A Deep Insight in Challenges of Natural Language Processing and Usage of Deep Learning

Varsha Mittal, Durgaprasad Gangodkar Bhaskar Pant, Nisha Chandran

Abstract*: Natural Language Processing (NLP) using the power of artificial intelligence has empowered the understanding of the language used by human. It has also enhanced the effectiveness of the communication between human and computers. The complexity and diversity of the huge datasets have raised the requirement for automatic analysis of the linguistic data by using data-driven approaches. The performance of the data-driven approaches is improved after the usage of different deep learning techniques in various application areas of NLP like Automatic Speech Recognition, POS tagging etc. The paper addresses the challenges faced in NLP and the use of deep learning techniques in different application areas of NLP.

Index Terms: Artificial Intelligence, Deep Learning, Natural Language Processing; Machine Learning.

I. INTRODUCTION

Natural Language Processing (NLP) is the branch of computer science that provides a way to empower computers to process, understand and analyze human language [1]. In the initial years of NLP, the data driven approaches including statistics and machine learning is used for computation [2]. In recent years, the advancement in computational capabilities and huge datasets has enabled the deep learning techniques to be more suitable for NLP tasks [3]. In the various fields like the computer Vision , Automatic image captioning and Speech recognition, deep learning approaches have shown better performance than conventional data-driven approaches [4]. This has shifted the paradigm from data-driven techniques to deep learning techniques, as the results are more promising and easier to generate. Researchers are focusing on leveraging the benefits and power of Deep Neural Networks (DNN) for core NLP tasks where they can be directly applied to achieve the results [5]. This survey provides a deep insight to the challenges of NLP and the use of deep learning techniques to address the challenges of NLP. The survey also discusses the present state of the DNN techniques before boarding to the advance research. The overview of NLP and the challenges of NLP are discussed in Section II. The core issues of NLP with deep learning techniques are discussed in Section III. Different applications of NLP with deep learning techniques are discussed in section IV. Finally, the paper is concluded in Section V by providing the suggestions to improve in the evolving field.

Revised Manuscript Received on October 20, 2019.

Varsha Mittal, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India.

Durgaprasad Gangodkar, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India

Bhaskar Pant, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India.

Nisha Chandran, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India.

II. NATURAL LANGUAGE PROCESSING AND CHALLENGES IN NLP

A. Natural Language Processing

Natural Language Processing can also be referred as Computational linguistics. It includes the computational algorithms to represent, process and understand human language. The different stages of NLP are: lexical analysis, parsing, semantic analysis and pragmatic analysis. The NLP work can be categorized into two areas- core areas and the application areas although there is no clear distinction between the two. The core area includes the problem of language modelling, morphological processing, syntactic processing and semantic processing [6]. The application area of NLP includes the extraction of important entities and relations, text summarization, question answering, language translation and document classification. It is required to handle the challenges of core areas and then implement those ideas to solve the problems in application area. NLP achieved success to solve the problems like part-of-speech tagging. But for the tasks like machine dialog, text summarization many challenges are still open. The problems of NLP can be divided into two groups:

1. Data-related problem

2. Understanding-related problem

B. Data-Related Problem

NLP uses data-driven approaches, but what type of data is required is a challenging question to be answered. The heterogeneous, incomplete, noisy and unbalanced data decrease the effectiveness of NLP tools. To define a NLP task, it is required to construct the dataset and design the evaluation methods to evaluate the progress. The different data-related problems are discussed below:

1. *Low-resource Language:* There are many languages which are very popular like English but there are also the languages for which the data is very scarce. A survey report shows that in Africa only, 1250-2100 languages exist for which the data is very scarce[7] . So, to transfer the tasks that entail understanding from high-resource language to a low-resource language is quite challenging.

To exploit the universal commonalities between different languages the cross-lingual Transformer model and cross-lingual sentence embeddings are used. But these models are sample-efficient since they need monolingual data or the word translation pairs. The development of cross-lingual datasets, have shown the improved performance in cross-lingual models. But building the efficient models for such low-resource language is still a challenging task and needs the attention to work on.

2. *Large or multiple documents:* Reasoning about large documents is another problem of NLP. The current model uses recurrent neural network that do not remember longer context. Working with longer context is related to NLU and needs the scaling of the current systems until they read the entire book[7] .

