European Journal of
Technology
(EJT)

A Comprehensive Survey of Deep Learning Techniques Natural Language Processing

Jasmin Praful Bharadiya

AJPO

A Comprehensive Survey of Deep Learning Techniques Natural Language Processing

Jasmin Praful Bharadiya[1*]
[1]Doctor of Philosophy Information Technology, University of the Cumberlands, USA
*Corresponding Author's Email: jasminbharadiya92@gmail.com

Crossref
Article history
Submitted 19.05.2023 Revised Version Received 23.05.2023 Accepted 23.05.2023

Abstract

In NLP research, unsupervised or semi-supervised learning techniques are increasingly getting more attention. These learning techniques are capable of learning from data that has not been manually annotated with the necessary answers or by combining non-annotated and annotated data. This essay presents a survey of various natural language processing methods. The discipline of natural language processing, which integrates linguistics, artificial intelligence, and computer science, was established to make it easier for computers and human language to communicate with one another. It is, as we can say, relevant psychopathology for the study of computer-human interaction. The understanding of natural language, which entails enabling machines to naturally interpret human language, is one of the many challenges this area faces. Discourse analysis, morphological separation, machine translation, production and understanding of NLP, part-of-speech tagging, recognition of optical characters, speech recognition, and sentiment analysis are some of the most frequent NLP tasks. As opposed to learning, which is supervised and typically yields few correct results for a given amount of input data, this job is typically quite difficult. However, there is a sizable amount of data available that is unannotated in nature, i.e. the entire contents are available on the internet, and it typically yields less accurate findings.

Keywords: *NLP (Processing of Natural Language), Natural Language Comprehension. Analysis of Text, Knowledge Acquisition, Class-Based Language Modeling*

1.0 INTRODUCTION

The goal of processing of natural language an area of linguistics, artificial intelligence, and computer science is to enable communication between computers and human language. It is associated with the field of computer-human interaction, to put it simply. (B. Manaris,) This topic has a variety of difficulties, including as the understanding of natural language, which involves enabling robots to comprehend human language naturally.

The most common tasks for NLP include discourse analysis, morphological separation, machine translation, generation and understanding of NLP, recognition of named entities, part-of-speech tagging, recognition of optical characters, recognition of speech, and sentiment analysis. Unsupervised or semi-supervised learning techniques are currently receiving greater attention in NLP research. (A. Lopez,) These learning approaches are capable of learning from data that has not been explicitly annotated with the necessary answers or by combining non-annotated and annotated data. As opposed to learning, which is supervised and often yields few correct outputs for a given quantity of input data, this job is typically quite difficult. However, there is a sizable amount of data available that is unannotated in nature, i.e. the entire contents are available on the internet, and it often yields less accurate findings.

Prior to the 1980s, many natural language processing systems relied on intricate rules that were handwritten. However, a revolution occurred in the field of natural language processing in the 1980s with the creation of learning methods based on machine learning approaches for processing natural language. (F. Jarray,) The study has then demonstrated interest in statistical methods that may apply probabilistic judgements based on the weights associated with various aspects for constructing input data. These methods are typically very reliable when we provide input that is unfamiliar, especially when the input contains errors (which are very common in case of data from the real world), and they produce very reliable results when integrating it with lengthy systems with numerous subtasks.

AI techniques have found applications in the health field due to their practical solutions, implicit feature engineering capabilities, word embedding integration capability (Ru et al., 2019, Ru et al., 2018), and ability to manage intricate and unstructured data. The availability of previously unheard-of amounts of health-related data, such as digital text in electronic health records (EHRs), clinical text on social media, text in electronic medical reports, and medical images, has also contributed significantly to the rise in popularity of DL in the healthcare industry. (F. Jarray,)The volume of publications reported between 2017 and 2020 illustrates the health-related attractiveness of DL.

