EasyChair Preprint

№ 10148

Natural Language Processing: Current Trend and Challenges

Kunal Saini, Gautam Juneja and Hardik Mehra

May 13, 2023

Research Article

Natural Language Processing: Current Trend and Challenges

Recently, Natural Language Processing has gained very high public attention for being able to mathematically analyze human lingual things. It comes with diverse range of ways to be used, which includes translation of machine language, fake emails detection, extraction of information and summarization, as well as in medical field. The article divides its discussion into four parts, which begins with discussing about different NLP levels and NLG components, followed by a presentation of the background and development of NLP, the phenomenon of art, a list of the numerous NLP applicabilities, and current scenarios and difficulties. NLP which is also known as "computational linguistics," employs both syntax and semantics to assist computers in comprehending human speech and writing and in making sense of what they say. Combining the power of computer programming and artificial intelligence, this field has a grasp so strong that programs can even reasonably accurately translate between languages. This field likewise incorporates voice acknowledgment, the capacity of a PC to comprehend what you say alright to properly answer.

Keywords:

Algorithm, Anthropomorphic, Artificial intelligence, Neural network, Semantics, Syntax

1. Introduction

The aim of NLP is to make computers understand human statements and words. NLP was developed to satisfy the desire of users to communicate with computers and to easify user work. NLP focuses on new skills of a person who don't have time or languages or communicational ways or become proficient in them because each user may not be that brilliant or understand the machine specific language well.

A language is what which can be classified as a bunch of rules or ways or cluster of image. In order to broadcast or convey information, symbols are combined and used. The Rules impose tyranny over signs. NLG (NL Generation) & NLU (Natural Language Unification) which completes the work of analyzing and understanding and generating text,

are the two main components of natural language processing (Figure 1).

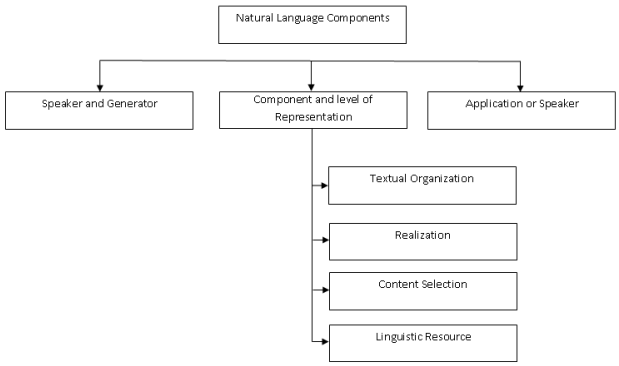

Fig 1. Broad Classification of NLP

Phonology, which is related to sound, Morphology, which is related to word formation, Syntax, which is related to sentence structure, Semantics, which is related to syntax, and Pragmatics, which is about understanding, are all parts of linguistics or languages.

Noah Chomsky, one of the principal etymologists of twelfth century that began syntactic speculations, denoted an extraordinary situation in the field of hypothetical phonetics since he upset the area of punctuation (Chomsky, 1965) [1]. Which can be roughly divided into two levels: the Higher Level including speech recognition, and the Lower Level including NL recognition.

A part of all of the above undertakings have directly linked accrediable applications, for example, interpretation by machine, Named substance addressing, Optical person addressal and so on. An understandable summarized article of a bunch of texts is generated by summarization which is kind of automatic, which is also responsible for providing information in detail or summaries of symbols of a particular type. Co-reference resolution is process of determining which words in a particular sentence or larger part of text refer to same area. Similarly the process of determining the discourse structure of connected text is referred to as discourse analysis. The term "machine translation" is defined as automated translation of text from one language spoken by humans into another.

Morphological segmentation is defined as the process of breaking down bunch of letters into its component morphemes and determining their class. NER (Named entity recognition) is a method for identifying the parts of a text stream that are related to proper names. An image of printed text is provided by optical character recognition, or OCR, which aids in identifying the corresponding or related text. It determines the part of lecture for

each word and describes bunch of words with speech tagging. Despite their obvious interdependence, NLP tasks are utilized often for easiness and convenience. Many of those tasks, like automatic summarization and co-reference analysis, serve as subtasks used to complete broader tasks.

