International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 12 | Dec -2020 ISSN: 2582-3930

NATURAL LANGUAGE PROCESSING

SANKET SUNIL BANDAWAR [19], YASH ARVIND GAIKWAD[23]

GUIDED BY MS.S.N.KAKARWAL

P.E.S COLLEGE OF ENGINEERING

COMPUTER SCIENCES AND ENGINEERING DEPARTMENT

I. INTRODUCTION

Natural language processing is a sub-branch of computer science, also related to artificial intelligence and linguistics. It deals with the inter-linkage of computers systems and natural languages spoken by humans

Natural languages are languages spoken by humans. Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of communication, natural languages are used to express our knowledge and emotions and to convey our responses to other people and to our surroundings. Natural languages are usually learned in early childhood from those around us. Currently, we are not yet at the point where these languages in all of their unprocessed forms can be understood by computers.

The field of tongue processing (NLP) is deep and diverse. Natural language processing (NLP) is a collection of techniques used to extract grammatical structure and meaning from input to perform a useful task as a result, natural language generation builds output based on the rules of the target language and the task at hand. NLP is useful in tutoring systems, duplicate detection, computer-supported instruction, and database interface fields as it provides a pathway for increased interactivity and productivity.

II.LITERATURE REVIEW

The research work in natural language processing has been increasingly addressed in recent years. The natural language processing is the electronic approach to scrutinizing text and being a very dynamic expanse of research and development. The literature extricates the main application of natural language processing and the methods to define it.

Natural language processing for Speech Synthesis:This is grounded on the text to speech transfiguration that is in which the textual information is the paramount feedback into the system. It uses high-level modules for speech synthesis. It routines the sentence dissection. Which deals with punctuation marks with a meek decision tree.

Natural language processing for Speech Recognition: Automatic speech recognition system makes use of natural language processing techniques based on grammar. It uses the context-free grammars for representing syntax of that language presents a means of dealing with spontaneous through the spotlighting addition of automatic summarization including

Indexing, which extracts the gist of the speech transcriptions to deal with Information retrieval and dialogue system issues.

III.LEVELS OF NLP

The most explanatory method for offering what actually materializes within a Natural Language Processing system is utilizing the 'levels of language' methodology. This is also referred to as the synchronic model of language and is illustrious from the earlier sequential model, which conjectures that the levels of human language processing follow one another in a sternly sequential manner. Psycholinguistic research advocates that language processing is much more dynamic, as the levels can co-operate in a multiplicity of orders. Introspection discloses that we recurrently routine information we gain from what is archetypally thought of as a sophisticated level of processing to succour in a lower level of analysis. For example, the pragmatic knowledge that the article you are reading is about biology will be used when a precise word that has numerous conceivable senses is encountered, and the word will be interpreted as having the biology sagacity. Of stipulation, the following depiction of levels will be presented chronologically. The vital point here is that meaning is conveyed by every level of language and that subsequently humans have been shown to use all levels of language to gain understanding, the more proficient an NLP system is, and the more levels of language it will exploit.

A. Phonology:

This level covenants with the elucidation of speech reverberations within and across words. There are three types of rules used in the phonological analysis:

1) Phonetic rules:

It is used for sound within words.

2) Phonemic rules:

It is used for variations of articulation when words are enunciated together.

3) Prosodic rules:

It is used to check for fluctuation in stress and intonation across a sentence.

In an NLP system that consents spoken input, the sound waves are scrutinized and encrypted into a digitized signal for elucidation by innumerable guidelines or by a judgment to the particular language model being exploited

B.Morphology:

Morphology is the principal phase of examination once the input has been acknowledged. It guises at the techniques in which words collapse into their constituents and how that distresses their grammatical prestige. Morphology is predominantly beneficial for categorizing the chunks of speech in a sentence and words that intermingle together. The following quote from Forsberg gives a little background on the field of morphology.