Another problem with large documents is that their supervision is expensive and scarce to get. We can visualize a document-level task that needs to predict the next chapter or next paragraph of book. However, achieving this objective is sample-inefficient. It is required to develop the methods through which we can represent the context more efficiently and should be able to track the relevant information of the document. Multi-document summarization is a step ahead in this direction.

C. Natural Language Understanding Problem (NLU)

The problem of natural language understanding is still the most critical for analysing and processing the text. Many researchers argued that NLU is a criterion for Natural language generation (NLG) tasks. None of the current models have the clear understanding of the natural language. Various issues in NLU are described below:

1. *Ambiguity:* Modeling of language elements within different context is the main challenge of NLP. Since in the natural language same word have different meaning depending on the context it is used. It results in ambiguity at the lexical, semantic as well as syntactic levels. Different approaches like POS tagging is proposed to address the issue [2]. However, understanding the meaning of a word in a phrase is still a challenging task.

2. *Synonymy:* In natural language we use the concept of synonyms that the same meaning can be expressed by different words depending on the context where it is used. For example: huge, large, vast and big can be synonym but they can not be interchanged in every context as it can be big sister but cannot be substituted by huge sister. For NLP tasks it is required to use the knowledge of synonyms and it becomes more challenging for huge and complex data especially when imitating human dialog.

3. *Coreference:* The process of extracting the expression that denote the same entity is known as Coreference resolution. It is a prominent step for many NLP tasks like document summarization, information extraction and automatic question answering. Use of deep learning techniques and reinforcement learning approaches have shown the better results in solving the problem of coreference resolution. At present, it is argued that the use of Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) architectures may further enhance the results.

NLP has taken a large leap from machine learning to technology that has faster pace of advancement and innovation. The collaboration of NLP with Deep Learning have begun to yield good results and can provide solution to the NLP problems.

III. DEEP LEARNING IN ADDRESSING THE ISSUES OF NATURAL LANGUAGE PROCESSING

To perform the NLP tasks like text summarization or automatic image captioning requires the understanding of the natural language. The understanding acquired from the language can be divided into four areas: language modelling, morphology, parsing and semantics. But all these areas are not completely disjoint rather they overlap each other. This section discusses the models with logical connections among all the above specified areas.

A. Language Modeling

Language modelling is one of the important task of NLP. It is the process of generating a model for predicting words or other linguistic components from the given words or previous components. It is used for the applications where user's input provides predictive capability for the entered text. The power of language modeling originates from the point that it captures sematic as well as semantic relationship among the linguistic components in a linear neighbourhood pattern. This capability of language modelling make it useful for the applications like machine translation or text summarization. Moreover, the predictions made through it can be used to produce more relevant sentences.

B. Morphology

The study of word formation identifying the different segments like roots, prefixes and suffixes of the words is called Morphology. It also considers other intra word devices to show the tenses, plurality and gender. A good morphological analyzer is required for different NLP task to achieve accurate results. A survey by Belinkov et.al [8] showed that the morphology is used and implemented in different neural network translation models to a larger extent. Many translation models are created for translation from one language to another like from English to French.

The current research in morphology focused on the development of the Universal morphological models. It is required to study the relationships among the different language's morphologies so that a universal morphological analyzer can be generated. The collaboration of deep learning with NLP will enhance the handling of morphological components in a better way and it will improve the performance of the morphological analyzer.

C. Parsing

The study of relationship of different words with other words and phrases in the sentence is called parsing. Parsing can be done by two methods either by constituency parsing or through dependency parsing [9].

In constituency parsing, the phrases or words within a sentence are extracted in a hierarchical style.

In dependency parsing the relationship between two individual words are extracted. In recent years, mostly all the deep learning techniques use the dependency parsing. The Graph-based methods are used to generate the parse trees which uses the formal grammar of the natural language [9].
Socher et.al [10] proposed the use of RNN to generate the Probabilistic context-free grammars (PCFGs). Dyer et.al [11] used LSTM in place of RNN. Since LSTM can remember the long context knowledge thus give better results in predictions on Stanford Dependency Treebank. The current research aims at the development of Universal parsing so that the standardized tags and relationship for all languages. Nivre [12] presented the challenges and the recent development in generating the treebank while using Universal parsing. Still there are many challenges that exist in universal parsing and is expected to receive focus.

D. Semantic

Semantic processing refers to the process of capturing the meaning of words, phrases, sentences or documents. Many deep learning techniques like Word2Vec [13], GloVe [14] showed an improved performance in capturing the meaning of words and used the distributed representation of words. The challenges still exist in integrating the deep neural techniques with distributed representation of words WordNet. The concept of knowledge graph and graph embedding is showing a better take off for improved machine understanding [15].

