A prior case study of natural language processing on different domain

Shruthi J.[1], Suma Swamy[2]
[1]Department of Computer Science and Engineering, BMS Institute of Technology and Management, India
[2]Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, India

Article Info

Article history:

Received Jun 22, 2018
Revised Mar 14, 2020
Accepted Mar 24, 2020

Keywords:

Artificial intelligence
Human language
Natural language processing
Text mining

ABSTRACT

In the present state of digital world, computer machine do not understand the human's ordinary language. This is the great barrier between humans and digital systems. Hence, researchers found an advanced technology that provides information to the users from the digital machine. However, natural language processing (i.e. NLP) is a branch of AI that has significant implication on the ways that computer machine and humans can interact. NLP has become an essential technology in bridging the communication gap between humans and digital data. Thus, this study provides the necessity of the NLP in the current computing world along with different approaches and their applications. It also, highlights the key challenges in the development of new NLP model.

Corresponding Author:

Shruthi J.,
Department of Computer Science and Engineering,
BMS Institute of Technology and Management,
Bengaluru, India.
Email: shruthij.research@gmail.com

1. INTRODUCTION

Naturally, human language is complex and to understand this language, a system must also know about corresponding grammatical rules, meaning and context, along with slang and acronyms utilized in the language. Natural language processing (NLP) is the mechanism which supports the computer machine by simulating the human capability to understand the language. NLP is the significant area in which by analyzing data, the system can extract the information from the contexts and provides the input information in several ways. Basically, it is a relationship among human language and computer machine, i.e. NLP operates machine understanding, analysis, generation or manipulation of natural language. However, natural language refers to analysis of text as well as audible speech, whereas machine captures or recognizes the meaning of input words in terms of structured output. NLP is basic element of artificial intelligence (AI). The initial goal of NLP is to provide such type of interaction so that nonprogrammers can produce useful information from computer machine. Such type of communication was popularized in the movie "A Space odyssey" in 1968. The NLP also has the ability to make insights from information contained in mails, video files and other unstructured content [1-4].

M. Maxson, said that, in future most of the useful information will be in unstructured form. The future BigData will be the combination of both structured and unstructured data and utilizing inherent data patterns that integrate from data itself and not from police imposed on data-sets by humans. It has been frequently noted that NLP predominantly is utilized to analyze, retrieve and summarize the pertinent data from large sets of data available. An exploration of NLP concept was introduced in 1950 when Turing-test on computer machine and intelligence was introduced [5]. It was able to exhibit intelligent behavior similar to, or non-differentiable from, that of a person. NLP need a combination of verbal and computational knowledge. This can be done for several languages for e.g. English; several challenges incurred when the information extraction contain paragraph phrasing, metaphors, idioms and rhetoric etc. [6].

Language disambiguation is the art to parse and extract the information in NLP. Analyze sentence-parsing, words identification using lexicon, Named entity recognition (NER) so on and extraction are the series of operations to be performed in NLP. Different approaches [7] example: NER, wrappers in web-corpus, tagging of parts of speech has been explored. The intention of information extraction is to extract the information for pre-determined blocks for a specific frame [8]. Generic models for NLP exploiting word joint probability and its tag or label, enhancing to Hidden-Markov approach and trigram markov approach has been introduced in [9]. It also defines usage of pseudo random words and their functioning for NLP of unidentified words for small dataset. In [10] Lapata et.al have illustrated about web-based approach for NLP that provided the knowledge of isolation of compound nouns, disorderly arrangement of adjectives, web interpolation and mass counts to maximize the flexibility. Thus, NLP falls into the domain of AI with the aim of understanding and making meaningful information in the human language [11].

Different terminologies are utilized in NL processing such as 1) Morphology: -This is used to construct a meaningful word from a primitive sentence, 2) Syntax:-utilized for word arrangement which results a meaningful sentence, 3) Semantics: - concerned with meaningful word and create the meaningful sentence by combining those words and phrases, 4) Pragmatics: - utilized to know the meaning of sentence for different situations and how the understanding of sentence is effected [12]. The remaining part of the study is organized as; section-2 describes the generic process of NLP which diagrammatically represents the procedural steps of language processing. Section-3 discusses scope and applications of NLP followed by different techniques and approaches utilized in NLP in section-4. Section-1.2 highlights the existing research work towards NLP implementation followed by critical challenges in NLP in Section 5. Finally, section-6 provides the conclusion of the research study.

