World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Natural Language Processing (NLP) in Artificial Intelligence

Sateesh Kumar Rongali *

Judson University.

World Journal of Advanced Research and Reviews, 2025, 25(01), 2515-2519

Publication history: Received on 16 December 2024; revised on 24 January 2025; accepted on 28 January 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.1.0277

Abstract

Natural language processing (NLP) which is a branch of Artificial Intelligence (AI) has experienced significant improvement in the recent past to allow machines to language comprehend and generate. They consist of uses like machine translation, sentiment analysis, chatbots, and virtual assistants, which form a cornerstone part of life. However, even with these advances, NLP still has numerous critical difficulties that affect its proficiency and usefulness in applying the systems. Some of the major problems include one language may mean different things to different people; every situation requires different approaches; and finally people from different cultures and languages will pose a significant problem. This paper presents evidence of lexical, syntactic, and semantic ambiguities that complicate the language understanding process. Besides that, NLP models are not able to comprehend the flow of human's dialogues which is important factor of the communication. The problem of language diversity in human dialogue makes it even more challenging to develop NLP since over 7,000 languages are characterized by unique structures and expressions. With the recent development in Machine learning, and deep learning, these challenges have been well addressed. Pretrained transformer models like BERT and GPT have greatly enriched the field's tech arsenal, since language comprehension and Boolean modernity loops very difficult to model and tackle. This journal provides a comprehensive look at these issues presents current technologies and examines new trends pertaining to NLP. Lastly, it brings focus on way ethical concerns like bias are paramount in making current NLP systems neutral. Looking forward more advances are expected in NLP which has the prospective of further improvement of interaction between human and computer.

Keywords: Natural Language Processing; Artificial Intelligence; NLP Challenges; Deep Learning; Machine Learning

1. Introduction

Natural Language Processing (NLP) is an important subfield of AI and its main concern is understanding, interpreting and generation of natural language by machines. NLP aims at embodying the gap between the manner through which human communicate and the manner in which a computer can understand textual and spoken content in a meaningful way. Usages of NLP are manifold, including personal assistants, moderation of content on social media, sales analysis of customer sentiment, and language translation services.

Thanks to the progress of the practice of machine learning, especially the idea of learning technology in recent years, which is deep learning, NLP system's performance has been greatly improved in the past decade. The previous models and methods of the NLP were based on the sets of rules and features that demanded more input from a human being to set up those rules and correlate language information. But with the availability of big data & more advanced versions of the machine learning capable algorithms, NLP has evolved to incorporate such models which are capable of learning patterns in a Language from large corpus data on its own.

The transformer models have been revolutionary starting from BERT and GPT. Nevertheless, there are certain problems, which NLP must address: the problem of disambiguation, the problem of coherence, and the problem of

* Corresponding author: Sateesh Rongali

crossing cultural and linguistic barriers. This journal will discuss such challenges in detail and, more specifically, how recent developments have attended to some of these matters; how the research in this field will look toward the future to surmount the remaining difficulties.

2. Key Challenges in Natural Language Processing

2.1. Ambiguity in Language

Indeed, complexity is one of the biggest problems in NLP since human language is, by nature, ambiguous. There are different types of ambiguity, lexical, syntactic and semantic, which makes them challenging for NLP systems. It is therefore imperative that NLP researchers learn how to manage ambiguity in order to create models that generate and comprehensively interpret language.

There is what is known as lexical variation, whereby a given word is used to have different meanings in the different situations. For instance, "bank" a noun, can mean financial institution, the side of a river or a place where something is kept "data bank" (Ruder *et al,* 2019). Often it requires the context to be defined or the words surrounding it for the meaning to pop out. However, if there is no context a NLP system is unable to know the right meaning to assign.

Syntactic ambiguity is a situation whereby the same sentence can be read in more than one way depending on the structure of the language used in the construction of the two interpretations in question. A vivid example is given in the form of a simple English sentence "I saw the man with the telescope." To reduce the number of syntactic ambiguities, the person must use several syntactic structures and choose the one which most adequately corresponds to the context.

Semantic ambiguity is used when even after analyzing syntactic meaning of the word or a sentence there are two meanings of the same concept. This often follows from the situation when one and the same word can produce different meanings due to figure of speech idiomatic language, cultural differences (Zhang & Wallace, 2017). For instance, the word "kick the bucket" is used to convey the meaning to die in one situation while in a different situation it is a literal meaning of the phrase.

For NLP systems to work effectively, the resolved and the relatively more complex multiple meanings together with context from rest of the text, users' profile, and general knowledge of world have to be captured. While the models such as deep learning, attention mechanism have made NLP solutions handle these to some degrees, there are still language specific challenges especially when dealing with languages that have complex syntax.

2.2. Context and Pragmatics

Culture and semantics are two major aspects of human communication that are where most of the difficulties in Natural Language Processing (NLP) originate. Language is particularly contextual and contextualized Most often, the exact signification of words, phrases or sentences depends on the context, given information and learning situations. Context modeling is another factor where standard NLP systems need to be proficient in order to best serve a user's needs.

