morpheme *–ize* to the root normal, the word normalize transforms from an adjective (normal) to a verb (usually).

(iii) Lexical

In Lexical, the interpretation of individual words is a task that can be accomplished by both humans and NLP systems. The use of various processing methods results in word-to-word comprehension, with the first being a part-of-speech tag for each term. During this process, the words that are capable of serving as more than one part-of-speech are assigned the most likely part-of-speech tag, which is determined by the context in which they appear. The substitution of semantic representations with words that have a single meaning can occur at the lexical level. Depending on the semantic theory employed, the representation in NLP system can differ. The analysis of word structure at the lexical level is conducted in accordance with the PoS and phrasal meaning. This analysis breaks down text into sections, such as paragraphs, sentences, and words. Hence, words that are associated with more than one PoS tag are aligned with the most likely (and hence predetermined) PoS tag in the context. At the lexical level, assigning the correct POS tag can replace semantic representation by providing a better understanding of the intended meaning of narrative information. Its primary uses are for cleaning and feature extraction, which involve the removal of stop words, stemming, lemmatization, and other techniques. Stop words such as 'in', 'the' 'and' etc are removed as they have a high frequency and do not contribute to any meaningful interpretation, which may take up computation time. By removing the suffix of a word, stemming is employed to stem the words of the text and obtain its root form. For example: *consulting* and *consultant* words are converted to the word *consult* after stemming, using word gets converted to us and *driver* is reduced to *driv*. Lemmatization does not remove the suffix of a word; in fact, it results in the source word with the use of a vocabulary. For example, in case of token *drived*, stemming results in "driv", whereas lemmatization attempts to return the correct basic form either drive or drived depending on the context it is used.

(iv) Syntactic

The lexical level involves PoS tagging, where words are classified as phrases and clauses, followed by the grouping of phrases to form sentences, and then the combination of phrase-pronouns at the syntactic level. By scrutinizing the grammatical structure of a sentence, it emphasizes the proper construction of the sentence. A sentence that exhibits structural dependency between words is the result of this level. The process of identifying phrases with greater meaning than the actual meaning of their respective words is also known as parsing. Words' order, stop-words, morphology, and PoS are examined at the syntactic level, which is not covered by the lexical level. Changing word order can cause variations in the dependency between words and may also impact the understanding of sentences. The sentences "Ram beats Shyam in a competition" and "Shyam beat is Ram in a competition", both of which have different meanings, are examples of syntax that only differ. The sentence's meaning is altered by the removal of the stopwords, which are still present. The grammar of a sentence is altered when words are transformed to their basic form, making it unsuitable for lemmatization and stemming. The focus is on pinpointing the appropriate PoS for sentences to reflect. For example: in the sentence "frowns on his face", "frowns" is a noun whereas it is a verb in the sentence "he frowns".

(v) Semantic

At a semantic level, the most important task is to determine the proper meaning of a sentence actually means. Humans rely on the understanding of language and concepts in sentences to comprehend the meaning of a sentence, but machines cannot use these techniques. Through semantic processing, the possible meanings of a sentence are determined by considering the logical structure of the sentence and identifying the most relevant words to understand the interactions among words or different concepts in the sentences. This system comprehends that a sentence can be about "movies" even if it lacks specific words and includes related ideas such as an actor, actress, dialogue, or script.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

This level of processing involves the incorporation of semantic disambiguation in words with multiple senses. When used as a noun, the word "bark" can have different meanings, such as a sound made by dog or covering of a tree. The semantic level scrutinizes words for their interpretation in the context of the sentence, or that their dictionary interpretation is derived from the situation. To give an example, consider the statement "Krishna is deemed worthy and honourable." This sentence could be interpreted as either about Lord Krishna or about an individual named Krishna. In order to understand the correct meaning of the sentence, it is necessary to examine the other parts of that sentence.