This section discusses some of the existing techniques of implementing NLP. Almada et al. [13], presents an identifying technical knowledge (TK) in software development. It helps in the hiring of software engineering positions and talent management. To develop this NLP technique is used and text mining (TM) is utilized to analyze formless text in resumes and curriculum. By implementing KP GENERATOR, specific data on resumes is analyzed and significantly reduce the time of work. Pelayo et al. [14] presented an enhancing hearing impaired-student reading skills in college level. The main objective of this subject is focused on only deaf students. The analysis of text, images and other information with respect to context NLP act as main role. Here, the user can imagine the information by gradually increasing the width of words and can improve understanding.

Khalid et al. [15], presented identifying the multi-word Urdu from a number of corpora. The improvement of this scheme C-value method and NLP is utilized along with Urdu WordNet. The outcomes of this method comprises of 735 synsets containing only multi-word- Urdu terms from Urdu-WordNet. Jimenez et al. [16] have proposed to predict the data or searching information using Big Data predictive analytics in social media. The result discovered and analyzed present status of the research that has been urbanized so far by educational background. Elahinia et al. [17], evaluated a prediction of complex and multidimensional health care problem using NLP technique. The main objective of this study, first analyzed the most comprehensive element of healthcare data because of many-sided nature of the problem. The output shows the possibility of patient-non adherence by person generated input-word (PGIW) and combination of machines.

Slater et al. [18], developed an effective measure of student appointment in tuition or class. The technique NLP utilized to identify the different linguistic features such as mathematics problems on students-affective states during work in a mathematic-tutor. The outcome is measured with a linguistic feature, and student problems on online tutor are evaluated. Calapodescu et al. [19] presented a semi-automatic de-identification of hospital discharge summaries. The proposed system is achieved with significant improvements for the annotation of the documents in quantitative, qualitative and homogeneous results. The work of Broniecki et al. [20] have focused on the upcoming arriving form of data and to make international practitioners grow. The presenting new data work on large question text and also worked on specific approaches. The outcomes are stable on human-derived coding and complete on correlate/predict and achieve the data program.

In the study of Zitnik et al. [21] have concentrated on the various types of applications and tools that is required for machine learning in Natural language process. A new arrived application called toolkit provide an end to end text investigation by the using of JavaScript technique based function. The state of the arts are comes in the terms of the result. The work of Baby et al. [22] have focused on hose control application in which the user can interface with web application or a newly arrived chatbot tools its control on all house electrical equipment. The presenting chatbot technique follow the statement of text information that's text by user in the form of command and then control the all-electric equipment at house by using the motion sensors. The study of Kaushik et al. [23] have focused on issues of NLP in which they discussed on deep learning which solves the problem of N-language processing. By the effective deep learning the secret issues like output-vector and input-word are removed. By the adding of Cross-Language future work is done.The study of Ahmed et al. [24] have concentrated on poetry analyzer on N- language processing which used different types of language and its consider the computer-based approach. This method considers all types of literary tested before presenting the final result is obtained.The work of pole et al. [25] have focused on the new proposed approach for the instruction of grouping documents' in which search results are in the terms of better way-group fuzzy apparatus for accurate compilation by the using of weighted method. In the study of Bartero et al. [26] described anomaly identification problem. To resolve this issue [26] introduced a Word-embedding method based on Google's Word2vec algorithm. The results illustrated 90% accuracy and efficient computation.

Suhaimin et al. [27] described a challenge which occurs in sarcasm detection into the texts. The outcomes show that a group of prosodic, syntactic and pragmatic structures developed an effective performance with F-measure result of 0.852. The study carried out by Alami et al. [28] has presented a novel Semi-automated methodology to develop a sequential structure from user necessities written in Arabic. The result shows that proposed approach provides an effective Unified Modeling Language (UML) diagrams from Arabic user requirements. Zhuang et al. [29] presented a Stroke Level approach to influence Chinese stroke for learning the Chinese word representation. The experimental outcomes show that presented method is superior for Chinese-Language processing. To reduce this type of problem the Deshmukh et al. [30] has developed Voice-based application to the users. The outcomes display the proposed system that provides correct answer to the user. It offers fast processing and maximum response capability to the users.

