Development of an Approach for Discourse Analysis of Natural Language

Animesh R. Tayal¹, S.R. Tandan², Rohit Miri³ and Madhuri A. Tayal⁴

¹Research Scholar, Department of Computer Science & Engineering,
Dr. C.V. Raman University, Bilaspur (CG), India

^{2,3}Associate professor, Department of computer science & Engineering,
Dr. C.V. Raman University, Bilaspur (CG), India

⁴Assistant Professor, Department of Information Technology, Shri Ramdeobaba
College of Engineering and Management, Nagpur (MH), India

ABSTRACT

We are alive in the age of brainy machineries where the whole thing is substituting by machinery, understanding of concealed meaning of text/speech turn into essential to comprehend by machine. Natural Language Understanding (NLU) is an emerging area of Computational Linguistics under the subarea of Natural language processing that is used to analyse and distinguish text or speech sentences written in human understandable language into machine understandable presentation. The purpose of NLU system is to interpret a text. The NLU expects that human languages must be interpreted statistically by the system. Understanding natural language is biggest challenge in Artificial intelligence field as well as Discourse processing. Discourse is "Language above the Sentence or above the Clause". This paper presents a method which designs and builds computer system that are able to analyse and understands human language and to generate output in same human understandable language. The system which is designed for discourse implementation is tested on IMDb movie review dataset. IMDb Stands for Internet Movies Database. The results are quiet promising. We have used Convolution Neural network concept for the same. The efficiency of this model is around 87%.

KEY WORDS: DISCOURSE, DISCOURSE ANALYSIS, TOKENIZATION, TAGGING, NAME ENTITY RECOGNITION (NER), SYNTACTICAL/ SEMANTICAL ANALYSIS.

INTRODUCTION

The Technology is making a great impact on us in every facet of our life whether it is through making our life relaxed by making better decision with the help of Artificial Intelligence or understanding the human and

understanding the language. In the hierarchy of artificial intelligence, Natural language understanding (NLU) plays important role, which empowers human computer interaction. NLU practices the system programs which must be understood either sentence text or speech. NLU forces computer to understand the instructions and then system joins back to humans in their particular dialects.

The NLU in our system will help to understand the conversation and provide the analytical thought to the conversation. The subdomain of it, is Discourse processing. Discourse processing is identifying the unseen sense from the portion of data. In this application, the system will analysis the reviews and check for the semantic and syntactic analysis, then based on these

ARTICLE INFORMATION

*Corresponding Author: animesh9777@gmail.com
Received 10th Oct 2020 Accepted after revision 29th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and
Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: <http://www.bbrc.in/>
Doi: <http://dx.doi.org/10.21786/bbrc/13.14/31>

analysis and training of these movie review datasets it will interpret how good or bad the movie is. Currently, our application can predict the movie review with that much accuracy as expected. The designed system is to be expected as:

Conversation-1: How was the movie U saw?

Conversation-2: Movie was fabulous and meanwhile story was good and after interval story makes sense.

The designed system should give Machine Interpretation as :- Movie was good and can be watch.

This way we want to analyse actual discourse from the text data.

This paper is arranged as follows: first section gives the idea of the concept Discourse processing along with Natural language Understanding. Second Section gives the existing methods and third section defines the method used for this system to be developed. Paper ends with results and conclusion.

Literature Review: There were many efforts made in this area of Discourse processing for identifying hidden meaning from given sentence/Text, listing here some of them. The concept of Neural Discourse Segmentation (Yizhong Wang et.al, 2018) which recommends an effective neural segmenter which is constructed on the basis of BiLSTMCRF framework. This concept is very fast and effective. They have used a word representation model which was trained on an enormous corpus. Altogether this is done for improving efficiency of system and solved the problem of data insufficiency. Overall the results for this investigation are good.

A comprehensive review based on Statistical learning for semantic parsing is specified by which has reviewed the current mechanism for semantic parsing with predictable deep learning techniques, machine learning techniques. For the received input order our system must give correct response, then we can say that machine understood well. The responses will be of various forms: a desired movement of robot, or a desired response for a given query. This can be accustomed through semantic parsing. The logical form, i.e semantic illustrations are used for machines so that it will be understood by it. In this work the system of parsing is given and further converted into learning statistically. (Xiyao Ma et.al).

The analysis of Named Entity Recognition (NER) is also required for this task, a complete survey of named entity recognition and classification given by David Nadeau et al. It focuses on Relation Extraction (RE), Co-reference Resolution (CO), Named Entity Recognition (NER), and Event Extraction (EE). All these concepts are utilised to categorize and extract the important entities like some expressions, proper names, organizations species etc. The various NER techniques have been presented in this paper. These NER techniques are in a position to classify Personnel to organization information as well as rare species, cells and proteins in various applications. (David Nadeau et.al)). Interpretation of the discourse and conversation prearrangement of SMS communication is

discovered in (Nianwen Xue, Qishen Su, 2016) similarly on chat dialog (Forsythand et.al, 2007). One of the algorithms considers information related to discourse for creation of categories clustering outcomes based on better syntactic parse trees. (Tatyana Makhalova et.al, 2015) (Ashraf Ullah et.al, 2014).

In one of the paper media contents were analysed from the newspapers "The Nation", of discourse from media. The different kinds of text were used in media discourse (close and open text) were analysed (Farah Benamara et.al, 2018) (M. Ayoub et.al, 2016). The various properties related to discourse of text have been explained along with development as present past and future in (Bonnie Webber et.al, 2012) (F.Cornish et.al, 2006). The method for detection of sarcasm which is based on various levels of analysing the text like, lexical, syntactic and semantic level is described in (Devamanyu Hazarika et.al, 2018). The pen tree bank is also been developed as wide-ranging corpus which uses discourse connectives are shown in (Miltakaki et.al, 2004). The next section explains detail methodology of the system that is designed.