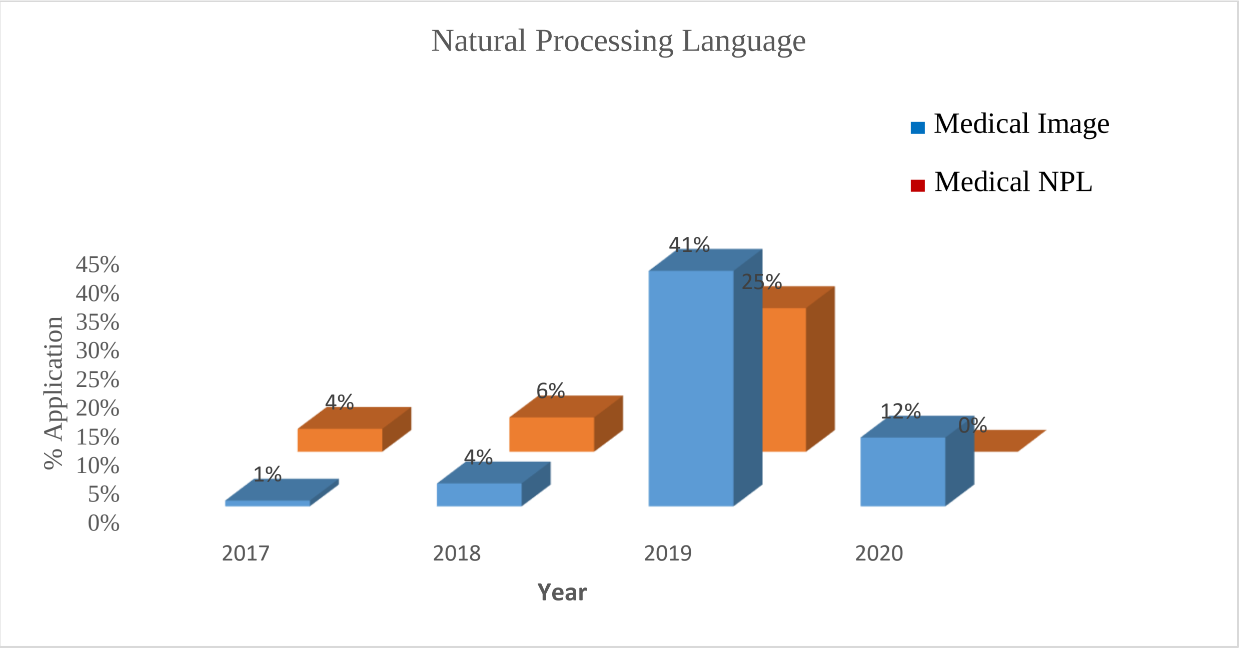

Figure 1: Natural Processing Language

A large portion of the previous success in machine translation was made possible by high-caliber research conducted at IBM, where extremely complex statistical models were frequently developed and put into practice. These methods benefited from earlier multilingual text corpora that were created by the European Union of European Parliament and the Canadian Parliament as a result of laws mandating the translation of all governmental proceedings into the various official languages used in various governmental systems. However, many other strategies also depended on this corpus, which was a major issue for their effectiveness. As a result, learning well from little amounts of data has been the subject of much research. This essay provides an overview of numerous NLP approaches.

Statement of the Problem

The statement of the problem refers to a concise description or formulation of the issue or challenge that needs to be addressed. It provides an overview of the problem, highlighting its key aspects and objectives. In the context of natural language processing (NLP), the problem statement would typically focus on a specific NLP task or goal that requires a solution. For example, the problem statement could revolve around improving sentiment analysis accuracy, developing a machine translation system, or enhancing question-answering capabilities. The statement of the problem sets the foundation for designing and implementing solutions or approaches to tackle the identified NLP challenge effectively.

Big Data for Natural Language Processing

Around 80% of the data accessible today is in its raw form. Big Data is created from information that is kept in both large organisations and businesses. Examples include data on personnel, firm purchases and sales history, business transactions, an organization's past performance, and data from social media. With the use of NLP, this vast amount of unstructured data may be exploited for growing patterns inside data to better understand the information contained in data, even if human language is unclear and unstructured for computers to read. By utilising Big Data, it may resolve key issues in the commercial sector. Be it a retail establishment, a medical facility, a workplace, or a financial institution.