2. GENERIC PROCESS OF NLP

Generally, there are five steps in NLP i.e., Morphological, Lexical analysis, Syntax analysis, Programmatics and Discourse. The process of natural language needs knowledge of how utterance are created, how they turn into clauses and sentences. Additionally, to effectively understand a group of sentences in the given context, it requires a high level knowledge. Such levels are described as follows;

a. Morphological: It deals with morphological-structure of words, e.g. prefix, suffix, word root, and infixes. Basically, morpheme is the basic unit of written word which provides the information about word formation.
b. Lexical Analysis: In this stage, each word is analyzed with their component and other non-words like punctuations, acronyms and abbreviations are separated from words. This analytical process is responsible for identifying and analyzes the word structure and separating the whole text into paragraphs, set of sentences and words.
c. Syntax Analysis: The primary purpose is to find that a words string is well arranged and divide it into a structural format which represents the relation between the distinct words. It deals with word structure and validity of sentences. A syntactic analyzer performs this byexploiting a word lexicon and syntax rules (i.e., Grammar).
d. Semantic Analysis: It provides exact meaning from the given text. The text is verified for meaningfulness. The analyzer rejects the sentence like 'Illegal-Law". The sentence structure created by syntax-analyzer is assigned meaningfulness.
e. Pragmatics: In a specific context, it deals with the sentences where the sentence structure represents the actual meaning of that sentence for example: "close the window?" It has been interpreted as appeal rather than order [31].
f. Discourse: The meaning of specific sentence may depends upon sentence preceding and might be influence the significance of that sentence. For example: "He wanted it" which depends on existing discourse context.

3. ESSENTIALS OF NLP

It is well known that, NLP is a branch of AI which has several significant associations between computer machine and person interaction. In this digital world, NLP plays a significant role which bridging the communication gap among human and digital world. Furthermore, this section provides scope and applications of NLP which is utilized in the days to come.

3.1. Scope of NLP

3.1.1. Machine translation

As the global information is online, the accessibility of information is becoming highly increasing. The main challenge is to provide access for globalinformation to every person with understandable language translation. Modern companies, e.g., “Duolingo,” recruit a number of employees to contribute, by concurring translation work with educating new language. But digital translation provides more scalable alternative information to corresponding global information. Google is the innovative company that has machine translation which translates the information based on the proprietary statistical information. The challenge of machine translation technique is not translating the words but understands and stores the meaning of the sentences to offer an actual translation.

3.1.2. Automatic summarization

During accessing of specific and significant information from the extensive knowledge, an information overloading is a complex challenge to retrieve such information. Automatic summarization is not only relevant to summarize the meaning of the document, but also to understand the emotional meaning inside the information, e.g., while collecting or accessing information from the social sites. This NLP application is becoming more useful as a demandable market asset.

3.1.3. Information extraction

During the summarization of some key aspects of specific kind of stories, we use patterns extraction. For example, the statistics on terrorist attacks can be summarized by seeing place, date, an id of perpetrators, injuries, and deaths, etc., are expressed, and therefore exploring out that information and explore it into a predetermined template.

3.1.4. Question answering

At present, QA is becoming more popular for the humans, and thanks to the applications like; OK-Google, Siri, and chat-boxes. These kinds of applications are capable to answer the human requests. It can be utilized as a text interface or like a linguistic dialog system. Moreover, this application remains more challenging especially for search engines, and it is the one of the significant application of NLP research field.