Morphology is a systematic description of words in a natural language. It defines a set of relations among words' surface forms and lexical forms. . A word's superficial form is its graphical or vocalized form, and the etymological form is an examination of the word into its lemma (also known as its dictionary form) and its grammatical description. This task is more precisely called inflectional morphology. Being able to identify the part of speech is essential to identifying the grammatical context a word belongs to. In English, regular verbs have a ground form with a limited set of modifications, however, irregular verbs do not follow these modification rules, and greatly increase the complexity of a language. The information gathered at the morphological stage prepares the data for the syntactical stage which looks more directly at the target language's grammatical structure.

1) Syntax:

Syntax involves applying the rules of the target language's grammar, its task is to determine the role of each word in a sentence and organize this data into a structure that is more easily manipulated for further analysis. Semantics is the examination of the meaning of words and sentences.

a) Grammar:

In English, a statement consists of a noun phrase, a verb phrase, and in some cases, a prepositional phrase. A noun phrase represents a subject that can be summarized or identified by a noun. This phrase may have articles and adjectives and/or an

implanted verb locution as well as the noun itself. A verb phrase epitomizes an action and may comprise an entrenched noun phrase along with the verb. A prepositional phrase describes a noun or a verb in the sentence. The majority of natural languages are made up of several parts of speech mainly: verbs, nouns, adjectives, adverbs, conjunctions, pronouns, and articles.

b) Parsing:

Parsing is the process of converting a sentence into a tree that represents the sentence's syntactic structure. The statement: "The green book is sitting on the desk" consists of the noun phrase: "The green book" and the verb phrase: "is sitting on the desk." The sentence tree would start at the sentence level and halt it down into the noun and verb phrase. It would then label the articles, the adjectives, and the nouns. Parsing governs whether a sentence is lawful in accordance with the language's grammar rules.

c). Semantics:

It builds up a representation of the objects and actions that a sentence is describing and includes the details provided by adjectives, adverbs, and propositions. This progression collects material vital to the pragmatic scrutiny to regulate which meaning was anticipated by the user.

d) Pragmatics:

Pragmatics is "the examination of the real meaning of an utterance in a human language, by disambiguating and contextualizing the utterance". This is done by identifying ambiguities met by the system and resolving them using one or more types of disambiguation methods.

1) Ambiguity:

Ambiguity is explained as "the problem that an utterance in a human language can have more than one possible meaning.

Types of Ambiguity:

- *Syntactic Ambiguity*: -is existent when more than one parse of a sentence occurs. "He upraised the branch with the red leaf." The verb phrase may comprise "with the red leaf" as part of the entrenched noun phrase recitation the branch or "with the red leaf" may be construed as a prepositional expression defining the act instead of the branch, inferring that he made use of the red leaf to lift the branch off. [11]
- *Semantic Ambiguity*: -is existent when more than one possible meaning exists for a sentence as in "He lifted the branch with the red leaf." It may mean that the person in question used a red leaf to lift the branch or that he lifted a branch that had a red leaf on it.
- *Referential Ambiguity:* - is the result of referring to something without explicitly naming it by using words like "it", 'he" and "they." These words require the target to be looked up and maybe unmanageable to decide such as in the sentence: "The interface sent the peripheral device information which instigated it to disruption", it could mean the peripheral device, the data, or the interface.
- *Local Ambiguity:* -occurs when a part of a sentence is unclear but is resolved when the sentence as a whole is examined. The sentence: "this hall is colder than the room," embodies local ambiguity as to the phrase: "is colder than" is indeterminate until "the room" is demarcated.

IV. METHODS AND APPROACHES

A. Natural Language Processing for Speech Synthesis: TTS synthesis makes usage of NLP procedures comprehensively meanwhile textual information is chief input into the system and thus it must be processed in the first place. [1]

Defines the diverse high-level modules convoluted in this sequential process: Text Normalization Acclimatizes the input text to be amalgamated.

