Survey on Deep Learning Approaches for Phrase Structure Identification and Classification of Sentences

HASHI HARIS[1], MISHA RAVI[2]

[1]M. Tech., Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.
[2]Assistant Professor, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.

Abstract- *Phrase structure is the arrangement of words in a specific order based on the constraints of a specified language. This arrangement is based on some phrase structure rules or we can say according to the productions in context free grammar. The identification of the phrase structure can be done by breaking the specified natural language sentence into its constituents that may be lexical and phrasal categories. These phrase structures can be identified using parsing of the sentences which is nothing but syntactic analysis. This paper explains about the existing deep learning techniques for identification and classification of sentences.*

***Key Words*: Phrase structure grammar, Natural Language processing, Deep learning, Indian languages, Context Free Grammar.**

1. INTRODUCTION

Natural Language is the mode of communication between humans. Humans interact with each other using a natural language by speech or text. We are surrounded by lot of data or texts that may be a web page, a document, an email or it can be even an sms. We speak to each other using a natural language which is more easier than text. The natural language that is understandable to humans can be also manipulated by a machine or a software which directs us to the broad field of Natural Language Processing(NLP). NLP is the ability of computer to process a natural language with the help of computer software. The researches on NLP is still going on and has been conducted from past 50 years. NLP is a component of AI and the most recent works are been conducted on developing NLP models using deep learning approaches. Since NLP can interpret and analyze text and deep learning is more flexible in which the algorithms learn to identify the speaker's intention from many examples as if a child would learn human language. Some of the Artificial Neural Network applications in NLP are Text Classification and Categorization, Part of Speech tagging, Paraphrase detection, Machine Translation, etc.

Text Analytics is a challenging field but deep nets have already made a big impact in NLP. Some of the drawbacks of NLP are its ambiguity and at times same words can have different meanings according to their usage in different context, it's also difficult to add a new word to existing words and thus NLP requires lot of manual curation. Deep learning is an important tool that overcomes some limitations of NLP. The basic difference between deep learning and traditional NLP is the use of vectors. A deep learning model uses one-hot vector representation in which each word is represented as a vector. This solves the problem of huge size of vocabulary. In this paper we reviewed the recent deep learning models such as Convolutional neural networks(CNN), Recurrent neural networks(RNN) and Recursive neural networks that are applied on NLP tasks. A large number of works have been done in languages like English, Chinese, etc. but only a few works have been done in Indian languages especially South Dravidian languages because they are morphologically rich and agglutinative. Another problem is that such languages are resource poor. This leads to a challenging task to identify the phrase structure of sentences.

1.1 Basic N- gram based approach

An n-gram is nothing but a longer set of strings sliced into an n character string. The benefit of using n gram is that the errors would be affected to only small number of parts since the strings are sliced into small strings the errors also would be only in that small portion. This approach is statistical so as to obtain the phrases. William B. Cavnar and John M. Trenkle proposed an approach which extracted phrases using n-gram model [1]. Gulila Altenbek, Ruina Sun used N-gram models for phrase structure extraction from unannotated monolingual corpus [2].

1.2 Rule based approach

Gulila Altenbek, Ruina Sun used rule based method [2] for noun phrase extraction from monolingual corpus. The phrases are extracted based on a rule set of the target language. The corpus is searched and the phrases that match the rule set are been extracted. Rule based approaches give a better perfomance accuracy than the statistical methods at the same time it requires special linguistic knowledge to develop. Ramshow and Marcus used transformation rule based learning [3]. They applied a heuristic prediction on the training corpus and then they applied these rules on the learned rule sequence. The rules were learned by deriving candidate rules, scoring these rules and based on the score the rules with maximum positive effect was selected.

2. DEEP LEARNING APPROACHES

Recently many deep learning approaches have been used in computer vision applications. Inspired from those

models deep learning algorithms and architectures are been used in various NLP tasks. Collobert et al. [4] in his work revealed that a simple deep learning framework performs better for NLP tasks than the existing approaches. In this paper we have presented a review on deep learning approaches which could be applied for the identification of phrase structure of sentences in Indian languages.