3.2. NLP techniques

The NLP is used in different a technique which is described below:

3.2.1. Named-entity recognition (NER)

The NER is a type of information/data extraction that goal to recognize and determine phrases or words in text format into pre-defined classes like as the names of places, persons, expressions, etc. [32]. The named entity (NE) is an order of words that describe real-world object such as Apple Incorporation, California, etc. In data extraction, the NER is an essential task. The NER contains three universally accepted types are; Places, Person and Organization. NER doesn’t generate templates, and it is not an event recognition [33]. There are some methodologies presented which are used in NER processes are:

a. Rule-based Methodology: The Rule-based approach work as a combination of rules either automatically or manually defined. A rule contains action and a pattern. The Amazigh-named entity recognition (AER) based on symbolic method uses linguistic rules manually to generate the improvement of gazetteers by using introduced approach [34]. The NER system is used to improve the new standard Arabic with the help of the Rule-based method [35].

b. Statistical Learning method: The much Statistical Learning method is based on NER algorithms, and it treats the work as a Sequence-labeling Problem (SLP). The SLP is a common machine learning (ML) problem and has been helped to create different NLP works also containing chunking, NER, and part-of-speech tagging. The Statistical Learning approach includes three modules like Hiddden Markov models, Maximum entropy Markov models and conditional random fields.

3.2.2. Speech recognition method

The NLP is a method that can minimize the distance between machine and human being. It makes human to communicate with the device without any complexity. The Speech Recognition has the capability of a program or machine to detect phrases or words in a verbal language and change them to a machine-understandable format. It is a technology which allows human communication with devices. These are beneficial in everyday life because a machine can understand by voice. The NLP performs very efficiently to create computer-interfaces that make more comfortable to use for humans. The modern NLP method allows the use of NL to manifest programming ideas [36]. The Figure 1 shows the whole speech recognition working and is distributed in four stages. The feature extraction (FE) process is applied by using Mel frequency cepstral coefficient (MFCC) in which voice features are extracted for every voice samples. After that, all functions are specified to Pattern training which is trained by hidden markov model (HMM) to generate HMM for each word. At last the Viterbi decoding is used to choose the one with maximum probability which is nothing but recognized word [37].

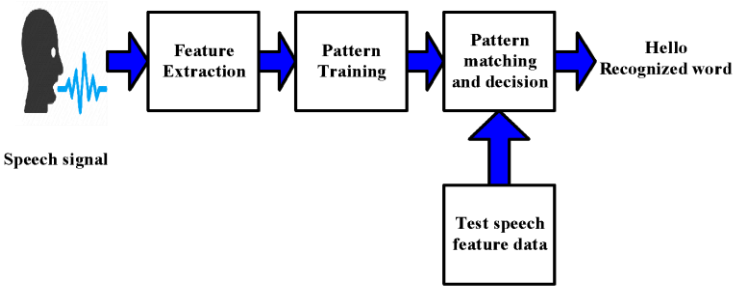

Figure 1. Speech-recognition system

The Speech-recognition system contains following techniques:

a. Dynamic Time Warping (DTW): The DTW method is a Dynamic Programming method where the whole issue is distributed into a small no. of steps and every step needing a decision to be created based on the Local-distance measures. The whole decision is depending on small decisions. The conquering and divide methodology stimulate the DTW. The DTW could find the minimum path distance with the help of matrix, and also can reduce the computation amount [38].
b. Hidden Markov Models (HMM): In HMM statistical design the system being designed uses Markov procedure with unknown parameters. The aim is to define the unknown parameters from the observable information. In HMM, the state is not directly visible, but the variables affected by the state which are visible.
c. Vector Quantization (VQ): The VQ is a procedure for mapping vectors from a big vector space to a limited number of cluster/regions in that space. The clustering/mapping the vectors can be achieved by using a clustering algorithm such as LBG (LindeBuzo and Gray) [39].
d. Ergodic Hidden Markov Models (E-HMM): The HMM is the optimal approach for demonstrating the voice signal because the HMM is signified in terms of states which is the characteristics of the voice signal and it is also shown in terms of states. The E-HMM is used for modeling the voice signal. E-HMM can directly create the order of verbal elements in a sequence [40].
e. Artificial Neural Network (ANN): An ANN is a program, which tries to copy the action of the biological activity of the human mind. The ANN has been effectual with overlapping, noisy and variable data streams. It provides effective performance in voice recognition because it delivers a faster communication as compared to "Text Writing," it has a hand free ability, beneficial for mentally or physically disabled peoples [41].