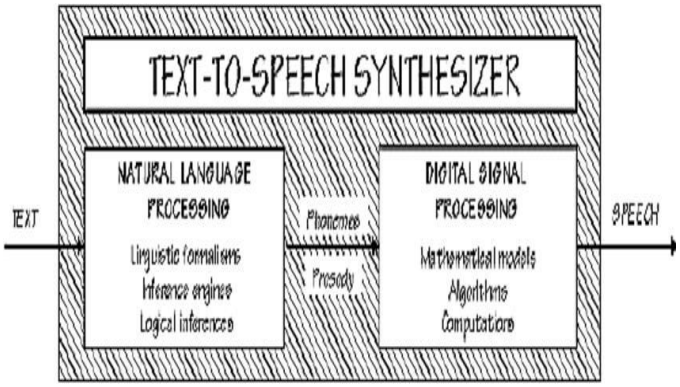

Figure1: TTS System [10]

It anticipates the characteristics that are routinely taken for granted when reading a text. The sentence segmentation can be accomplished through dealing with punctuation marks with a modest decision tree. But more confusing situations require more complex methods. Some instances of these difficulties are the period marking, the disambiguation amongst the capital letters in suitable names and the commencement of sentences, the acronyms, etc. The tokenization separates the units that build up a piece of text. It routinely ruptures the text of the sentences at white spaces and punctuation marks. This process is accomplished with a parser. Finally, nonstandard words such as certain abbreviations (Mr., Dr., etc.), date concepts, mobile numbers, abbreviations or email and URL addresses need to be expanded into more tokens (units) to be synthesized correctly. Rules and dictionaries are of use to deal with non-standard words. Part-of-Speech Tagging assigns a word-class to each token. Thus this process consecrates the Text Normalization. Part-of-Speech taggers have to deal with unknown words (Out-Of Vocabulary problem) and text with abstruse POS tags (same arrangement in the sentence) such as nouns, verbs, and adjectives. As an example, the usage of a participle as an adjective is aimed at a noun in “broken glass”.

Grapheme-to-Phoneme Transfiguration dispenses the precise phonetic set to the token stream. It must be stated that this is a continuous language-dependent process since the phonetic transcriptions of the token boundaries are influenced by the transcriptions of the neighboring token boundaries. Thus, accounting for the influence of morphology and syllable structure can improve the performance of Grapheme-to-Phoneme conversion [5].

Word Stress Allocates the stress to the arguments, a procedure firmly bound to the language of study. The phonological, morphological, and word class features are essential characteristics in this assignment: the stress is mostly determined by the syllable weight (phonological phenomena which treat some syllable sorts as substantial than others [6]). See [1] for an extensive set of orientations for this procedure.

B. Natural Language Processing for Speech Recognition: Automatic Speech Recognition systems make use of NLP procedures in a legitimately circumscribed way: they are based on grammar. This paper denotes a grammar as a set of guidelines that regulate the arrangement of texts written in a given language by defining its morphology and syntax. ASR takes for granted that the incoming speech utterances must be created according to this prearranged set of rules established by the grammar of a language, as it happens for a formal language. In that case, Context-Free Grammars (CFG) play a significant role since they are well capable of representing the syntax of that language while being efficient at the analysis (parsing) of the sentences. For this motive, such language cannot be deliberated as natural. ASR systems assume though that a large.Instruction set permits any (sternly formal) language to be taken for natural. NLP methods are of use in ASR when exhibiting the language or sphere of interaction in question.

Through the manufacture of a precise set of guidelines for the grammar, the organizations for the language are demarcated. These guidelines can both be

1) hand-crafted or 2) derived from the statistical scrutinizes executed on a branded corpus of information. The prior entails an abundant deal of hard-work since this practice is

Neither simple nor brief because it has to embody the entire set of grammatical rules for the application. The latter is generally the chosen one because of its programming flexibility at the expense of a trade-off between the complexity of the process, the accuracy of the models, and the volume of drill and test data obtainable (notice that the body has to be labeled, which implies a considerably hard workload). Since hand-crafted grammars depend solely on linguistics for a specific language and presentation they have slight importance in machine learning research in general. Thus, the literature is extensive on the data-driven approaches (N-gram statistics, word lattices, etc.)

