International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

FEATURE EXTRACTION AND ANALYSIS OF NATURAL LANGUAGE PROCESSING FOR DEEP LEARNING ENGLISH LANGUAGE

Ravuri Seshu (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

NLP (Natural Language Processing) is a technology that enables computers to understand human languages. Deep-level grammatical and semantic analysis usually uses words as the basic unit, and word segmentation is usually the primary task of NLP. In order to solve the practical problem of huge structural differences between different data modalities in a multi-modal environment and traditional machine learning methods cannot be directly applied, this paper introduces the feature extraction method of deep learning and applies the ideas of deep learning to multi-modal feature extraction. This paper proposes a multi-modal neural network. For each mode, there is a multilayer sub-neural network with an independent structure corresponding to it. It is used to convert the features in different modes to the same-modal features. In terms of word segmentation processing, in view of the problems that existing word segmentation methods can hardly guarantee long-term dependency of text semantics and long training prediction time, a hybrid network English word segmentation processing method is proposed. This method applies BI-GRU (Bidirectional Gated Recurrent Unit) to English word segmentation, and uses the CRF (Conditional Random Field) model to annotate sentences in sequence, effectively solving the long-distance dependency of text semantics, shortening network training and predicted time. Experiments show that the processing effect of this method on word segmentation is similar to that of BI-LSTM-CRF (Bidirectional- Long Short Term Memory-Conditional Random Field) model, but the average predicted processing speed is 1.94 times that of BI-LSTM-CRF, effectively improving the efficiency of word segmentation processing.

1. INTRODUCTION

With the rapid development of Internet information technology and the continuous advancement of science and technology, a large amount of data of various types and structures have been accumulated in the real life and scientific research fields. In the real world, for the observation target of the same semantic conceptual ontology, multiple observation methods can often be used to obtain data information from multiple different observation channels, and these data from different information channels describe the same concept. Each of these kinds of information data can be called a different modal, or different observation perspectives. Different information

modalities together constitute multi-modal data for the same problem. NLP (Natural Language Processing) is one of the key technologies for realizing human-computer interaction and artificial intelligence [1]–[3]. It is listed as the three major elements of artificial intelligence research with voice processing and image processing [4], [5]. In the early days of NLP research, the main focus was on the analysis of language structure, technology-driven machine translation, and language recognition [6]–[9]. The current focus is on how NLP can be used in the real world. The corresponding research areas include dialogue systems and social media data. However, the training of deep frames is a difficult task, and traditional shallow proven methods that have proven effective cannot be transplanted into deep learning to ensure their effectiveness [10]–[13]. Another realistic problem is that there is no necessary connection between increasing the layer structure and obtaining better feature representations. For example, in a neural network, the more hidden layers, the less impact the first layer in the backpropagation algorithm. When using the gradient descent algorithm, it will also fall into the local optimum and lose the effect of continued transmission.

Related scholars have proposed a word segmentation algorithm based on supervised machine learning. This method implements a word-based word segmentation system. The main innovation is to use the maximum entropy model as a tokenizer to automatically label characters. This method has the highest recall rate of 72.9% in the AS2003 closed test experiment [14]–[17]. In the method of English word segmentation based on the dictionary and rules, it mainly focuses on the word segmentation algorithm and dictionary structure [18], [19]. The advantages of dictionary-based and rule-based methods are simple, easy to implement, and suitable dictionaries can be formulated according to special scenarios. In addition, in systems that require real-time performance, dictionary-based and rule-based methods are often more suitable because of their high efficiency [20], [21]. The disadvantages are: there is a problem of word segmentation ambiguity; there is no universal standard for word division, so the quality of the dictionary cannot be clearly defined. The dictionary has a great impact on the segmentation result [22]–[25]. With the advent of the era of big data, data has become more and more in natural language processing problems [26], [27]. Improving these labeling problems to support parallel computing and being able to perform parallel learning on large-scale training data has also become a research hotspot [28]–[32]. Parallel learning is currently supported, including maximum entropy models and conditional random field models. Some researchers have proposed the technical route of "understand first and then segmentation" [33]–[35]. The idea of understanding the segmentation first is to solve the lack of global information in the traditional matching segmentation, while the statistical method lacks the structural information of the sentence [36]–[39]. Relevant scholars use deep learning to perform sequence labeling in the NLP field [40], [41]. It can also add a sequence labeling model to combine with the output of the previous neural network to extract the best labeling sequence through the Viterbi algorithm. Related scholars have proposed an open domain question answering system based on relationship matching [42], [43]. The problem analysis problem based on relation matching is

solved through the associated data in the question answering system. The fragments in the question match the binary relationship in the triples and are automatically collected using the relational text pattern. Existing models do not take into account the importance of different modalities for the current learning task, but only focus on how to effectively use multiple modalities for feature extraction at the same time. Moreover, the selection of modals and the filtering of harmful modals are not involved, and this issue is also an important issue addressed in this paper. In terms of word segmentation processing, in view of the problems that existing word segmentation methods can hardly guarantee long-term dependency of text semantics and long training prediction time, a hybrid network English word segmentation processing method is proposed. Experimental results show that this method improves the efficiency of natural language processing.

In terms of English word segmentation, since traditional machine learning methods cannot solve the long-distance dependencies of texts, it is difficult to analyze the information contained in the problem as a whole and grasp the user's true intention. In order to solve the above problems and save the relevance of the forward and reverse information of the text, this paper uses BI-GRU (Bidirectional Gated Recurrent Unit) neural network and combines the CRF (Conditional Random Field) model to solve the problem of sequence labeling at the sentence level analysis, based on BI-GRU-CRF (Bidirectional-Gated Recurrent Unit- Conditional Random Field) hybrid network English word segmentation processing method. Specifically, the technical contributions of this article are summarized as follows:

Firstly, a multimodal fusion feature extraction method is proposed. The problem of heterogeneity of multi-modal data is solved through the feature transformation of deep neural networks.