3.2.3. Optical character recognition (OCR) system

The OCR system permits to scan hand or typewritten text and change the scanned picture into a computer process format. It could be either in a word document or plain text form. The transformed document can be modified, or used in different documents. Thereby, the documents file becomes editable. The OCR is used when renewing a similar document file in a page as a document format in electronic form taking extra time. It is extensively used in transforming books and texts format into the digital form. Hence, it can be used in text mining, machine translation, character-recognition and electronically search. The three important techniques for OCR processing [42] are given below:

a. Pre-Processing: The Pre-processing stepis to generate data that are very easy and efficient for the OCR to operate and perform accurately. The Pre-processing stage is to enhance the picture quality for optimal recognition by the system. There are few techniques which implement this process are De-Skew, Binarization, Line-removal, etc.
b. Segmentation: The next OCR processing step is segmentation, where the character picture is segmented into its sub-component form. It is essential because if once extent enters into a separation of the different lines in the characters, it affects the recognition-rate.

c. Post-Processing: The Post-processing doeserror-correction/detection and grouping. It refines OCR outcomes with the help of spell or grammar check and other specific knowledge-source comparisons. The result of plain symbol recognition in text form is a group of individual symbols.

3.3. Applications of NLP

In this section, the applications of NLP are discussed.

3.3.1. Role of the NLP in the field of AI applications

An artificial intelligence (AI) is a machine which thinks like a human and it becomes dependent part of everybody’s life by providing a wide form of services. Various workis still in progress to make AI an efficient system to interact with human. One of the main techniques in AI system is NLP which plays a crucial role for the communication task. Without NLP an intelligence system cannot understand the meaning of words in context**.** So the NLP can be called as fundamental unit of the artificial intelligence which enables intelligent system to interact efficiently with humans by performing communication process such as automated speech and automated text writing by using natural language. Some of the recent techniques are discussed below:

a. NLP for speech synthesis: In this process NLP is applied for text-to-speech conversion in which text information is taken as input to the system and after conversion, machine replies into speech. The process involves various high level modules for performing speech synthesis and the sentence understanding and segmentation process is done by using decision tree.
b. NLP for speech recognition: In this NLP includes context free grammars to understand the sentences and phrase in spoken language and then extract meaningful information and transform them to digital pattern that machine can understand[43].

3.3.2. Role of the NLP in the field of big-data and classification

The text classification has continuously been an interesting topic in the research field of NLP, and while entering in the age of big data, efficient text classifiers are difficult to obtain NLP for the scientific data analytics. To provide the high computational performance to text classification in the era of big data [32] a deep neural network concept is with hybrid outliers in which text classifiers are classified as rules-based and statistical classification methods. While most of the machine learning techniques such as support vector machine, decision tree KNN, neural network, kernel learning have applied in the process of text classification [44].

3.3.3. NLP in the area of business and software management

The field of business process management faces a lot of complexities to provide quality of services due to the growth of enormous demand and the technology. Many researchers have drawn attention to redesigning process via NLP to produce better result and performances to the business organization. A domain model is designed to handle the issue of customer feedback through the support of social multimedia [39]. An NLP is implemented with geolocation context to achieve efficient interface by using fuzzy model [41].

3.3.4. NLP for ontology learning

The term “ontology” is considered within different senses in different areas. For an artificial intelligence, ontology can be defined as a specific concept such as things, events, and relations that are used to support information exchange between different entities. NLP is closely related to the Ontology’s it can be used to define essential rules for semantic analysis and question and answering system. Various approaches such as symbolic, statistical and hybrid have drawn to information extraction, and retrieval from the knowledge resource and some of these techniques have been taken into consideration for text mining via ontology learning and which is having an effective result. The symbolic method exploits human natural language (linguistic information) data to mine useful information from the text. Such as noun phrases are taken as to lexical idea to denote in an ontology.

3.2.5. NLP in the area of Bioinformatics

Bioinformatics is class of science which uses the computer technology to extract the knowledge or information from the collection of biological data. The process of information extraction involves the collection, storage, manipulation, modeling and retrieval for analyzing, predicting and visualizing through the development of software and the algorithm. Text mining [41] and NLP techniques play the very important role to recover user preference knowledge from growing databases.