First of all we have to represent the document or a natural language sentence in a distributed representation. The distributed representations are of different types like the continuous bag-of-words model (CBOW), skip gram model or count based model. These are called the word embedding techniques which could represent the words as a real vector. This is a neural representation of words. Recently many researchers also found that there is no need for the deep neural networks to process or change the document in such a way because already the words in the document are represented in one-hot vector representation. This is a method to represent a word in the form a vector. Each dimension of the vector represents a word. The traditional methods that we discussed above like n-gram are statistical and they have data sparsity issues. To solve such data sparsity issues deep neural nets use such word embedding representations. These embeddings have proven to be effective in capturing the context similarity, syntactic, semantic information and since they are having smaller dimension they are fast in computing core NLP tasks. So we can say that this word embedding is the first data processing layer in deep neural nets. Word2vec[5] by Mikolov et al is such a model that computes the vector representation of words. In his work he proposed the CBOW and skip gram models. CBOW model computes the conditional probability of the target words from the surrounding context words within a window size k. Skip gram model but does the opposite of the CBOW model. It predicts the context words from the central target words. Apart from word embeddings character level embedding has also been used specially in morphologically rich languages. Bojanowski et al. [6] tried to improve the representation of words by using character-level information in morphologically-rich languages. Thus a number of researches are going on the word representations in neural networks itself in order to improve the understanding of the concept behind words.

2.1. Convolutional Neural Networks

Convolutional neural networks have been used in many NLP tasks so far such as the pos tagging, chunking, sentence modeling, etc. CNNs are designed in such a way that can capture more important features from sentences. Works have been reported in literature on sentence classification. Such a work done by Kim[7] on sentence classification tasks such as question type classification, sentiment, subjectivity classification. But this vanilla network had difficulty in modeling long distance dependencies. This issue was solved by designing a dynamic convolutional neural network. CNNs require large training data in developing semantic models with contextual window. This becomes an issue when data scarcity occurs. Another issue with CNNs is they find difficulty in preserving sequential data.

2.2. Recurrent Neural Networks

RNN's are used to process sequential data. The peculiarity of recurrent networks is that they store previous computation information and using that they repeat the same task in several epochs. The output in each instance depends on the previous data. They possess memory of the previous computations so that it can be used for current computations. Due to such advantages RNN's are recently been successfully used in many NLP tasks such as language modeling, speech recognition, etc. RNN's are suitable for many NLP tasks because it has an inherent nature for handling context dependencies in a language. It can handle variable length text that may be sentences, paragraphs or documents. RNN's are suitable for semantic modeling in sentences. Simple RNN network models are based on Elman network [8]. In order to overcome disadvantages of the simple RNN network like the exploding and vanishing gradient problems RNN's with LSTM [9], GRU [10] are used for many NLP tasks.

2.3. Recursive Neural Networks

Natural language possesses a hierarchical recursive structure. Such a hierarchical structure can be modeled by a recursive neural network where words are combined to phrases and the phrases join to form a sentence. Such a syntactic interpretation can be represented by a consistuency or dependency parse tree structure. The sentence structure can be modeled by such tree structured models. For more interaction between the input vectors recursive neural tensor networks (RNTN) have been introduced. Recursive models are widely used in applications like natural language parsing [11]. The logical relation between sentences was classified by Bowman et al. [12]. Phrase sentiment analysis was done by Socher et al. [13] where each node was named by a sentiment label in the parsing tree.

3. CONCLUSION

Deep learning has been widely used in many NLP tasks as it needs only little engineering by hand. A word embedding representation in the dimensional space makes easier for the neural nets to capture input data and so they are been widely used in various natural language processing applications. But still there are various areas in which deep learning is still in its childhood stage as in case of processing with unlabeled data. But with the developing researches it is expected that deep learning would become more enhanced and can be applied in different areas.

REFERENCES

[1] Cavnar,W.B.and Trenkle, J. M, (1994) "N-grambased text categorization", In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval ,161-175

[2] Gulila Altenbek & Ruina Sun, (2010) "Kazakh Noun Phrase Extraction based on N-gram and Rules", Internation conference on Asian Language Processing

[3] L. Ramshaw and M. Marcus. (1995) "Text Chunking using Transformation-Based Learning" In Proceedings of the Third Workshop on Very Large Corpora

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing from scratch," Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," arXiv preprint arXiv:1607.04606, 2016.

[7] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[8] J. L. Elman, "Finding structure in time," Cognitive science, vol. 14, no. 2, pp. 179–211, 1990.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[11] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 129–136.

[12] S. R. Bowman, C. Potts, and C. D. Manning, "Recursive neural networks can learn logical semantics," arXiv preprint arXiv:1406.1827, 2014.

[13] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts et al., "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, 2013, p. 1642.

BIOGRAPHIES

Hashi Haris, she is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Elavumthitta, Kerala, India. Her research area of interest includes the field of natural language processing, Deep learning.

Misha Ravi received the master's degree in Software Engineering from Cochin University of Science and Technology, Kerala. She is an Assistant Professor in Department of Computer Science and Engineering, at Sree Buddha College of Engineering.