(vi) Discourse

NLP's discourse level encompasses multiple sentences, whereas the sentence-length units are studied at the syntax and semantics level. The study of logical organization is focused on connecting words and phrases to guarantee its validity. *Anaphora Resolution* and *Coreference Resolution* are the two most common levels. By identifying the entity referred to by an anaphor and solving for its references within the text, analogy resolution is accomplished. For example, (a) Ajay topped in the class. (b) He was intelligent. Here a) and b) together form a discourse. Human beings can quickly understand that the pronoun “he” in (b) refers to “Ajay” in (a). Earlier, the word "Ram" was used to interpret "He". It is not possible to determine the relationship between these two structures and Ram's intelligence in order to top the class. By analysing all expressions that describe the same object in a text, coreference resolution can be accomplished. This is a crucial step in many NLP applications that deal with high-level NLP tasks, such as document summarization and information extraction. Co-reference is responsible for encoding anaphora.

(vii) Pragmatic

The grammatical level is the focus on the knowledge or content that originates from outside sources in the document. It relates to what the speaker means and what his listener wants. It scrutinizes the sentences that are not spoken informally. The text's content is based on practical experiences. A context analysis is used to determine the meaning of the text. The absence of specificity in a sentence and the lack of context leads to pragmatic ambiguity. When individuals have varying interpretations of the text, it can result in pragmatic ambiguity. By referencing other sentences within the context of a text, interpretations of the text and background knowledge can be determined, giving importance to the concepts expressed in that text. While semantic analysis is focused on the literal meaning of words, pragmatic analysis focuses on their practical implications. For example, the sentence “Do you know what time is it?” is interpreted to “Asking for the current time” in semantic analysis whereas in pragmatic analysis, the same sentence may refer to “expressing resentment to someone who missed the due time” in pragmatic analysis. While semantic analysis is concerned with the connection between different linguistic expressions and their meanings, pragmatic analysis pertains to the context within which we understand linguists' expression. Through pragmatic analysis, users can apply contextual knowledge to determine the intended meaning of the text.

NLP is designed to cater to one or more areas of an algorithm or system. By using the NLP assess metric on an algorithmic system, language comprehension and generation can be integrated. Multilingual event detection makes use of it. A new modular system was created for extracting events from English, Dutch, and Italian Texts, by utilizing different pipelines for each language. A set of top-notch multilingual NLP tools is incorporated into the system. Both basic NLP processing and advanced tasks like cross-lingual named entity linking, semantic role labelling. Cross-lingual frame- work enables the interpretation of events, participants, locations, and time, as well as the relationships between them. The output of individual pipelines is intended to be input for a system that produces graphs that focus on events. The output of all modules is standard and serves as the input for their subsequent pipelines. They have a data-centric architecture where the pipelines are modular and can be replaced. Modular architecture enables flexible configuration and dynamic distribution.

Natural language's major issue is that a single sentence can have multiple meanings. This is commonly tackled at the syntactic, semantic, and lexical levels. When faced with syntactic ambiguity, a sentence can be translated into various syntactical forms. When words have ambiguous meanings, semantic equivalence occurs. Lexical level ambiguity is the obscurity of one word that can be challenged by multiple claims. Each level can create ambiguity that can be resolved by knowledge of the complete sentence.

B. Natural Language Generation (NLG)

NLG is a technique for creating meaningful phrases, sentences, and paragraphs from an internal representation. Four phases are involved in Natural Language Processing, with identifying the goals, planning for achieving them, and realizing them as a text. (Figure 3) It is the opposite of Understanding.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

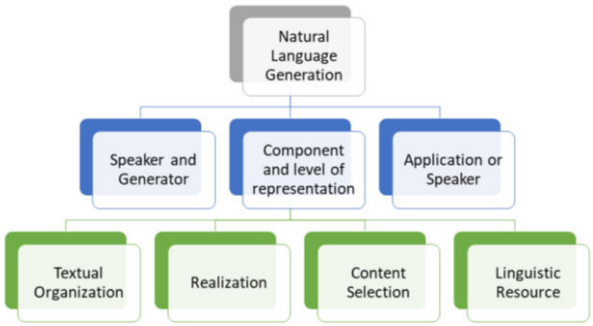

Figure 3 Components of NLG

(i) Speaker and Generator

The creation of a text requires the presence of an individual speaker or application and specialized software or generator that can translate their intentions into relevant and applicable phrases.