3.2.6. Text mining approaches to protein research

DNA and protein sequences are the types of genetical language codes in which NLP and text mining methods are applied to analyze the bioinformatics. A latent semantic analysis was applied to protein remote homology detection, and analysis of protein spectral originates from the statistical frequency in NLP, and various web servers were designed to extract knowledge features and grammar rules of protein, DNA, and RNA sequences. In the field of predicting protein structure, some research works have used machine learning prediction methods for predication of protein disorder which based on AI and neural network, maximum entropy, support vector machine and random forest. For prediction of protein function some another work is carried out in terms of identifying region of sub-cellular protein, protein remote homology detectionobserving multifunctional enzymes, classifying protein functions and classifying trans-membrane protein by using machine learning algorithm such as SVM , kernel method, decision-making tree, random forest, Bayesian network Text processing methods which are as ontology annotation and sample weighting, are used to detect features and process training data and grammatical analyses hidden Markov model (HMM) [38, 39].

4. KEY CHALLENGES IN NLP

The key challenges in NLP are:

a. High Scale Information Acquisition: For the prediction of linguistic structure information, knowledge of idioms, lexical and general patterns recognition, mostly thousands of millions of knowledge information will be required.
b. Ambiguity Problem: Ambiguity is the most significant challenge in NLP. NLs are riddled with ambiguities at each level of description, i.e., from phonetic to sociological. Until now, human's natural languages, system is unaware of this pervasive ambiguity—it comes to researcher's attention in the guise of such misunderstandings, logistical marginal concept and contested libel suits.
c. Language Variability: Human languages are highly rich, and those have different ways to express a specific meaning like, e.g. "which is the nearest hotel?" & "can you please tell me the address of nearest hotel?" both sentences have a similar meaning. Thus, NLP approach will figure out those two sentences mean the same. Therefore, there are several ways to represent the text semantics or to distinguish two texts.
d. Semantics Problem: The fundamental challenge is that we don't have the knowledge of humans semantic structures are like; one significant method is to study that they are very nearer to human syntactic structures. In this method, a suitable study of semantics is supposed to provide high constraint the level of interpretations.
e. Knowledge Representation: How do human represent the background world information, by pointing that it requires to interact with the structure of semantic representation in the language understanding process. Have to analyze that knowledge and meaning representation is the same, it means that language is a mirror of mind or mental representation of the world.
f. Learning of language: What are the words, and how they are categorized and what is its pattern structure (i.e., idioms, syntactic, etc.) and is a very complex task, therefore the role of language learning and knowledge representation is one of the common challenges in NLP approach.

5. CONCLUSION

From this study can conclude that NLP is a significant approach which has unique features with excellent communication approach, i.e., NLP is the set of tools and methodology of knowing to achieve the goals and get results during knowledge discovery. Therefore, the case study provides the necessity of the NLP in the current computing world along with different approaches and their applications. Also, it highlights the critical challenges in the development of new NLP model.

REFERENCES

[1] H. Lane, et al., "Natural Language Processing in Action," *Manning Publications*, 2018.

[2] S. Singh, "The role of speech technology in biometrics, forensics and man-machine interface," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 281-288, 2019.

[3] I. El Bazi and N. Laachfoubi, "Arabic Named Entity Recognition Using Deep Learning Approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2025-2032, 2019.

[4] Pratheek I. and J. Paulose, "Prediction of Answer Keywords using Char-RNN," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2164-2176, 2019.

[5] "Natural Language Processing," [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing.

[6] "Syntax- English sentence structure," [Online]. Available: esl.fis.edu/learners/advice/syntax.htm.

[7] Ratinov, Lev, and Dan Roth, "Design challenges and misconceptions in named entity recognition," In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 147-155, 2009.

[8] R. J. Mooney and R. Bunescu, "Mining Knowledge from Text Using Information Extraction," *SIGKDD Explorations*, vol.7, no. 1, pp. 3-10, 2005.

[9] M. Collins, "Chapter 2: Tagging Problems, and Hidden Markov Models," Course note for NLP, Columbia University, 2013.