Keeping in mind that by designation a grammar-centered exemplification of a language is a subsection of a natural language. Pointing at a flexibly sufficient grammar to simplify the most archetypal sentences for an application, [2] and [3] end up building N-gram language models. N-grams model a language through the approximations of arrangements of N successive words. While the former tackles the problem with a binary decision tree, the latter elects to use further conventional Language Modeling theory (smoothing, cutoffs, context clues, and vocabulary types) also makes usage of N-gram structures but it pursues a unified model integrating CFGs. Refer to the cited articles for further information. Lastly, [4] presents a means of dealing with spontaneous-speech through the spotlighting addition of automatic summarization together with indexing, which excerpts the essence of the language transcriptions to deal with Information Retrieval and dialogue system issues.

V. CONCLUSION

Whereas NLP is a comparatively new expanse of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that NLP-based information access technologies will carry on to be a chief area of research and development in information systems now and far into the future. The state-of-the-art Natural Language Processing techniques useful to speech technologies, specifically to Text-To-Speech synthesis and Automatic Speech Recognition. In 3TTS. The importance of NLP in processing the input text to be amalgamated is reflected. The naturalness of the speech utterances produced by the signal-processing units is firmly bound to the performance of the previous text-processing modules. In ASR NLP is harmonizing [7].

It abridges the recognition task by presuming that the input speech utterances must be created conferring to a predefined set of grammatical rules. Its capabilities can though be enriched through the usage of NLP pointing at more natural interfaces with a certain degree of knowledge. Analyses the key methodologies projected in language model edition in order to yield from this precise knowledge

VI. FUTURE WORK

NLP's future will be redefined as it faces novel technological dares and an impulse from the market to generate more user-friendly systems. The market's influence is prompting fiercer competition among existing NLP based companies. It is also assertive NLP more to Open Source Development. If the NLP community embraces Open Source Development, it will make NLP systems less proprietary and therefore less costly. The systems will also be constructed as easily disposable components, which take less time to build and more user-friendly [9].

Chatterbots – even though they occur at present, novel generations of them are being constantly developed. Chatterbots use natural language processing to mimic conversations with consumers.Web sites are commencement to mount

chatterbots as Web guidesand customer service agents.

REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "A tree based statistical language model for natural language speech recognition," in Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, Issue 7, (Yorktown Heights, NY,USA), pp. 1001–1008, July 1989.

[2] P . Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit," in Proceedings EUROSPEECH (N. F.G. Kokkinakis and E. Dermatas, eds.), vol. 1, (Rhodes, Greece), pp. 2707–2710, September 1997.

[3]] J. Tejedor, R. Garca, M. Fernndez, F. J. LpezColino, F. Perdrix, J. A. Macas, R. M. Gil, M. Oliva, D. Moya, J. Cols, , and P. Castells, "Ontology-based retrieval of human speech," in Database and Expert Systems Applications, 2007. DEXA '07. 18th International Conference on, (Regensburg, Germany), pp. 485– 489, September 2007.

[04] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," vol. 42, no. 1, pp. 93–108, 2004.

[5] Y.-Y. Wang, M. Mahajan, and X. Huang, "A unified context-free grammar and n-gram model for spoken language processing," in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. III, (Istanbul, Turkey), pp. 1639–1642, Institute of Electrical and Electronics Engineers, Inc., 2000

[6] L. Zhou and D. Zhang, "NLPIR: a theoretical framework for applying natural language processing to information retrieval," J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 2, pp. 115–123, 2003

[7] L. Zhou and D. Zhang, "NLPIR: a theoretical framework for applying natural language processing to information retrieval," J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 2, pp. 115–123, 2003.

[8] Wohleb, R. "Natural Language Processing: Understanding Its Future," PC/AI, November/December, 2001.

[9] Guerra, A. "T. Rowe Price to hone in on voice systems," Wall Street and Technology, Vol. 19, No. 3, 2000.

[10] "TTS SYSTEM" Internet https://www.researchgate.net/figure/A-simple-but-general-functional-diagram-of-a-TTS-system-2_fig1_276195975

[11]https://www.experts-exchange.com/questions/26671252/Pragmatics-Ambiguity-in-NLP.html