Secondly, in view of the problems that traditional neural network models cannot capture the long-distance dependencies of text and the long cost of training and prediction of LSTM (Long Short Term Memory) neural networks, a word segmentation processing method based on BI-GRU-CRF hybrid network is proposed.

Thirdly, the proposed method is tested from two aspects, accuracy and timeliness. According to these two sets of experiments, the proposed hybrid network word segmentation processing method has good performance in English word segmentation processing.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

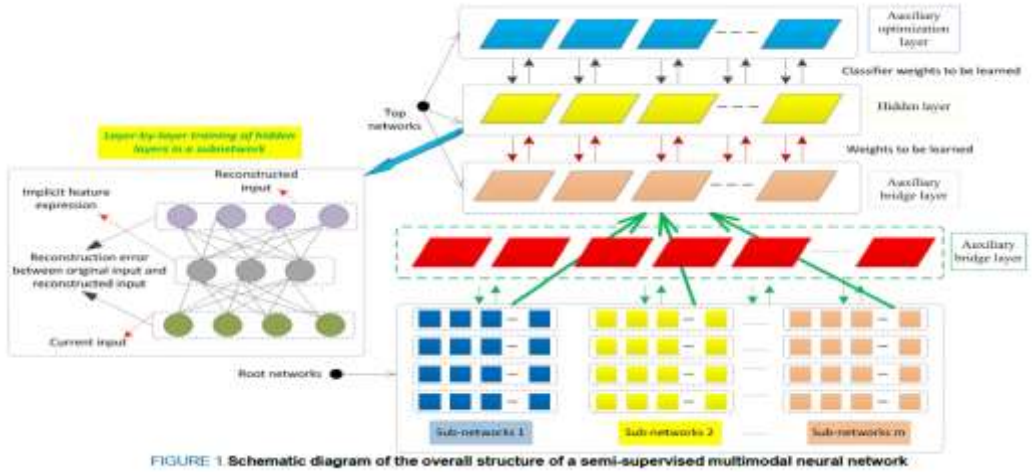


FIGURE 1. **Schematic diagram of the overall structure of a semi-supervised multimodal neural network**

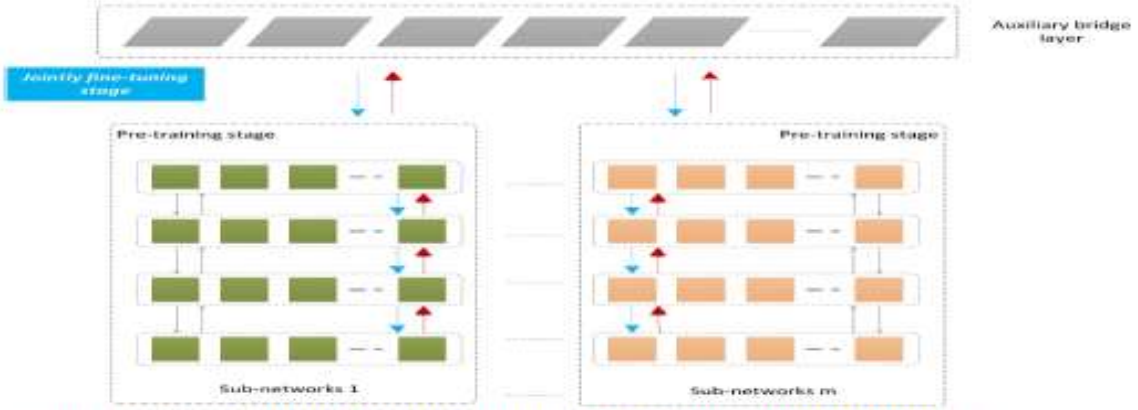


FIGURE 2. **Schematic diagram of the root network structure**

3. SYSTEM DESIGN

Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language

In this paper author is using Natural Language Processing with deep learning to detect English words segmentation and then evaluating performance of two deep learning neural networks called BI-GRU (Bidirectional Gated Recurrent UNIT) and BI-LSTM (Bidirectional Long Short Term Memory) and from both algorithm BI-GRU is taking less execution and giving less LOSS compare to BI-LSTM. Neural network model which give less LOSS can be consider as best model.

Word segmentation is identifying meaningful information form give data for example if we got data as ‘commentsunderquestioning’ then segmented output will be ‘comments under questioning’ and to get this output using neural network we will train neural network with all possible words and their ID’s and whenever we gave such input then neural network will predict

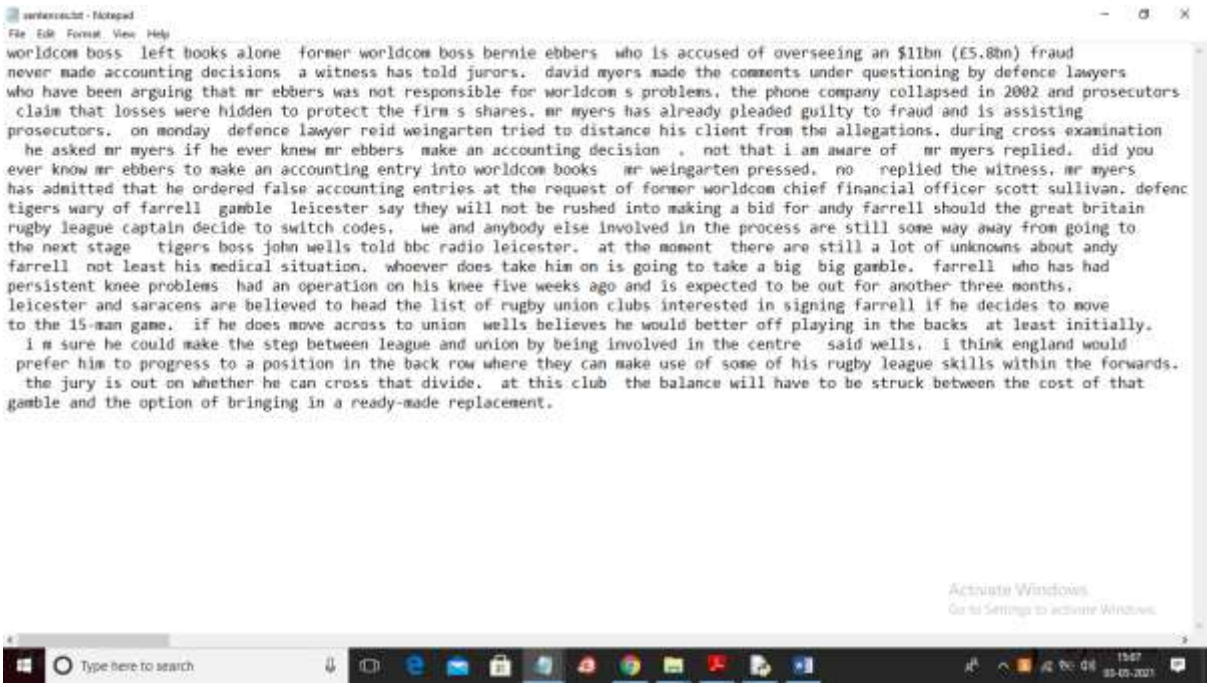
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