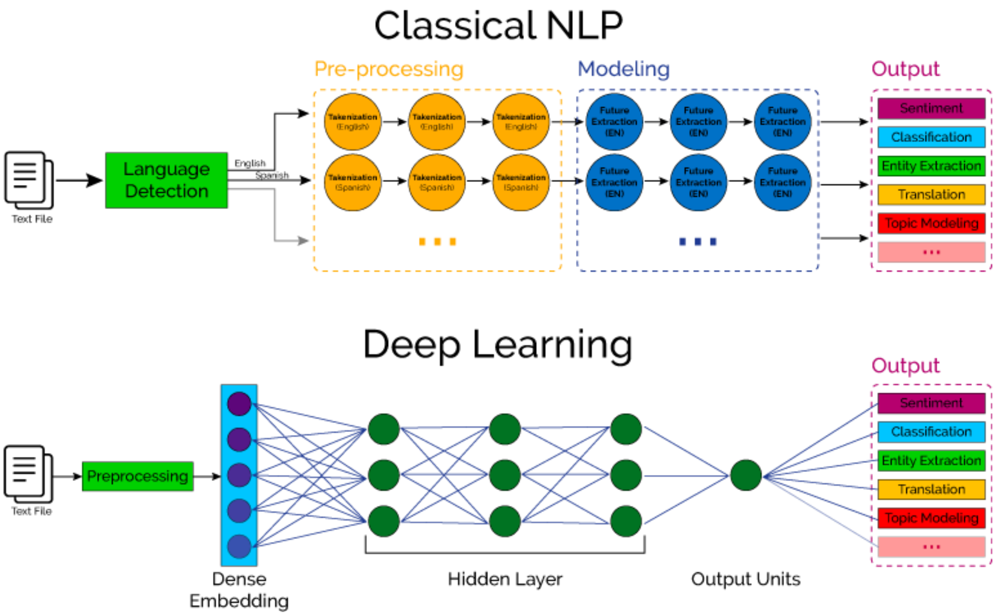

Figure 2: Classical NLP; Deep Learning

As per the pragmatic understanding the nature language of processing is to deal with different sub-modules are:

Improved Training Data

The use of big data enables researchers to collect and analyse enormous volumes of textual information from a variety of sources, such as social media, online articles, and books. The generalisation abilities of NLP models are improved by the availability of varied and substantial training data, which increases their efficiency in comprehending and producing natural language.

Multilingual and Cross-Lingual NLP

Big data offers the chance to make use of enormous volumes of multilingual text data for creating models that can handle various languages. Future research in multilingual and cross-lingual NLP will concentrate on methods to efficiently use big data resources to enhance cross-lingual information retrieval, machine translation, and other language-related activities.

Domain-Specific NLP:

Big data may be used to create NLP models that are specialised to a certain domain. Researchers may create models that are especially designed to comprehend and produce text in specialised areas like medical, law, finance, or scientific research by gathering and analysing enormous volumes of domain-specific text data. For applications that are domain-specific, these models can provide answers that are more precise and pertinent.

Ethical and Responsible NLP

Research on the moral and responsible usage of NLP models is becoming more and more important as they become more potent and are able to analyse enormous volumes of data. Future studies are anticipated to look into methods for dealing with biases in training data, making sure that NLP applications are fair and inclusive, and incorporating concepts of accountability and transparency into the creation and use of NLP models.

These are hypothetical study fields based on the trends and difficulties in NLP at the moment. New directions and study horizons may open up as the area develops, impacted by improvements in big data accessibility and processing power.

Class-Based Language Modeling

By utilising class-level data, the class-based language modelling (CBLM) approach in natural language processing (NLP) enhances language modelling. Traditional language models use each word or subword as a distinct unit during training and prediction, such as n-gram models or neural network-based models like recurrent neural networks (RNNs) and transformers. Contrarily, CBLM classifies words according to how similar they are and models the language at the class level.

The fundamental goal of CBLM is to increase vocabulary efficiency by grouping words into classes according to their semantic or syntactic similarity. Frequently, unsupervised methods like Brown clustering or K-means clustering are used for this clustering process. The language model learns to anticipate the following word or series of words given the preceding context and the current class information once the words have been classified into classes.