Use of deep learning techniques has enhanced the performance of the NLP tasks and also provides a better way to handle the core issues of NLP.

IV. APPLICATION OF NLP WITH DEEP LEARNING

This section discusses the use of deep learning for different NLP tasks. Various algorithms to solve NLP tasks and improvement by the use of deep learning approaches are summarized below.

A. Information Retrieval

Information Retrieval (IR) system aims at providing the right information at right time in right format. One of the major problems of IR is to rank a document with respect to the query submitted [16]. The deep learning models got better scores for retrieving the matched documents with respect to the text in query. Deep learning models uses two types of approaches: representation-focused approach or interaction-focused approach. In representation-based approaches, initially deep learning models are used to generate the good representation for the text and later on compare the representation [17]. Whereas in interaction-based approaches initially, the local interactions are build and then deep neural models are used for text matching [18].
The queries are shorter as compared to the documents and also have less information than the document so the representation of query should be denser. Thus, a hybrid approach called CEDR (Contextualized Embeddings for Document Ranking) is used to obtain BERT token [19]. The hybrid approach using BERT token representation has shown the better results in text matching.

B. Information Extraction

Information Extraction refers to the process of extracting information from the text. The extracted information contains the named entities, events and the participant of the event, and finally the relationship between the entities and events. The extracted components of the information are discussed below:

1. *Named Entity Recognition (NER) :* It aims the extraction of proper nouns and other important information like date, time etc. The LSTM model was used by Hammerton [20]. But due to the lack of computing resources at that time it showed a slight improvement from the baseline methods. An architecture using the bi-directional LSTM was developed by Lample et al. [21]. The character-level inputs and word embeddings are used in bi-directional model. Another improved bi-directional model was proposed by Akbik et al [22]. The proposed model uses contextual embedding for each word that is passed to Bi-LSTM sequence labeller to improve the performance of NER

2. Event Extraction: Event extraction is concerned with the extraction of the words that are related to the occurrence of an event including the participant of the event. While using CNN for event extraction it is identified that only the most important information of a sentence is captured in a max pooling layer. Whereas other valuable facts are missed which later on can be used to relate the events [23]. The drawback is addressed by Nguyen et.al [24] by using a RNN based encoder-decoder to extract the event and role of event trigger.

C. Text Classification

Text classification is one of the important applications of NLP. It refers to classify the documents to the predefined labels or classes. A CNN model using pretrained word vector was developed for sentence-level classification [25]. It was shown that a dense layer following the convolutional layer with drop-out and softmax function in output layer could obtain better results. CNN models have improved the performance in sentence classification, question classification and sentiment classification. Later on [26] showed that CNN also works well for document classification also but the number of layers should be increased.

A hybrid architecture called Dynamic Convolutional Neural Network (DCNN) uses k-max pooling to capture semantic modelling of sentences [27]. An LSTM-RNN framework has been used for sentence embedding especially for web search [28]. Both RNN and CNN models are combined in some models used for text classification. Here recurrent architecture along with max-pooling achieved superior results as compared to simple neural-based models. Another model C-LSTM is also proposed for sentence and document classification [29]. The model uses the long-term dependency to improve the accuracy of the text classification

V. CONCLUSIONS

In this paper, a comprehensive survey is presented that includes the challenges of Natural language processing, core issues of NLP with respect to deep learning and the achievement in NLP using deep neural techniques. It has been observed that the use of deep learning for NLP tasks has enhanced the performance.

The accuracy in text classification and other NLP related areas has also improved. NLP and Deep neural are the two most rapidly growing areas. So it is hoped that the collaboration of both the fields will help the researchers to develop new models that supersede the current approaches. This survey will help the researchers to get insight of the problems of integration of NLP with deep neural network.