NLP focuses on accommodation of one or more algorithmic specialties. Rospocher et al. even use it for multilingual event detection [2] purposed an original secluded framework for inter-language occasion extracted for English, Italian and Dutch texts by involving various pipelines for various dialects. A leading collection of multilingual NLP tools is included in the system as a modular set. The pipeline includes modules for lower levelled NLP processing as well as for high levelled tasks like semantic role labeling, time normalization, inter-lingual named entity linking, and more. As a result, the cross-lingual framework makes it possible to interpret the relationships between events, participants, locations, and times. Result of all these singular pipelines is planned for utilization as contribution for framework that gets data driven information charts. In UNIX, all modules behave like holes: After annotating standard input, they all produce same and standard output, which serves as input for the subsequent module pipeline. This makes it possible to modify and adapt modules. Additionally, dynamic distribution and a variety of configurations are made possible by modular architecture.

The larger portion of work in Normal Lingual Handling is proctored by researchers while other experts have shown interest, for example, analyst and savants and so on. The fact that NLP expands one's understanding of human language is one of the most ironic aspects of the practice. The field of Typical Language Taking care of is associated with different speculations and methodology that game plan with the issue of customary language of talking with the laptops. One of the most common issues in natural language is ambiguity, which typically arises at the level of syntax, which has subtasks like lexical and morphology, which deal with the study of words formation and words. The complete sentence can be used to resolve any ambiguities that may arise at any of these levels. A portion of the strategies proposed by scientists to eliminate equivocalness is saving uncertainty, for example (Shemtov 1997; 1998, Emele and Dorna; Knight & Langkilde, 2000) [3], [4], [5] Their goals are very similar to the last one:

They address a wide range of ambiguities and incorporate a statistical approach implicitly.

Literature Review

The literature review revealed that NLP has a wide range of applications in lot of domains, including health department, finance, marketing, and social media. The studies highlighted the importance of using machine learning techniques, such as neural networks and deep learning, to improve the accuracy of NLP models. The important applications of NLP in recent research include:

Sentiment Analysis:

Sentiment analysis is the method of identifying, scanning and extracting opinions and sentiments expressed in textual format. Several studies have focused on using NLP techniques for sentiment analysis in social media, customer feedback, and product reviews. The studies showed that NLP can help businesses improve customer satisfaction by analyzing customer feedback and identifying areas of improvement.

Text Summarization:

Text summarization is the method of reduction of length of a text while retaining its most important details. Recent studies have used NLP techniques to develop automated text summarization models that can be used in news articles, research papers, and legal documents.

Chatbots:

Chatbots are automated computer programs designed to simulate interaction or conversation with human users. NLP is a vital component in the development of chatbots, as it allows them to understand natural language queries and provide relevant responses. Recent studies have shown that chatbots can be used in healthcare, customer service, and e-commerce, to improve efficiency and reduce costs.

2. Levels of NLP

Best ways to represent NLP are the "levels of language," which help to generate NLP text by stating the Contextual Planning, Words Planning, and Surface Realization phases (Figure 2):

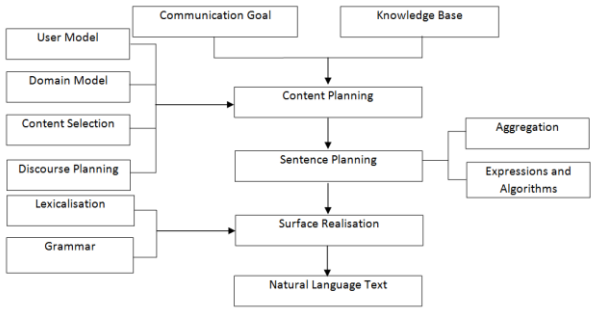

Fig. 2 Phases of NLP

The study of language meaning, language context, and various language forms is known as linguistics. These are some of the most important words used in natural language processing: -

2.1 Phonology

The discipline of linguistics called as **phonology** is concerned with systematic arrangement of voice or sound. 'Phono-'means voice or sound and '-logy' means speech or words and these are the origins of the term phonology. Phonology, according to Nikolai Trubetzkoy in 1993, is "the study of sound pertaining to the system of language." "Phonology proper is concerned with the function, behavior, and organization of sounds as linguistic items," according to Lass's 1998 article, "phonology refers broadly to the sounds of language, concerned with the lathe sub discipline of linguistics." The semantic use of sound to encode the meaning of any human language is part of phonology.