ID's of that word and then convert that ID into word and then look that word in vocabulary and if word found in vocabulary then segmented word will be identified.

About algorithms and other details you can read from paper and to implement this paper author has used WIKI dataset and this dataset contains lots of sentences and to train all those sentences may take days of time so I took few sentences which consists of 3000 words and then train both LSTM and GRU and then calculate LOSS of each model.

Reading all words from dataset and then preprocessing them is called as Features Extraction and then converting this features into vector is called as Natural Language Processing (NLP). Generating vector from features is referring as assigning unique ID to each word.

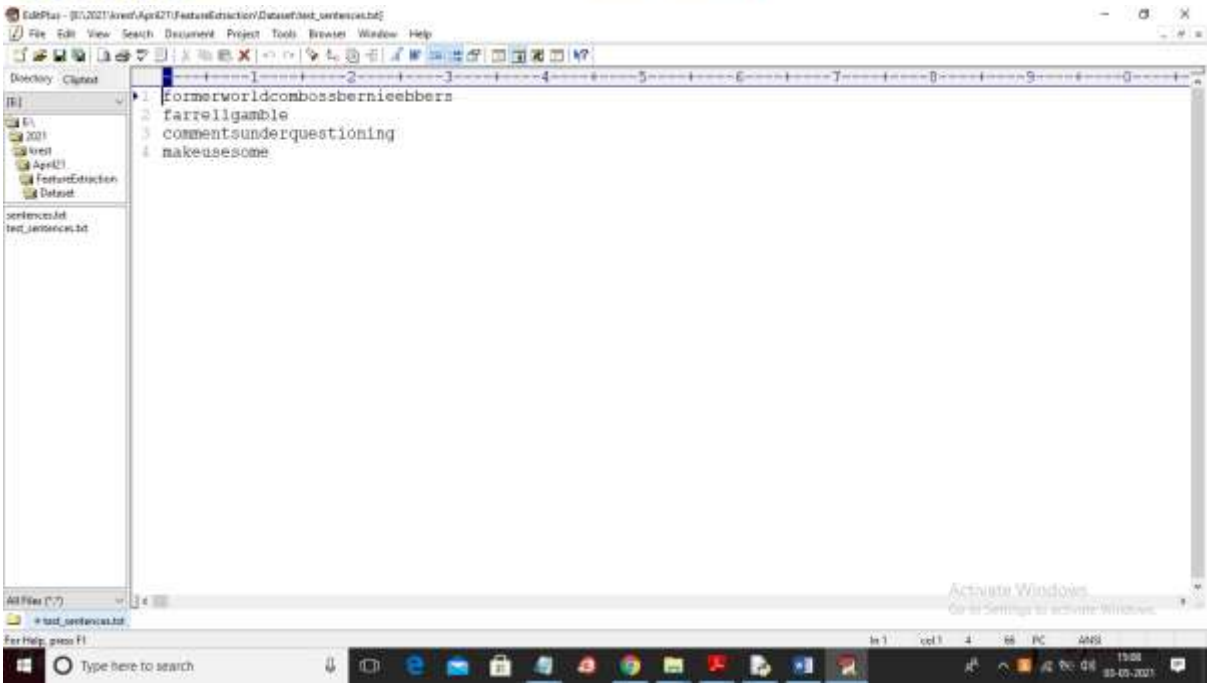
Below screen shots showing dataset sentences used to train above algorithms



We are using above sentences to train models and you can give any word from above sentences to get segmented output.