CBLM offers a number of benefits by employing classes as opposed to specific phrases. First, because there are often fewer classes than there are distinct words in a language, the complexity of the language model is reduced. This results in quicker inference and more effective training. Second, it aids in handling words that are not in the user's vocabulary (OOV) by classifying them according to context. As opposed to treating OOV words as wholly unknown, CBLM can offer accurate predictions based on words of the same class that are comparable. Third, by learning from the class-level data, CBLM can capture more universal patterns and enhance generalization.

CBLM has been used for a variety of NLP tasks, including sentiment analysis, part-of-speech tagging, named entity identification, language modelling, and machine translation. It has proven to be very helpful in situations where there is a big amount of language and little training data available.

Techniques of Natural Language Processing

These are most common techniques

1. Machine Translation

It is the process of mechanically translating text from one human language to another. It is a really challenging problem and is included in the category of AI-complete challenges. It requires the many types of information that people have (such as understanding of semantics, syntax, and real-world concepts) in order to fully resolve the challenges of translation.

2. Examination of Discourse

There are several tasks that are linked to the discourse analysis task. Determining the related text's discourse structure—that is, the links between lines of speech, such as contrast and explanation—is one such task. Identifying and classifying speech activities in the specific text is another task. For instance, a statement or an assertion, a yes-or-no question, or a content inquiry, etc.

3. Splits Morphologically

Split words to identify distinct morphemes and their respective categories. The primary issue with this task is that it heavily depends on the complicated word structures in the language that we are thinking about. Since the morphology of the English language is so straightforward, especially when it comes to morphology connected to inflection, it is frequently possible to ignore the task entirely and create many plausible versions of any phrase. Using the phrases opened, opens, and opening differently, for instance. However, this method cannot be used with the Turkish language since each entry in the dictionary might have a huge number of workable word forms.

4. Generation and Understanding of Natural Language

The process of generating natural language entails converting data from computerised databases into the easily understandable language of humans. To understand natural language, text portions must be converted to much more formal notations, such as first-order logical structures, which are considerably simpler for computer programmers to manage. Additionally, it deals with the recognition of semantics utilizing a variety of plausible semantics that are acquired using natural language expressions, which are often in the form of organized notations found in notions of natural languages. The creation of ontologies and meta-models in languages are excellent remedies that are empirical in character. In order to generate formalization's of natural language, it is necessary to formalize its semantics by making assumptions about things like closed vs. open terms and objective vs. subjective in the absence of ambiguities.

5. Identification of Named Entities

With the use of the input text, identify phrases may be classified as named entities, such as organisations, locations, or persons, as well as the kinds to which these named entities belong. Although it is useful for recognising names that are found in languages like English, capitalization is not useful for distinguishing the types of names.

6. Parsing of Text

By creating a parse tree of the sentences, parsing analyses their grammar. We are aware that natural language grammars frequently contain ambiguity and that certain lines might support many viable interpretations. There are millions of different parses for each given line, many of which are entirely incomprehensible to humans.

7. Recognition of Speech

Recognising written notation of any speech while listening to a sound recording of any individual is known as recognition of speech. It is a highly challenging problem and completely different from the text to voice conversion assignment. Furthermore, there won't be many gaps between phrases in genuine speech, therefore we can argue that segmenting speech is a crucial stage in the recognition of speech. The process of turning an analogue sound signal into discrete textual characters is known as co-articulation, and it occurs when sounds signifying successive words are said together in many spoken languages.

8. Analysis of Sentiments

Recognition In order to determine the polarity of certain items, sentiment analysis [2] works with extracting information that is subjective in nature often from collections of text documents like internet reviews. This technique is heavily utilised to determine It is often used in marketing to ascertain the attitudes or reviews of the general population towards social media.

9. Finding Words Boundary

It deals with dividing parts of text with continuous characters into distinct phrases. This exercise is fairly easy to do in English because spaces are typically used to separate phrases. However, several languages throughout the world, including Japanese, Chinese, and Thai, lack clear term borders, making it difficult to segment words in these languages because doing so requires knowledge of their morphology and lexicon.

10. Word Sense Disambiguation

One word can have a variety of meanings depending on the situation, according to word sense disambiguation. It might be challenging to determine the proper meaning of a word when it is employed in a particular statement. We may utilise Word-Net, which has a list of terms and related word senses, for a given language to help us solve this issue.