(ii) Components and Levels of Representation

The process of language generation involves several interconnected tasks:

- *Content Selection*: Relevant information must be chosen and included in the set. Depending on how this information is parsed into representational units, certain parts may need to be omitted, while others may be added by default.
- *Textual Organization*: The selected information must be textually organized according to grammatical rules. This includes ordering it both sequentially and in terms of linguistic relations, such as modifications.
- *Linguistic Resources*: Appropriate linguistic resources must be chosen to support the realization of the information. This involves selecting specific words, idioms, syntactic constructs, and other language elements.
- *Realization*: Finally, the selected and organized resources must be realized as actual text or voice output.

(iii) Application or Speaker

This is only to preserve the model of the situation. This is where the speaker starts off, without any involvement in language generation. It archives the past, organizes potentially pertinent content, and employs a portrayal of its understanding. All of these constitute the situation, as we select a subset of propositions that the speaker has.

IV. CONCLUSION

Ultimately, this study presents an in-depth exploration of Natural Language Processing (NLP) that highlights its two primary components: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU handles analysing the level of human speech (phonology, morphology and syntax), as well as semantics, discourse, and pragmatics. Understanding language beyond its basic levels, such as sound, word structure, context, and practical meaning, is a complex task. Machines can comprehend and react effectively to human language through this comprehensive analysis. On the other hand, NLG concentrates on extracting meaningful text from structured data. The tasks encompass content selection, textual arrangement, linguistic resource selection (i.e. the outputs produced by these processes are not only grammatically correct but also contextually relevant. By drawing on information from both NLU and NLG, the paper highlights their importance in enabling advanced technologies of NLP. With the progress of deep learning and computational linguistics, these approaches have significantly improved the natural and effective interaction capabilities of NLP systems. To push the boundaries of NLP, future research should focus on improving existing techniques and exploring new applications.

REFERENCES

1. J. Hutchins, "Machine Translation: A Brief History," in Concise History of the Language Sciences: From the Sumerians to the Cognitivists, E.F.K. Koerner and R.E. Asher, Eds. Pergamon, 1995, pp. 431-445.
2. A. Turing, "Computing Machinery and Intelligence," Mind, vol. 59, no. 236, pp. 433-460, Oct. 1950.
3. A. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.
4. A. Jurafsky and J. H. Martin, Speech and Language Processing, 2nd ed. Pearson, 2009.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 9, September 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1109034 |

5. R. Navigli, "Word Sense Disambiguation: A Survey," ACM Comput. Surv., vol. 41, no. 2, pp. 1-69, Feb. 2009.
6. Toutanova et al., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," in Proc. of NAACL-HLT, pp. 173-180, May 2003.
7. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of NAACL-HLT, pp. 4171-4186, June 2019.
8. A Vaswani et al., "Attention Is All You Need," in Proc. of NeurIPS, pp. 5998-6008, Dec. 2017.
9. R. Mihalcea and P. Tarau, "Textrank: Bringing Order into Texts," in Proc. of EMNLP, pp. 404-411, July 2004.
10. R. Collobert et al., "Natural Language Processing (almost) from Scratch," J. Mach. Learn. Res., vol. 12, pp. 2493-2537, Nov. 2011.
11. Lass R (1998) Phonology: An Introduction to Basic Concepts. Cambridge, UK; New York; Melbourne, Australia: Cambridge University Press. p. 1. ISBN 978–0–521-23728-4. Retrieved 8 January 2011Paperback ISBN 0–521–28183-0

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details