Below is the test words used to get segmented output

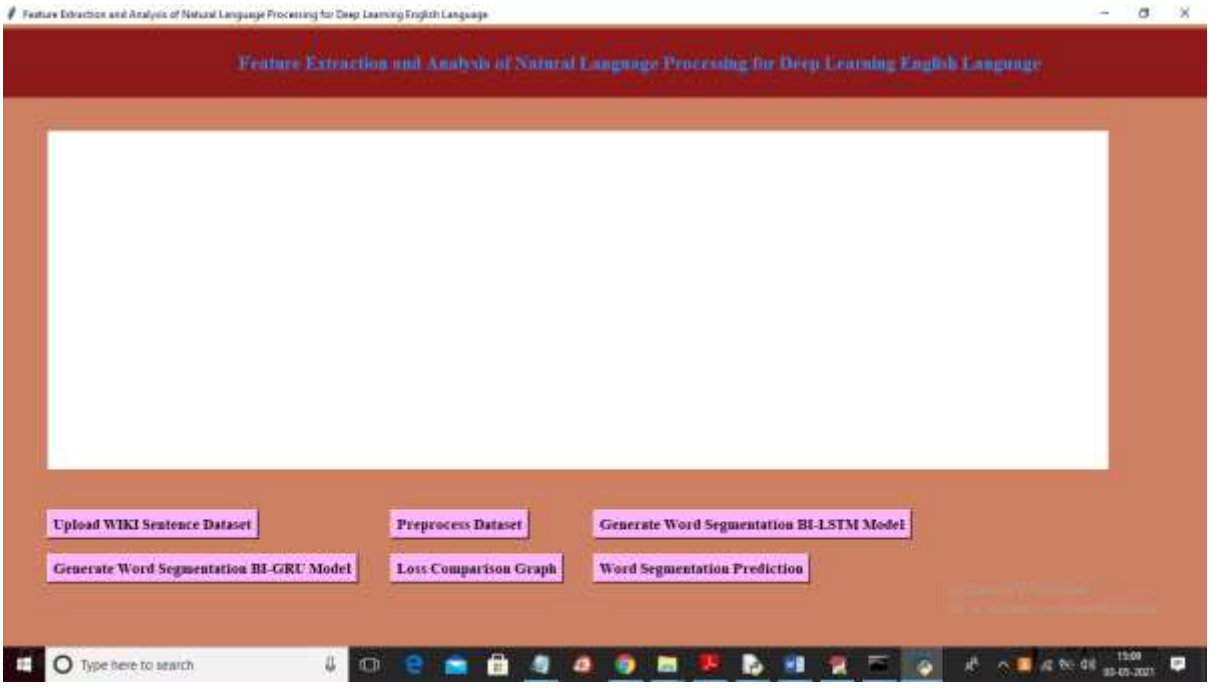
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362



In above test data we cannot get any meaningful information so by applying GRU or LSTM we can get segmented words from above data.