2.2 Morphology

The various parts of the word refer to Morphemes, which are the smallest units of significance. Study of nature of words is known as morphology, preceded with morphemes. Morphemes are there with similar importance throughout all words since people can transform any unknown word into morphemes to understand its significance. Adding the successor - ed to an action word, for example, shows the activity has occurred before. Words which cannot be divided further and are there with their own meanings are called lexical morphemes.

2.3 Syntactic

Here level is placed at an emphasis on examining the words in a group to discover the sentence's grammatical way. In this level, grammar and a differentiator are required. This kind of processing produces an overall representation that reveals the words' textual dependency relationships. There are a number of grammars that can be blocked, whacking the parser's option. In most of the languages, order and dependency leads to connotation, so syntax conveys meaning.

2.4 Semantic

The majority of semanticians believe that meaning is determined; however, all levels confer meaning. The communication between word-level meanings in a words group are the focus of connotation processing, which determines the possible meanings of a sentence. According to Elizabeth D. Liddy (2001), [7] the noun "file" can, among other things, refer to a binder used to organize papers, a material used to shape one's fingernails. Words are scrutinized at the semantic level not only for their elucidation in the dictionary but also for the extraction they derive from the context of the words group.

2.5 Pragmatic

Pragmatic language uses nub above and beyond the text's nub to state how additional sense is read into texts without being literally coupled in them and is concerned with the firm use of language in situations. This required extensive global knowledge, including comprehension of purpose, plans, and objectives.

3. Natural Language Generation

The method of producing senseful phrases, group of words, and paragraphs from an insider representation is known as natural language generation (NLG). It's a part of Regular Language Handling and gets executed in four stages: determining the objectives, making a plan for how those objectives might be achieved by analyzing the situation and the possible communication channels, and putting the plans into writing [Figure 3]. It's in opposition for comprehension.

Fig. 3 Components of NLG

The following are NLG's components:

Speaker and Generator: In order to create a text, we want a speaker, also known as a usage, and a generator, also known as automation, which

translates the app's goals into understandable, situation-specific language.

Components and Stages of Representation:

The following intertwined work make up the language generation method. Choosing content: The pile should consist of wanted information. Portion of the units need to be abandoned, whereas others may be added automatically, which in turn depends on the way that in which way this information should be parsed into units that are to be represnted. Structure of the Text: Semantically, the information must be arranged in accordance with grammar, both unidirectionally and in terms of lingual relations, such as alterations. Resources for Languages: It is necessary to select linguistic resources to support the understanding of the data. Ultimately, the selection of particular words, idioms, syntactic constructs, etc., will determine these resources. Realization: It is necessary to produce an actual text or voice output from the selected and arranged resources.

4. Related Work

NLP's tools and systems, developed by a large number of researchers, are what make it what it is today. NLP was a good subject for research thanks to tools like the Feeling Analyzer, Parts of Speech, Chunks, Named Entity Recognitions (NER), detection of emotions, and Labeling of Semantics.

Analyzer of emotions (Jeonghee et al. ,2003) [26] works by extracting sentiments regarding a particular subject. Two lingual resources are used in analyzing emotions: the emotional pattern database and sentiment lexicon. It tries to give ratings on a scale from -5 to +5 by looking at the documents for both positive and negative words.

Parts of speech taggers for European languages, as well as research into the development of parts of speech taggers for Arabic and Sanskrit (Namrata Tapswi, Suresh Jain), 2012) [27], Hindi (Pradipta Ranjan Beam et al., 2003)[28] etc. It is able to effectively tag and categorize words as nouns, pronouns, actions, and so on. Most procedures for part of speech work well for languages spoken in Europe, but not for Asian or Eastern languages. Arabic purposes Backing Vector Machine (SVM) (Mona Diab etal. ,2004) [29] method for

automatically tokenizing, tagging base phrases with parts of speech, and annotating Arabic text.