REFERENCES

1. C. D. Manning, and H. Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999.
2. E. D. Liddy, "Natural language processing," 2001.
3. R. Collobert and J. Weston, "A unified architecture for Natural Language Processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning, pp. 160–167, ACM, 2008.
4. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017, pp. 2242–2251.
5. Y. Liu and M. Zhang, "Neural Network Methods for Natural Language Processing by Yoav Goldberg," Comput. Linguist., vol. 44, no. 1, pp. 193–195, Mar. 2018.
6. T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," Ieee Computational Intelligence Magazine, vol. 13, no. 3, pp. 55–75, 2018.
7. N. Ranjan, K. Mundada, K. Phaltane, and S. Ahmad, "A Survey on Techniques in NLP," Int. J. Comput. Appl., vol. 134, no. 8, pp. 6–9, Jan. 2016.
8. Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass, "What do neural machine translation models learn about morphology ?",Proc. of 55th Annual Meeting of Association of Computer Linguistic, Vancouver, Canada, pp. 861–872, July 2017.
9. D. Jurafsky and J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, vol. 2. 2008.
10. R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Oct. 2013, pp. 1631–1642,
11. C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory", Proc of 7th International Joint Conference on National Language Processing, Beijing, China, pp. 334-343, July 2015.
12. J. Nivre, "Towards a Universal Grammar for Natural Language Processing," in Computational Linguistics and Intelligent Text Processing, Cham, pp. 3–16, Aug 2015.
13. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.
14. J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, , pp. 1532–1543, Oct. 2014.
15. Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," IEEE Transactions on Knowledge & Data Engineering, vol. 29, no. 12, pp. 2724–2743, 2017.
16. T. Kenter, A. Borisov, C. Van Gysel, M. Dehghani, M. de Rijke, and B. Mitra, "Neural Networks for Information Retrieval," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, , pp. 1403–1406, Aug. 2017.
17. L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text Matching as Image Recognition," in 30th AAAI Conference on Artificial Intelligence, Feb. 2016.
18. J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, pp. 55–64 2016.
19. S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "CEDR: Contextualized Embeddings for Document Ranking," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, pp. 1101–1104,July 2018.
20. J. Hammerton, "Named entity recognition with long short-term memory," in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Edmonton, Canada, pp. 172–175, May 2003.
21. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, pp. 260–270,Jun 2016.
22. A. Akbik, D. Blythe, and R. Vollgraf, "Contextual String Embeddings for Sequence Labeling," in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1638–1649, Aug 2018.
23. Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing Beijing, China, pp. 167–176, July 2015..
24. T. H. Nguyen, K. Cho, and R. Grishman, "Joint Event Extraction via Recurrent Neural Networks," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California,pp. 300–309, Jun 2016.
25. Y. Kim, "Convolutional Neural Networks for Sentence Classification", Proc, of 14th International Conference on Empirical Methods of Natural Language Processing, Doha, Qatar, pp. 1746-1751, July 2014.
26. A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very Deep Convolutional Networks for Text Classification," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, Apr. 2017, pp. 1107–1116.
27. N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," ,Proc. of 52th Annual Meeting of Association of Computer Linguistic, Baltimore, Maryland, pp. 655-665,June 2014.
28. H. Palangi et al., "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval," IEEEACM Trans. Audio Speech Lang. Process., vol. 24, no. 4, pp. 694–707, Apr. 2016.
29. C. Zhou, C. Sun, Z 33. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," ArXiv151108630 Cs, Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08630.

AUTHORS PROFILE

Varsha Mittal received the M.Tech degree in Computer Science and Engineering from Graphic Era University, Dehradun, India. She is Pursuing her Phd in Computer Science & Engineering from Graphic Era University, Dehradun, India. She is working as Assistant Professor in Computer Science and Engineering in Graphic Era University since 2007. Her research interest include Big Data Analytics, IOT, Machine Learning, Natural Language Processing and soft computing. She has published many research papers in International Journal and conferences in this area

Durgaprasad Gangodkar received the B.E. degree in electronics and communication engineering from Karnatak University, Dharwad, India, and the M.Tech. degree in computer network engineering from Visvesvaraya Technological University, Belgaum, Karnataka, India. He received his Ph.D. degree with the Department of Electronics and Computer Engineering, Indian Institute of Technology (IIT), Roorkee, India. He is currently with the Department of Computer Science and Engineering, Graphic Era University, Dehradun, India. He is the author or a coauthor of papers that have been published in international journals and conference proceedings. His research interests include high-performance computing, computer vision, video analytics, and mobile agents.

Bhasker Pant has received his Ph.D degree from NIIT Bhopal ,India. Currently he is working with the Department of Department of Computer Science and Engineering, Graphic Era University, Dehradun, India. He is having the teaching experience of more than fifteen years. He is the author or a coauthor of papers that have been published in international journals and conference proceedings. His research interests include Bio Informatics, Data Mining, Soft Computing Big Data Analytics.