[10] M. Lapata and F. Keller, "Web-based models for Natural Language Processing," *ACM Transactions on speech and Language Processing*, vol. 2, no. 1, pp. 1-30, Feb. 2005.

[11] N. Ranjan, et al., "A Survey on Techniques in NLP," *International Journal of Computer Applications*, vol. 134, no. 8, pp. 6-9, Jan. 2016.

[12] A. Copestake, "Natural Language Processing," [Online]. Available: http://www.cl.cam.ac.uk/users/aac/.

[13] R. V. Almada, et al., "Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions: Generating Knowledge Profiles from Resumes," *2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT),* Mérida, Mexico, pp. 97-106, 2017.

[14] C. B. Q. Pelayo, et al., "Natural language processing for improving hearing impaired student reading skills," *2017 International Conference on Information Systems and Computer Science (INCISCOS),* Quito, pp. 201-206, 2017.

[15] K. Khalid, et al., "Extension of Semantic Based Urdu Linguistic Resources Using Natural Language Processing," *2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Orlando, FL,* pp. 1322-1325, 2017.

[16] J. L. J. Marquez, et al., "Challenges And Opportunities In Analytic-Predictive Environments Of Big Data And Natural Language Processing For Social Network Rating Systems," in *IEEE Latin America Transactions,* vol. 16, no. 2, pp. 592-597, Feb. 2018.

[17] H. Elahinia, et al., "Predicting medical nonadherence using natural language processing," *2017 IEEE MIT Undergraduate Research Technology Conference (URTC),* Cambridge, MA, pp. 1-4, 2017.

[18] S. Slater, et al., "Using natural language processing tools to develop complex models of student engagement," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII),* San Antonio, TX, pp. 542-547, 2017.

[19] I. Calapodescu, et al., "Semi-Automatic De-identification of Hospital Discharge Summaries with Natural Language Processing: A Case-Study of Performance and Real-World Usability," *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData),* Exeter, pp. 1106-1111, 2017.

[20] P. Broniecki and A. Hanchar, "Data Innovation for International Development: An overview of natural language processing for qualitative data analysis," *2017 International Conference on the Frontiers and Advances in Data Science (FADS),* Xi'an, pp. 92-97, 2017.

[21] S. Žitnik, et al., "nutIE—A modern open source natural language processing toolkit," *2017 25thTelecommunication Forum (TELFOR),* pp. 1-4, 2017.

[22] C. J. Baby, et al., "Home automation using IoT and a chatbot using natural language processing," *2017 Innovations inPower and Advanced Computing Technologies (i-PACT),* pp. 1-6, 2017.

[23] A. R. Sharma and P. Kaushik, "Literature survey of statistical, deep and reinforcement learning in natural language processing," *2017 International Conference on Computing, Communication and Automation (ICCCA),* pp. 350-354, 2017.

[24] M. A. Ahmed and S. Trausan-Matu,"Using natural language processing for analyzing Arabic poetry rhythm," *2017 16th RoEduNet Conference Networking in Education and Research (RoEduNet),* pp. 1-5, 2017.

[25] K. R. Pole and V. R. Mote, "Improvised fuzzy clustering using name entity recognition and natural language processing," *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM),* pp. 123-126, 2017.

[26] C. Bertero, et al., "Experience Report: Log Mining Using Natural Language Processing and Application to Anomaly Detection,"*2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE),* Toulouse, pp. 351-360, 2017.

[27] M. S. M. Suhaimin, et al., "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts,"*2017 8th International Conference on Information Technology (ICIT),* Amman, pp. 703-709, 2017.

[28] N. Alami, et al., "A semi-automated approach for generating sequence diagrams from Arabic user requirements using a natural language processing tool,"*2017 8th International Conference on Information Technology (ICIT),* Amman, pp. 309-314, 2017.

[29] H. Zhuang, et al., "Natural Language Processing Service Based on Stroke-Level Convolutional Networks for Chinese Text Classification,"*2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI,* pp. 404-411, 2017.

[30] S.Deshmukh, et al., "Sia: An interactive medical assistant using natural language processing,"*2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon,* pp. 584-586, 2016.