Chunking, also known as Shadow Parsing, is a method of sentence segment labeling that employs syntactically co-related keywords like "Noun Phrase" & "Vocabulary Phrase". Each split has its own tag, which is typically referred to as the Preceded Chunk (B-NP) tag or the Insider Chunk (I-NP) tag. The CoNLL 2000 shared task is frequently used for evaluating chunking. Chunking test results are available from CoNLL 2000. A number of systems have emerged since then (Sha and Pereira, 2003; McDonald and other, 2005; Sun and Co., 2008) [30], [31], and [32], with each reporting an F1 score of around 94.3%. Word-based features, POS tags, and tags are used in these systems.

5. Applications of NLP

Machine translation, spam detection in emails, extraction of information, making summaries, question answers, and other applications for natural language processing exist:

5.1 Machine Translation

Making data accessible to everyone is difficult given that most of the world is online. The language barrier is a major obstacle when it comes to making data accessible. There are numerous languages with distinct grammar and sentence structure. Utilizing a statistical engine like Google Translate, machine translation typically entails translating phrases from one language to another.

5.2 Text Categorization

A lot of data is entered into categorization systems, including official documents, military casualty reports, market data, newswires, and so on. and place them in predetermined indices or categories. Take, for instance, the Construe system from The Carnegie Group (Hayes PJ, Westein; 1991)[44], inputs Reuters articles and saves much time by doing this work that staff or human indexers would normally do. Trouble tickets and complaint requests can now be routed to the appropriate desks using categorization systems from some businesses.

5.3 Information Extraction

The process of locating key phrases in textual data is focus of extracting information. Extracting entities like places, names, events, dates, times, and rates is a powerful way to summarize relevant information for many applications. On account of a space explicit web search tool, the programmed distinguishing proof of significant data can build exactness and productivity of a coordinated hunt. Hidden Markov models (HMMs) are used to extract relevant research paper fields. These sections of extracted text are used to match references to papers, present search results in an efficient manner, and conduct searches in specific fields.

6. Approaches

The rationalist or symbolic approach is based on the idea that a significant portion of the human mind's knowledge is established in advance, probably through genetics, and is not derived from senses. This method was strongly supported by Noam Chomsky. It was believed that lingual knowledge is directly involved in rules or other representational forms, so that machines could function similarly to human brains by imparting some fundamental knowledge and reasoning mechanism. Natural language processing is made easier by this. [98] The algorithmic development that enable a logic to detect patterns is central to statistical and machine learning. The underlying algorithm of a given algorithm is characterized by an iterative procedure that is amended by a numerical measure that identifies parameters (numerical) and the learning phase. The majority of ML models fall into either the generative or discriminative categories. Because they are able to generate synthetic data, generative methods produce rich probability distribution models. Based on observations, discriminative methods are more effective and have right estimating posterior probabilities.

6.1 Hidden Markov Model

An HMM is a system in which, with each switch, possible output symbols are generated by shifting between multiple states. Despite their size, the sets of viable states and unique symbols are limited and well-known. The system's internals are hidden, but we can complain about the outputs. A specific sequence of output symbols can be used to calculate the probabilities of one or more candidate states, solving a small number of the

problems. It is likely that a particular output-symbol sequence was generated by pattern matching the state-switch sequence. Calculate the state-switch/output probabilities that best fit the data after training the output-symbol chain data.

6.2 Naïve Base Classifiers

The decision of region is colossal covering normal things like word division and interpretation yet additionally strange regions like division for newborn child learning and recognizing records for feelings and realities. Additionally, the exclusive article was chosen for its use of Bayesian techniques in the design of algorithms for the study.

Conclusion

The process of instructing machines to comprehend and interpret human conversational input is known as natural language processing. It is possible to establish channels of communication between machines and humans using NLP that is based on machine learning. NLP has already proven useful in numerous fields, despite its ongoing development. NLP's various applications can help individuals and businesses save time, increase productivity, and improve customer satisfaction.

References

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

2. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.

3. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing

(EMNLP) (Vol. 1631, pp. 1642-1654).

4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

5. Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751).

6. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Association for Computational Linguistics (ACL) System Demonstrations (pp. 55-60).

7. Ruder, S., Peters, M. E., Swayamdipta, S., Wolf, T., & Vulic, I. (2019). Transfer learning in natural language processing. arXiv preprint arXiv:1901.08163.

8. Goldberg, Y. (2015). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.

9. Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NIPS) (pp. 5998-6008).