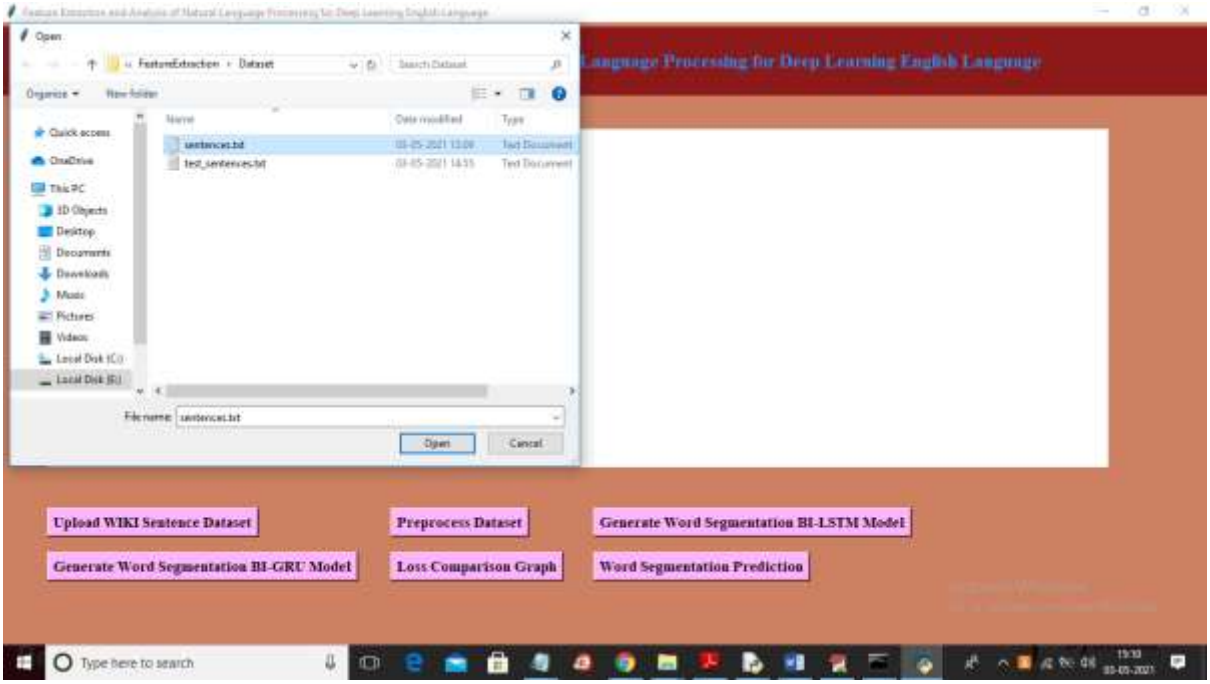
SCREEN SHOTS

To run project double click on 'run.bat' file to get below screen

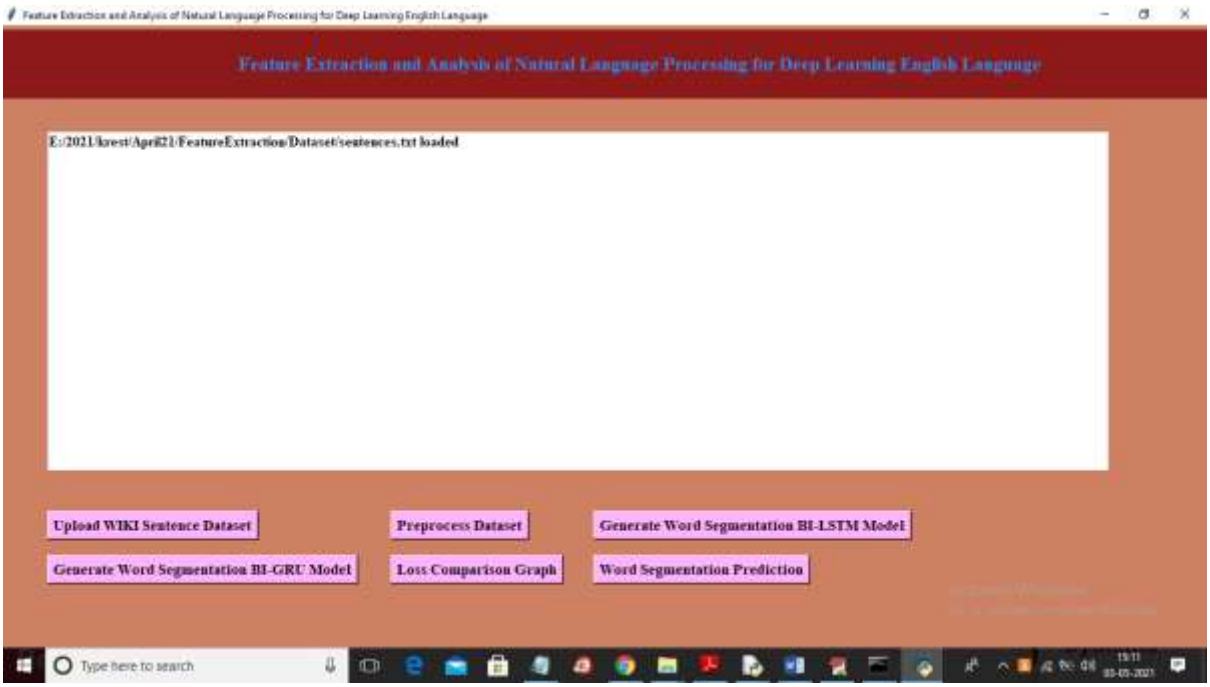


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

In above screen click on 'Upload WIKI Sentence Dataset' button to upload sentences dataset

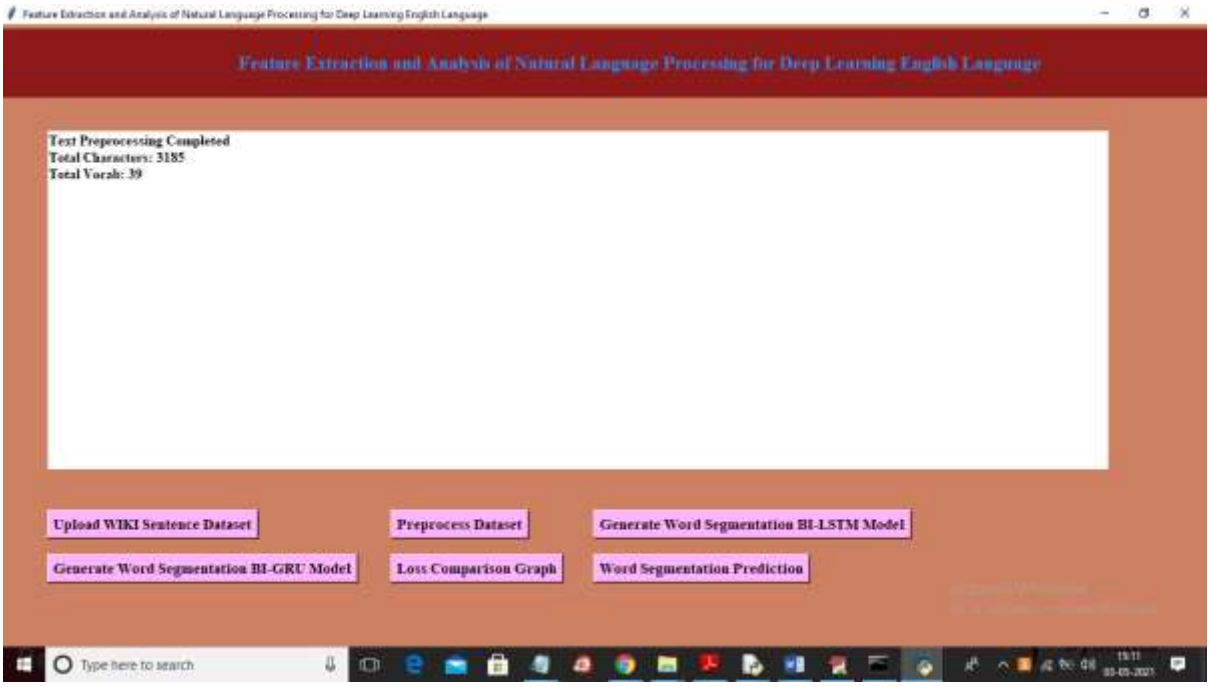


In above screen selecting and uploading 'sentences.txt' file and then click on 'Open' button to load dataset and to get below screen