Culturally sensitive context involves linguistic background, interactional features or practices and communicative setting within which language is employed. For instance when speaking of such a primitive type of a sentence as "Can you pass the salt?" differences when it is said in a dinner with friends compared to when it is said before an audience. In the dinner scenario, it is possibly a call, while in the other one it will be quite a question. They have to know such context when it comes to identifying intent and providing the right outputs in NLP models.

Pragmatics is the branch of linguistics that deals with language in use, including how speakers rely on knowledge, presumption, and culture in particular social encounters. For instance, irony, sarcasm and humor entail presuppositions that goes beyond the surface of the words used. Of course, any such pragmatic issues are challenging for the NLP systems since solving them comes understood of references to the speaker's intentions, their emotional state, which might be expressed with body language etc. Progress in dealing with context in deep learning, especially the transformers like BERT, GPT has been made but capturing human pragmatics using a model is still a challenge in current NLP.

2.3. Cultural and Linguistic Diversity

The last major challenge of Natural Language Processing (NLP) is cultural and linguistic diversity. Nonetheless, many NLP systems have made considerable advances; unfortunately, most of them have been designed mainly in English since it is the most examined and implemented language to date. Therefore, they tend to do poorly if they are implemented into other languages with different grammatical structures, syntactical structures and even cultural values. Two

observations could be made: First, NLP systems enable understanding and generation and second, these capabilities are highly dependent on linguistically and culturally contextualized data.

In sum, there are more than 7,000 documented languages that signify phonetic, syntactic and semantic patterns different from one another. English also has elements unlike some of the languages which are; complex verb conjugations, agglutinative word formation, and also the tones which in turn prove complex to most of the NLP models. For example, the languages like Mandarin Chinese or Arabic or Hindi, which are totally different from English and even also from Spanish and French etc., need models understanding not only grammar but also culture of the countries these languages belong to. For example, in Mandarin English word order, can be much more free compared to the English language or in Arabic writing from right to left. They entail that it is possible to have significant errors whenever one wants to apply any model derived from data originating from English languages on these languages.

The terms and expressions that are familiar in one culture may not certainly exist in another; similarly, tastes if humor, sarcasm, and politeness are not universal. NLP models face a problem of transferring knowledge from one cultural set to another since this would not be easily possible without a hitch. To address these issues, researchers are working on multilingual models, such as mBERT and XLM-R that aim to bridge linguistic gaps. These models are designed to understand and process multiple languages simultaneously, often leveraging shared semantic spaces across languages. However, achieving high performance across the vast array of linguistic and cultural diversity remains a significant challenge, requiring extensive data, resources, and innovative approaches to model training and evaluation.

2.4. Scalability of Models

A main issue related to Natural language Processing (NLP) models is scalability. The traditional NLP system or models get complex as the concept is evolving and this necessitates high computational power, larger corpus, and long training periods. This is especially true in advanced models such as transformers that have boosted NLP by outcompeting rivals in most tasks. However, the use of these models leads to billions of parameters making training and deployment a very costly process.

Processing such models like GPT-3 or BERT implies the application of highly effective hardware such as distributed computing structures and graphic or tensor processing unit. In the case of environment and financial costs related to all these resources, they are usually very expensive for many organizations and researchers.

Furthermore, the requirement of massive and diversified data sets to train these models increases the problem too. Making sure that these models can also generalize when the number of instances to process is large without heavy computational resources is still an open issue in the field. (Conneau & Lample, 2019). It is noted that utilizing approaches as model pruning, knowledge distillation, and other efficient architectures helps the researchers to increase the scalability and decrease the usage of resources.

3. Advancements and Solutions to NLP Challenges

3.1. Transformer Models

A major compounding factor in recent years has been the creation of transformer-based models. Recent advances in the Transformers like OpenAI's GPT, and Google's BERT, have caused a dramatic shift in the performance of NLP machines.

The transformer model eliminates many of the issues encountered with prior approaches, such as RNN and LSTMs. The use of Transformers makes easier parallel computation compared to RNNs and hence faster in training and are very efficient in a broad spectrum of NLP tasks. This self-attention ability enables the transformer to capture long-range dependency in the text, which is crucial when resolving the ambiguous text or when trying to understand the context.

3.2. Transfer Learning and Pre-trained Models

Transfer learning is now a basic technique for contemporary Natural Language Processing (NLP), enabling models to use knowledge from one task or domain and use it in another. Such technique is especially beneficial in NLP because of limitations implied by the data and time-saving when building models from scratch. Transfer learning usually consists in training a model on a large corpus on text data and then adapting it to a particular task, for instance sentiment analysis, named entity recognition, machine translation and so on (Radford et al, 2018)

Now, with models like BERT, GPT, and T5, NLP has been transitioned into having general purpose language understandings which can be fine-tuned on a variety of tasks with relatively little need for task-specific data. In turn, they have a great understanding of the language, which can be adapted to reach even higher results on certain tasks.