[31] T. Eftimov, et al., "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS one,* vol. 12, no. 6, pp. e0179488, 2017.

[32] J. Jiang, “Information Extraction from Text,” in Aggarwal C. and Zhai C. (eds), “Mining Text Data,” *Springer*, Boston, MA, pp. 11-41, 2012.

[33] S. Boulaknadel, et al., “Amazighe Named Entity Recognition using a A rule based approach,” *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, Doha, pp. 478-484, 2014.

[34] H. Elsayed and T. Elghazaly, “A named entities recognition system for modern standard Arabic using rule-based approach,” *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, pp. 51-54, 2015.

[35] Zaykovskiy, Dmitry, “Survey of the speech recognition techniques for mobile devices,” *Proc. of DS Publications*, 2006.

[36] R. S. Chavan and G. S. Sable, “An Overview of Speech Recognition Using HMM,” *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 6, pp. 233-238, 2013.

[37] T. B. Amin and I. Mahmood, “Speech recognition using dynamic time warping,” *2nd International Conference on Advances in Space Technologies, 2008 (ICAST 2008)*, pp. 74-79, 2008.

[38] M. Saleem, et al., “Self learning speech recognition model using vector quantization,” *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, Dublin, pp. 199-203, 2016.

[39] S. T. Pan, et al., “Speech recognition via Hidden Markov Model and neural network trained by genetic algorithm,” *2010 International Conference on Machine Learning and Cybernetics*, Qingdao, pp. 2950-2955, 2010.

[40] P. P. Patange and J. S. R. Alex, “Implementation of ANN based speech recognition system on an embedded board,”*2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, Chennai, pp. 408-412, 2017.

[41] A. Chaudhuri, et al., “Optical Character Recognition Systems for Different Languages with Soft Computing,” *Springer*, 2016.

[42] A. Reshamwala, et al., “Review on natural language processing,” *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, vol. 3, no. 1, pp. 113-116, 2013.

[43] A. Trilla, “Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition,” Departament de Tecnologies Media Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, 2009.

[44] C. Strapparava and R. Mihalcea, “Learning to identifyemotionsintext,” in*SAC’08: Proceedings of the 2008 ACM symposium on Applied computing*, New York, pp. 1556-1560, 2008.

BIOGRAPHIES OF AUTHORS

Shruthi J. completed B.E in CSE from VTUin the year2006, and completed Mtech from VTU in the year 2008. Currently pursuing Ph.D in Computer Science and Engineering, Sir MVIT under VTU.Currently working as Assistant Professor, Department of CSE, BMSITM, Bengaluru, with 10 years of teaching experience and also has 01 year of industry experience. Areas of interest are Speech Processing, Data mining, Big Data, Machine Learning and Natural Language Processing.

Dr. Suma Swamy completed her B.E in ECE from KBP College of Engineering, Satara under Shivaji University, and Kolhapur in 1990. She completed Post Graduate Diploma in Advanced Computer Technology (ASSET) from Aptech, M G Road, Bengaluru in 1996. She completed her M.Tech in ECE from Sir MVIT, Bengaluru under VTU in 2005. She completed her Ph.D in Information and Communication Engineering, Anna University Chennai in 2014. She is currently working as Professor, Department of CSE, Sir MVIT, Bengaluru. She is guiding 6 research scholars under VTU. She has 27+ years of experience in teaching with 2 years of industry experience. Her areas of interest are Speech Processing, Data mining, Big Data, Machine Learning, Natural Language Processing, Algorithms, Database Management Systems and IoT. She has 21 publications in her credit in renowned peer reviewed Journals with good impact factor and National and International Conferences. She has also authored a book “PRACTICAL APPLICATIONS OF SPEECH SIGNALS" published by LAP LAMBERT, Germany. She has 42 citations for her publications with h index of 2 with top h cited research publication titled “Speaker Independent Digit Recognition System “(18 citations) and RG score of 1.29. She is reviewer for many IEEE International Journals. She was Session Chair for many International and National Conferences. She is an Editorial Board member of Science Publishing Group, USA. She is a Life Member of CSI and ISTE professional bodies.