Here are a few well-known NLP strategies.

1. Named Entity Recognition (NER)

NER techniques aim to identify and classify named entities (such as names of people, organizations, locations, etc.) in text. Advanced NER models employ deep learning techniques, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), to achieve better accuracy in entity recognition tasks.

2. Attention Mechanisms

Attention processes: In order to improve NLP models, attention processes have been extremely important. They enable models to generate or decode output while concentrating on pertinent segments of the input sequence. Attentional processes help with tasks like machine translation, text summarization, and question answering as well as the comprehension of long-range relationships.

3. Neural Machine Translation (NMT)

To improve the accuracy of machine translation, NMT uses deep learning methods. NMT uses neural networks to translate text across languages directly rather than using conventional statistical methods. Better accuracy and fluency in translation tasks have been demonstrated using this method.

4. Transfer Learning

Transfer learning is a method that allows pre-trained models to be improved on certain NLP tasks. With this method, models may first learn from broad general language patterns before being modified for use on smaller datasets for more specialised tasks. By utilising the information learned during pre-training, performance in NLP has been greatly enhanced.

2.0 CONCLUSION AND RECOMMENDATION

Conclusion

The goal of processing of natural language, a subfield of linguistics, artificial intelligence, and computer science, is to enable communication between computers and human language. We may say that it has something to do with computer-human interaction. The most common tasks for NLP include: discourse analysis, morphological separation, machine translation, generation and understanding of NLP, recognition of named entities, part-of-speech tagging, recognition of optical characters, recognition of speech, and sentiment analysis. Unsupervised or semi-supervised learning techniques are currently receiving greater attention in NLP research.

Recommendations

Transformer Models: Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), are powerful deep learning models that have shown great success in natural language processing (NLP) tasks. They excel at capturing contextual relationships in text, making them suitable for tasks like machine translation, text generation, and language understanding.

BERT (Bidirectional Encoder Representations from Transformers): BERT is a pre-trained transformer-based model that can be fine-tuned for specific NLP tasks. It learns contextual representations of words by considering the surrounding words on both sides. This bidirectional approach allows BERT to capture rich semantic information and perform well on various tasks such as sentiment analysis, named entity recognition, and question answering.

LSTM (Long Short-Term Memory): LSTM is a type of recurrent neural network (RNN) architecture that is commonly used in NLP. It is designed to capture long-term dependencies in sequential data, such as text. LSTM models are effective in tasks that require understanding of sequential patterns, such as language modeling, sentiment analysis, and text classification. They are especially useful when dealing with tasks that involve understanding the context of previous words or phrases in a sentence.

REFERENCES

B. Manaris, "Natural Language Processing: A Human–Computer Interaction Perspective," Appears in Advances in Computers (Marvin V. Zelkowitz, ed.), Academic Press, New York, vol. 47, 1998, pp. 1-66.

Jasmin Praful Bharadiya https://journalajorr.com/index.php/AJORR/article/view/164

P. M Nadkarni, L. O. Machado and W. W. Chapman, "Natural Language Processing: An Introduction,"J Am Med Inform Assoc, vol.18, 2011, pp. 544-551.

R. Mihalcea, H. Liu, and H. Lieberman, "NLP (Natural Language Processing) for NLP (Natural Language Programming), " In Proceedings of CICLING'06, LNCS, Springer, 2006, pp. 319-330.

Jasmin Praful Bharadiya https://journaljerr.com/index.php/JERR/article/view/858

W. Fan, L. Wallace, S. Rich and Z. Zhang, "Tapping into the Power of Text Mining", International Journal of ACM, Blacksburg, 2005.

A. Lopez, "Statistical Machine Translation," Iternational Journal of ACM Computing Surveys, vol. 40, 2008

https://ijisrt.com/convolutional-neural-networks-for-image-classification

Z. B. Wu, L. S. Hsu, and C. L. Tan, "A survey on statistical approaches to natural language processing," Technical Report, 1992.

B. B. Ali and F. Jarray, "Genetic Approach for Arabic Part of Speech Tagging," International Journal on Natural Language Computing (IJNLC), vol.2, 2013, pp. 1-12