Pre-trained models make a tremendous contribution towards the time and effort that is required to build NLP applications. Introducing fine-tuning approach to AI enthusiasts means that rather than building models anew, developers can use smaller datasets with specificity to the task required to fine-tune a pre-existing model, making it possible to develop statistically excellent NLP systems for far less data and computational power. Such an approach has crested large enhancements into account NLP application in terms of accessibility and scalability across industries and Languages.

3.3. Data Augmentation and Synthetic Data

Data augmentation and availability of synthetic data have become the important approaches when it comes to enhancing the performance of Natural Language Processing (NLP) models in particular when dealing with limited or imbalanced datasets. Compared to the general strategies for other NLP tasks, obtaining huge corpus, especially for certain topics or languages, can be expensive or take a lot of time. The problems stated above are solved with data augmentation and the generation of synthetic data as they artificially expand the training dataset.

Augmentation is a technique in which various procedures are applied to existing data to generate new variants of the identical input and does not require the labelling of new data. Various strategies used in NLP for translated text may employ paraphrasing, back translation, the random re-arrangement of a few of the words and text summarization. For instance, the fourth paraphrase of an example “It is beautiful in this weather” is “Today is a lovely weather”. Such transformations assist models in better generalizing by training them on the different ways by which the same semantics can be represented.

This experimental form of text data is specific in that it derives from natural language data but is not original and was created artificially. Two of such methods include GANs or the likes of text generation models like the GPT can be employed to generate high quality synthetic data for training (Sun et al, 2019). These can generate different sentences, phrases or even full documents which can be used in addition to real data, or in cases where real data for models of low resource languages are hard to come from.

With the help of data augmented and synthetic data, training data become larger and more diverse for NLP systems there by improving the systems performance for a variety of tasks and domains. However, these techniques are most useful in relieving situations of short supply of data, particularly for languages that are not well catered for or for specific applications.

3.4. Multilingual Models

Multilingual models have become a major breakthrough to combat the issues of language diversities in Natural Language Processing (NLP) systems. These models are developed to handle multiple languages at the same time and work several languages through a single model called multilingual models. Conventional systems of NLP involve language-specific models, they are expensive and also have low generalizations (Devlin et al, 2019). However, as it will be discussed in detail in the following sections, multilingual models such as mBERT and XLM-R allow for cross-lingual transfer that is for using a model learned on data from one language directly in other languages with slight adjustments.

These models employ a multimodal architecture where a single multilingual embedding space is used when translating to more generally applicable embedding’s. For instance, any methods introduced to mBERT have been tested in dozens of languages and even in low resource language shows high performance, thus making these techniques more available for various uses. And for tasks like machine translation or cross-lingual sentiment analysis, named entity recognition, or multilingual question answering, which are at the heart of the globalized digital world today, multilingual models are the only way to go. Despite the improvements brought by these models, there are still limiting factors that constrain developers to achieve high performance in all languages; not to mention the under-represented languages and those far from large languages.

3.5. Ethical Considerations and Bias Mitigation

Researchers assured that with the further spread of NLP models, the issues of both ethical practices and bias will remain significant. NLP systems can learn from a huge amount of data that can be representative of any society, including developmental prejudices. These gender-associating, race-associating or class-associating biases can be unconsciously

'learnt and 'taught' to the NLP models and reflected in the outputs (Zhao & Mao, 2020). For example, language model may make bias in self-employment recommendation or in moderating the content of the text.

Solving such problems is only possible with bias detection and mitigation throughout the development and deployment of these models. Additionally, nowadays we need diverse and representative datasets which would help to avoid discrimination in terms of performance between different groups of people. Scientists are also trying to design best practices for NLP practice and its applications that can also prevent the misuse of this technology

4. Conclusion

Several years, Natural Language Processing remains a promising, growing area with substantial contributions to Artificial Intelligence. Nevertheless, NLP still has many fundamental issues despite the recent improvement in machine learning and deep learning approaches such as: Ambiguity, Context Sensitive and Multilingualism. Nevertheless, it has been improving many problems like extractive, translation, transfer, multilingual, and augmentation with new techniques like transformer, transfer learning, multilingual, and augmentation.

As the use of NLP systems continues to increase in ability and application, there are clear potential conflicts with those principles of ethics. Reducing bias, enhancing culture understanding and aperture, and embracing fairness will be vital to the sound execution of NLP ventures. The potential of NLP in particular can be looked at in terms of creating more natural interface between the human and the computer up to the possibility of increasing the level of intercultural communication.

Over the years, many researchers will scale up and develop their NLP approaches taking perception and language creations to another level using AI. We have seen this is an interesting area for development and NLP is set to revolutionize industries like healthcare, finance, and education.

References

[1] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. NeurIPS 2019.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019.

[3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog.

[4] Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2019). A survey of cross-lingual embedding models. Journal of Artificial Intelligence Research, 65, 1-43.

[5] Sun, T., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).

[6] Zhang, Y., & Wallace, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Proceedings of the 8th International Conference on Learning Representations (ICLR 2017).

[7] Zhao, S., & Mao, X. (2020). Learning in adversarial settings for natural language processing. IEEE Transactions on Neural Networks and Learning Systems.