Methodology: Identification of unseen text from given text is a crucial and complex task in Discourse analysis. It is related with social media insides as well as interaction between groups or two persons. Following are the steps that are performed for this task. For training and testing datasets from Kaggle and Keras library (IMDB ratings dataset) having 1,10,000 reviews is used. Figure 1 shows step-by step procedure.

Figure 1: Flowchart Designed system of Discourse processing using Neural Networks

Pre-Processing of the datasets: In this module, the dataset will be breaking down into separate sentences/ words from paragraphs in order to make programming a lot easier and this process is known as Tokenization. Cleaning is done through stop word removal. The words

like 'a', 'the' etc. are removed. Next, task is to identify part of speech to know a noun, a verb, adverb model, proper noun, an adjective and so on. This task is very important as this will convey the major information in the given sentence. In addition to that it informs the user its purpose also.

The working of the model is shown in figure -1 mentioned below where each step is elaborated.

Text Lemmatization and Removing Stop-words: This module will perform the lemmatization of each tokens in a sentence i.e. identifying the base form of each word/text in the given sentence or paragraph. Then it identifies the stop words from the datasets and filters out them as they don't add any meaning to the sentence. (e.g: 'a', 'the', 'and', etc.)

Parsing of datasets and NER: This module will identify relatedness of each word from the sentences. This is called dependency parsing. Then it will perform NER (Named Entity Recognition) that gives the entity information like: person information, industries, species, organizations, locations, quantities, Information of time, monetary values, percentages, etc.

Training and Testing of datasets: The IMDb (Internet Movie Database of data related to movies, television programs, home videos, video games etc.) dataset for movie reviews is used for this purpose. We have used 60 percent data for the training and 40 percent data for testing purpose. The system uses all total 1,10,000 reviews. The network is trained and various Activation function are used to govern the output of neural network like yes or no. It plots the resulting values in between 0 to 1 or -1 to 1 etc. We have used the ReLU function which is the most used activation function as in convolutional neural networks or deep learning. Using these functions model is designed which uses test dataset in order to test our model's accuracy.

RESULTS

The developed application is a prototype of Implementation of discourse processing for identifying hidden meaning. We have used 60 percent data for the training and 40 percent data for testing purpose. The system uses all total 1,10,000 reviews from IMDb. Sample output is shown below in figure 2. The system's efficiency is around 87% and error rate is 0.311. Referring Figure 2. First part, depicts for the given input data: as the movie started it was interesting, but after the interval movie was slow and it seemed to live theatre.

Output: Movie is below Average and story has no meaning and cast has no impact.

Referring Figure 2. Second part, Sample epochs are shown along with in between accuracies during runtime. The system is tested through human evaluation also. The systems output is verified against human experts. We have evaluated five hundred samples for this purpose.

And found that around 82 percent systems results are matching with human experts output results.

Figure 2: Sample output of Designed system

CONCLUSION

So, the developed application can be installed on any system which will meet the average hardware & software configuration. This application is a prototype of initial level of Natural Language Understanding, which is initially works on the movie reviews which can later be extended to common human conversation or chatting platform. In this application, the system will analysis the reviews and check for the semantic and syntactic analysis, then based on these analysis and training of these movie review datasets it will interpret how good or bad the movie is. Currently, our application can predict the movie review with that much accuracy as expected. The system can be applied for different twitter databases and review databases. Further efficiency also can be increased with different activation functions and techniques.

REFERENCES

- Ashraf Ullah, K. B. (September 2014) Mining opinion components from unstructured reviews: A review. Journal of King Saud University. Computer and Information Sciences, 26, Pg.258-275.
- Bonnie Webber, A. J. (2012) Discourse Structure and Computation: Past, Present and Future. Association for Computational Linguistics (ACL).
- David Nadeau, Satoshi Sekine, A survey of named entity recognition and classification, National Research Council Canada / New York University.
- Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, Rada Mihalcea (2018) CASCADE: Contextual Sarcasm Detection in Online Discussion Forums". 27th International Conference on Computational Linguistics. .
- F.Cornish (2006). "Understanding spoken discourse". Elsevier Ltd. Encyclopaedia of Language and Linguistics

(2nd edition) Pg.227-230.

Farah Benamara et al., (2018) Introduction to the Special Issue on Language in Social Media: Exploiting Discourse and Other Contextual Information. Special Issue on Language in Social Media. Volume 44 | Issue 4, Pg.663-681.

Forsythand et al., Lexical and Discourse Analysis of Online Chat Dialog. 2007, International Conference on Semantic Computing. Pg.1837–1848.

M. Ayoub et al., (2016) Textual Analysis of Print Media Text Discourse” New Media and Mass Communication, IISTE.

Miltsakaki et al., (2004). The Penn Discourse Treebank. Proceedings of the Association for Computational Linguistics (ACL).

Nianwen Xue, Qishen Su. (2016). Annotating the discourse and dialogue structure of SMS message

conversations. Proceedings of Association for Computational Linguistics (ACL).

Tatyana Makhlova, Dmitry Ilvovsky, Boris Galitsky. (2015). News clustering approach based on discourse text structure”. Proceedings of Association for Computational Linguistics and The Asian Federation of Natural Language Processing. Association for Computational Linguistics (ACL).

Xiyao Ma, Qile Zhu and Xiaolin Li, “Statistical learning for semantic parsing: A survey” -, National Science Foundation Center for Big Learning, University of Florida, Gainesville, FL 32608, USA.

Yizhong Wang et.al, Toward Fast and Accurate Neural Discourse Segmentation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (ACL).