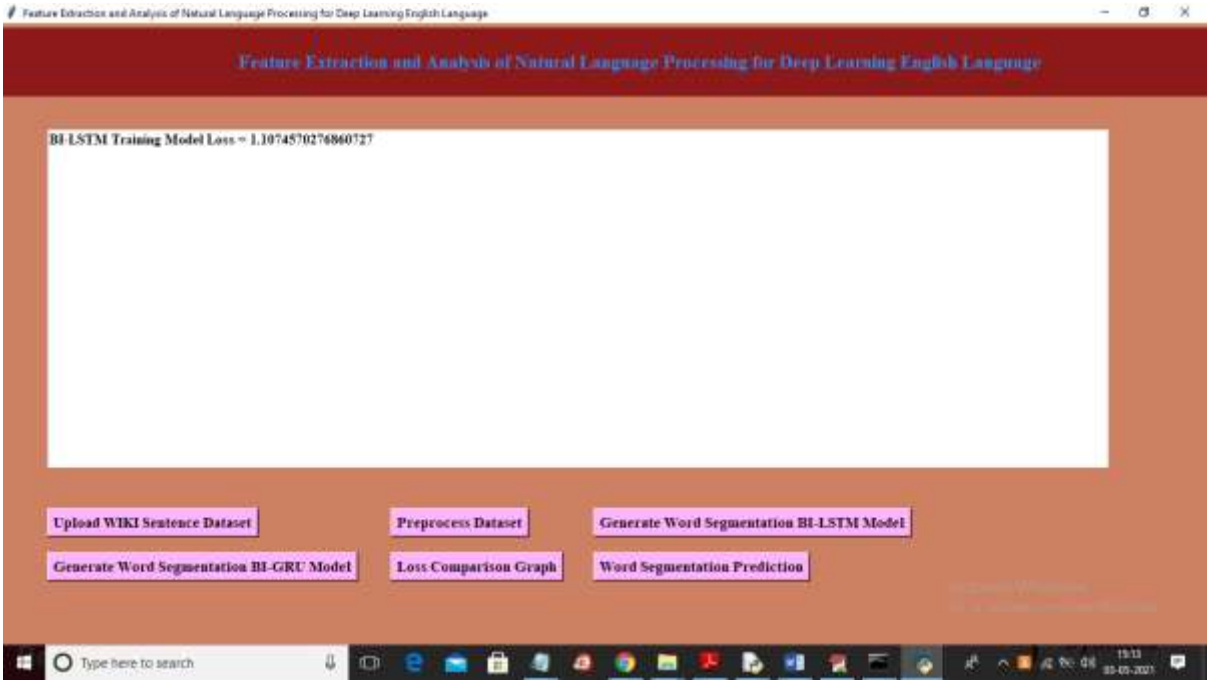


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

In above screen dataset loaded and now click on 'Preprocess Dataset' button to read and process dataset such as features extraction and generating vector

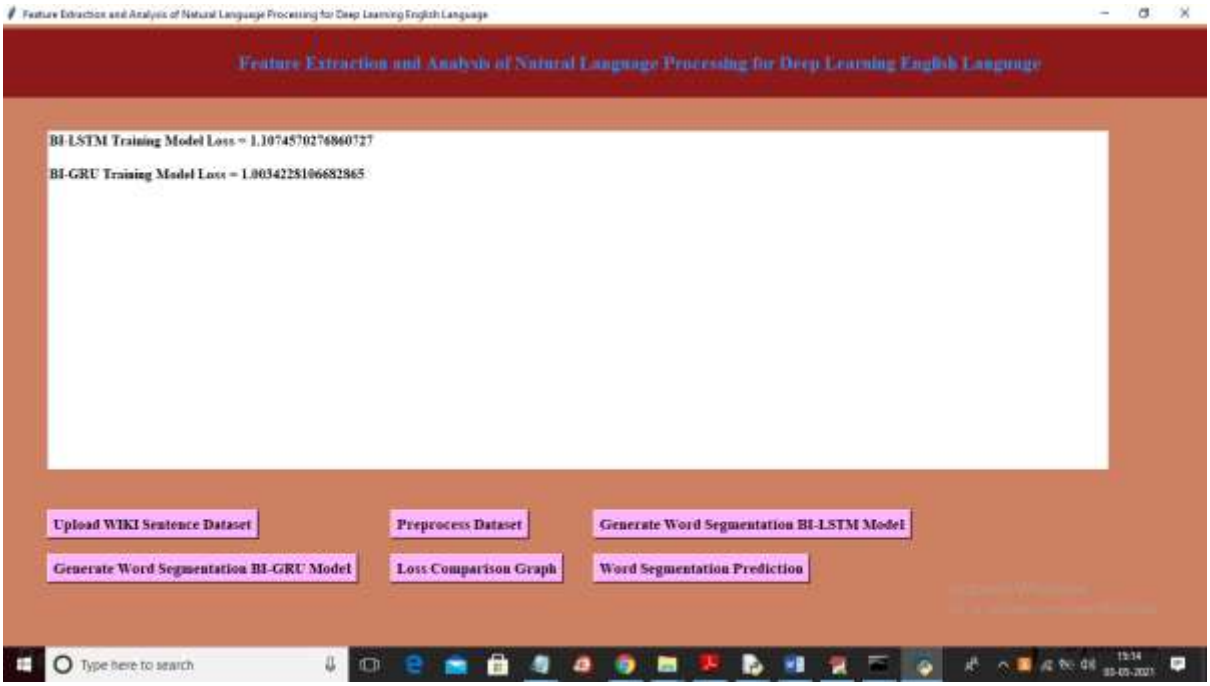


In above screen total features or characters extracted from dataset is 3185 and the vocabulary or total unique words found in dataset is 39. Now click on 'Generate Word Segmentation BI-LSTM Model' button to build LSTM model on above dataset and then calculate loss

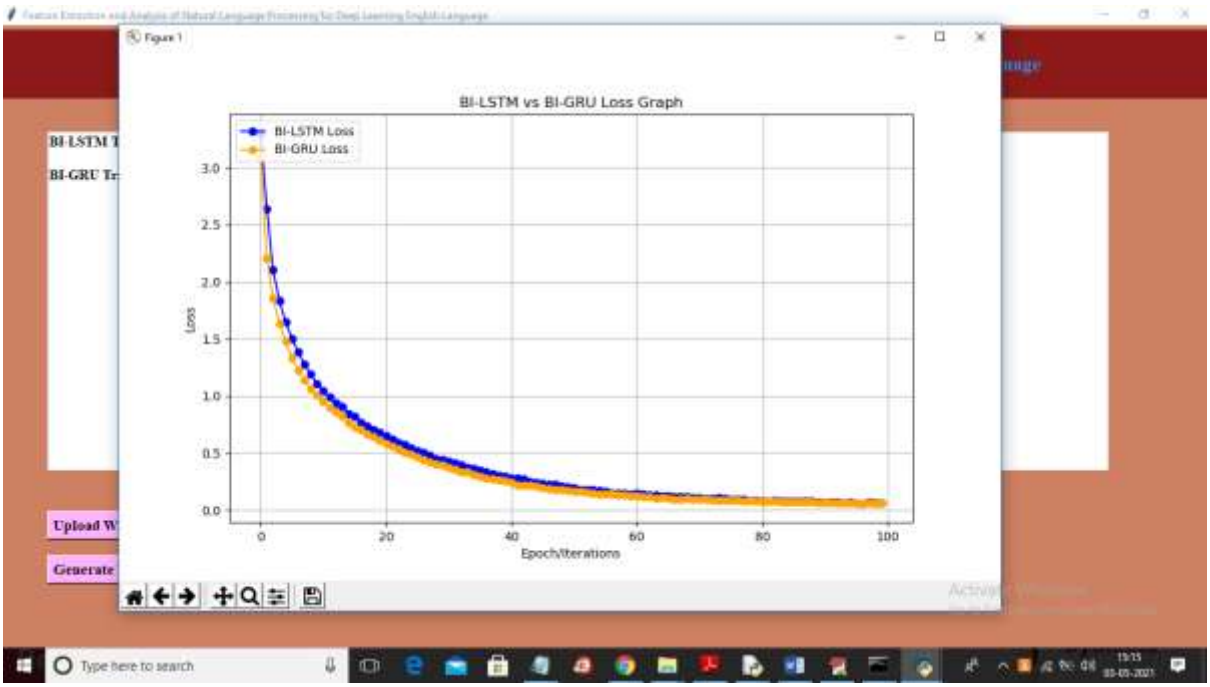


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

In above screen out of 100% LSTM loss reduce to 1.10% and now click on 'Generate Word Segmentation BI-GRU Model' button to build GRU model

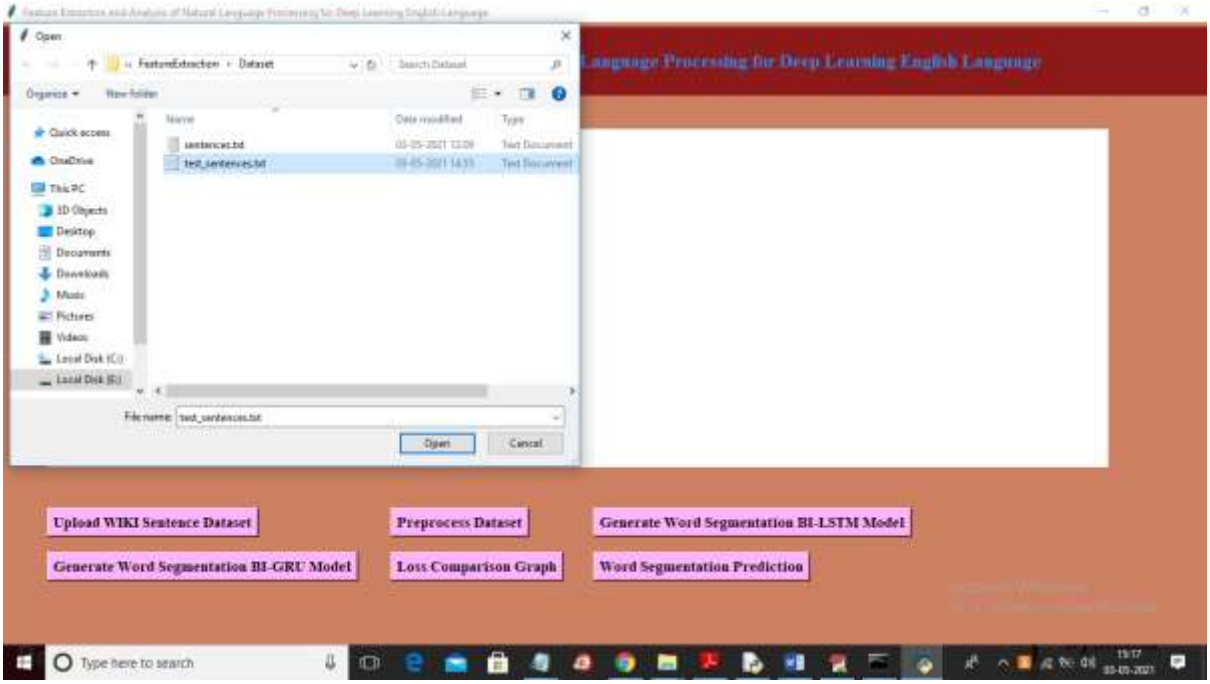


In above screen GRU loss reduce to 1.00% so GRU is better than LSTM and now click on 'Loss Comparison Graph' to get below graph of both LSTM and GRU loss

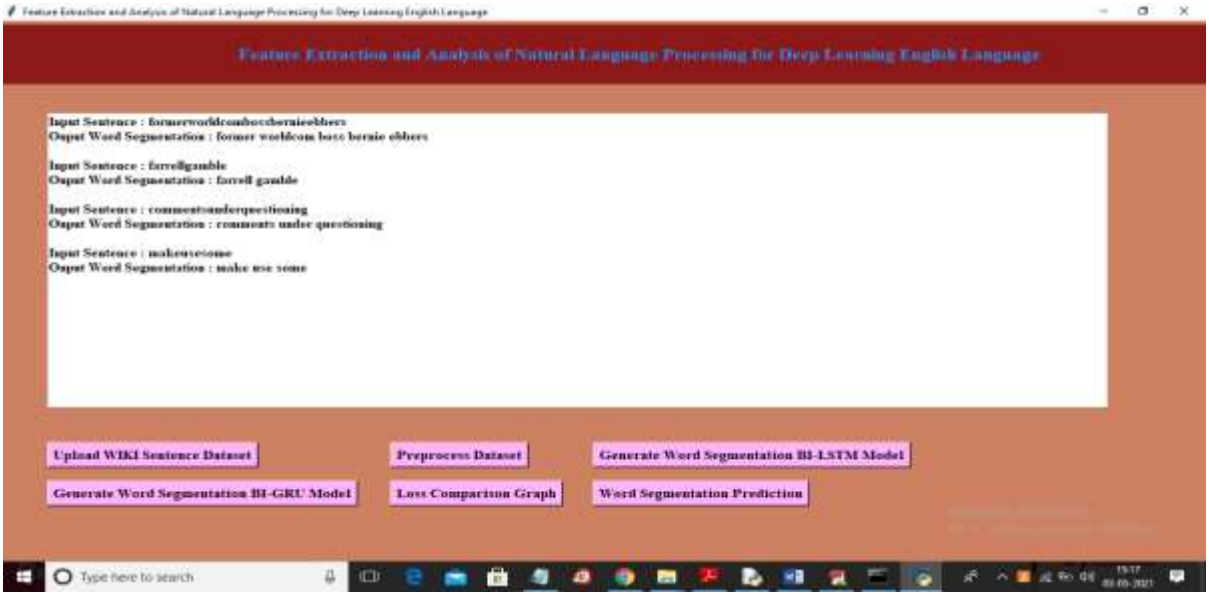


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

In above graph x-axis represents Epoch/Iterations and y-axis represents LOSS and blue line represents LSTM and orange line represents GRU and for both algorithms we too 100 EPCOH and both algorithms loss is reduce per increasing EPCOH but GRU loss is little less compare to LSTM so GRU is better and now click on 'Word Segmentation Prediction' button to upload test file and then will get segmented word



In above screen selecting and uploading 'test_sentences.txt' and then click on 'Open' button to get below result



In above screen we can see Input Sentence and then predicted segmented output from GRU and from above result we can see we extracted meaningful information from give data.

Note: I used few sentences to train both algorithms so application will perform segmentation on dataset words only. Actually to train large dataset application taking hours of time

4. CONCLUSION

This paper proposes a multimodal shared feature expression extraction algorithm based on deep neural network, gives the entire model structure of the algorithm, and details the design of the model structure and the model training method. In order to verify the effectiveness of the proposed model, a series of comparative experiments were carried out. The experimental results show that the proposed multimodal fusion feature extraction model can effectively extract low-dimensional fusion features from the original multiple high-dimensional data. The obtained fusion feature expression has a strong discriminative ability while possessing a lower feature dimension, thereby proving the validity of the proposed model. In terms of English word segmentation, this article has studied LSTM and GRU in depth. After analysis and research, both networks can solve the problem of traditional word segmentation in the long-range dependency relationship of text. However, due to the complexity of its structure, LSTM consumes a lot of time in the process of training and predicting the data set. The GRU is a simplified version of the LSTM. It has a simple structure and consumes less time in training and prediction. Based on the two-way network's ability to better capture the contextual relationship between semantics, this paper combines BI-GRU and CRF models, and proposes a hybrid neural network word segmentation processing method. The experimental results show that the model proposed in this paper is better than most previous models in terms of accuracy, and in terms of timeliness, the method proposed in this paper is 1.62 times faster than the BI-LSTM-CRF network word segmentation method in training speed. The average speed is 1.94 times that of the word segmentation method based on BI-LSTM-CRF network. Based on these two sets of data, the hybrid network word segmentation method proposed in this paper has good performance in English word segmentation. In future work, we can consider analyzing the impact of different feature extraction methods and feature selection methods on the model, thereby further enhancing the learning ability of the model. The proposed method treats different features obtained by different extraction methods in each kind of raw data as an independent mode, and does not learn directly on the raw data. How to get the multi-modal fusion low-dimensional features directly from the original multi-modal data needs further research.

5. REFERENCES

1.
V. K. Ha, J.-C. Ren and X.-Y. Xu, "Deep learning based single image super-resolution: A survey", *Int. J. Autom. Comput.*, vol. 16, pp. 413-426, 2019.
Show in Context CrossRef Google Scholar

2.
F. Meng, P. Chen, L. Wu and X. Wang, "Automatic modulation classification: A deep learning enabled approach", *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10760-10772, Nov. 2018.
Show in Context View Article
Google Scholar

3.
Q. Xia, S. Li, A. M. Hao and Q. P. Zhao, "Deep learning for digital geometry processing and analysis: A review", *J. Comput. Res. Develop.*, vol. 56, no. 1, pp. 155-182, 2019.
Show in Context Google Scholar

4.
Y. Chen, Y. Zhang, S. Maharjan, M. Alam and T. Wu, "Deep learning for secure mobile edge computing in cyber-physical transportation systems", *IEEE Netw.*, vol. 33, no. 4, pp. 36-41, Jul. 2019.
Show in Context View Article
Google Scholar

5.
K. A. Weber, A. C. Smith, M. Wasielewski, K. Eghtesad, P. A. Upadhyayula, M. Wintermark, et al., "Deep learning convolutional neural networks for the automatic quantification of muscle fat infiltration following whiplash injury", *Sci. Rep.*, vol. 9, no. 1, pp. 7973, May 2019.
Show in Context CrossRef Google Scholar

6.
O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved", *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514-2525, Nov. 2018.
Show in Context View Article
Google Scholar

7.
M. Abdughani, J. Ren, L. Wu, J.-M. Yang and J. Zhao, "Supervised deep learning in high energy phenomenology: A mini review", *Commun. Theor. Phys.*, vol. 71, no. 8, pp. 955, Aug. 2019.
Show in Context CrossRef Google Scholar

8.
R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.-C. Chen, V. M. Patel, et al., "Deep learning for understanding faces: Machines may be just as good or better than humans", *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 66-83, Jan. 2018.
Show in Context View Article
Google Scholar

9.
Y. Tian and X. Liu, "A deep adaptive learning method for rolling bearing fault diagnosis using immunity", *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 750-762, Dec. 2019.
Show in Context View Article
Google Scholar

10.
C. Wu, R. Zeng, J. Pan, C. C. L. Wang and Y.-J. Liu, "Plant phenotyping by deep-learning-based planner for multi-robots", *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3113-3120, Oct. 2019.
Show in Context View Article
Google Scholar