Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021), 2577– 2585

Research Article

Survey on Text Transformation using Bi-LSTM in Natural Language Processing with Text Data

Preethi V[a], and Usha Kiruthika[b]

[a] Research Scholar, SRM Institute of Science and Technology,Kattankulathur

[b]Assistant Professor, SRM Institute of Science and Technology,Kattankulathur

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Communication using text can be understood by human beings as language plays a vital role in their life. Computers need to understand such human readable languages in order to explore documents in a wider manner. For such cases, machine learning algorithms as well as natural language processing algorithms can be used. Computers understand the human readable language and convert them as machine readable form. Replication of data stands as a most challenging problem in document exploration as the document may contain many repetitive words. The system makes use of classifier named LSTM to remove the words appearing frequently in the entire document. Firstly the text is transformed using NLP techniques followed by feature extraction and finally Bi-LSTM is used to classify the text data. LSTM mainly focus on reduction of repetitive words from the entire document and also maintains the integrity of the document. Bi-Lstm is an effective method to connect two different independent RNN together that helps to follow the sequence of statement front and back path every now and then.Multi directional ways of running the code will help in future prediction by saving the sentence for future purpose also.When the independent hidden states are joined it will be preserved for future and also past search. Although the word count is reduced, content of the word document is maintained in the LSTM method which stands as a major advantage. The experimental results showcase the efficacy of the system.

Keywords: Feature Extraction, Long Term Short Memory, Tokenization, Stop Words

1. Introduction

Smart devices such as mobile devices and computers, laptops are dealing a massive amount of data that are produced as structured information which are gathered from spreadsheets and database such as object data, relational database, etc. Human languages are via communication as language in the form of words and they are not framed as rows and columns such as tables. As the language varies for each people they are considered as unstructured and semi structured data's that are not in format of extraction as well as framed as raw information compared to other people language. A language based domain that comes as sub-field of Artificial intelligence is Natural Language Processing. NLP is a version of communication of human languages as machine understands language. Also neural machine translation and questionnaire session using ML techniques follows NLP operation such as interruptions and tough operations that are focused. The packages used in NLP are shown as Table.1.

Table.1. NLP Packages

S.No	NLP Packages	Library usage	Real Time Applications
1.	SpaCy	Text preparion for deep learning. Linguistic for a real time NLP problems.	Uber, Quora, Retriever, STITCH FIX, Chartbeat, etc
2.	Textacy	Tokenization, POS, tagging, dependency and text parsing, etc	Wikipedia, Twitter, Whatsapp,
3.	Neuralcoref	Neural coreference and neural nets, accuracy and speed	Health care system, automatic car parking, inconvenience stores, driver less car testing etc

4.	Natural language toolkit(NLTK)	Linguistic and text processing, classification, tokenization, stemming	Twitter, Facebook, Large scale data interactions
5.	Gensim	unsupervised topic modeling, NLP, Machine Learning	DTU, tailwind, IBCN, EuDML, issuu, Roistr
6.	Polygot&TextBlob	Embedding2, ner2,sentiment,morph2, transliteration2 together for massive multilingual applications	Language translator, text editor etc
7.	CoreNLP	Human language varieties, Annotators/models;	Human language conversion applications ,social media

1.1 NLP on Language Processing

English language makes different meaning based on the sentence defines. more or less its very much difficult for a computer to understand human language what emotionally they thought about. Parsing is the technique that helps in text extraction from a structured data. Splitting the data called chunking data into many pieces and combining them through the pipelines makes difficulty for a machine to process the human language.

Document exploration is an important step in NLP as system takes several steps to transform the text followed by extracting the features and finally LSTM is applied to the document.

Natural language processing deals with words whereas system understand numbers, the current real time data such as time series data, GPS tracking direction predictions are all some more applications of NLP. Finding phishing mails and spelling checks are more to understand the NLP. As an example certain paragraph or meaningful sentence are taken, as input initial steps to be applied is classifying and analyzing the data. Social media includes various activities such as twitter, Instagram have similar activities to find with classifications. Even online shopping sites such asFlipkart or Amazon gives us similar data. Also tourism data are considered as raw text and discussed to categories which form it is involved to.

Consider a man has travelled all over countries starting from his native to England as his destination. Many different techniques such as support vector machine, Naive Bayes gives better results and perform well. We identify the classes as below

a) Location
b) Expenses
c) Booking history
d) Food
e) Packages

1.2. Collecting data:

Data collected from as many as passenger with his meta data, Data set created with a piece of real time travel details are as follows,

Table 2.Data Collection

S.No	Location	Expenses	Booking history	Food	Packages
1	Delhi	60590	City	North	12
2	India	300075	State	Indian	14
3	London	867555	Region	British food	18
4	England	129875	Country	England food	24

The mandatory information in data is Location and booking history, in the available column several steps takes place is preprocessing.

The contributions of the paper are as follows:

1. Text transformation using NLP intends to identify the POS (Parts of Speech) from the document.
2. Extracting the features is mainly used to identify the count of frequent terms in the document.
3. Classifiers utilized for the document in the existing system is Bi-LSTM (Bi-directional Long Short Term Memory)

Section2 reviews the literature related to NLP and Section 3 discusses the NLP system. Section 4 presents the results obtained and Section 4 concludes the paper.

2. Literature Review:

CNN coined as Convolution neural network can be interpreted and is termed as modal interpretability in the field of machine learning. The main use of the above method is to capture the features effectively such as identification of patterns [1]. Visualization of feature stands as one of the important methods along with another method that is mainly oriented in contributing to input to the label predicted. Local model interpretability (LIME) [2] helps by predicting at samples of data levels and helps users to make decisions based on the predictions they make. The main focus of LIME is to change only one data sample and then observe the changes made in output. The output of LIME actually is the explanation list which is obtained from the role of input sample in data sample prediction. Though LIME has many advantages, it does not reveal the internal state and working of the model and acts as a black box model [3] limiting the access of to the applications. In order to overcome such disadvantages, existing method allows accessing the model internally as well as manipulating it.

In case of interpreting neural network model, visualization [4-6] has been used to identify relationships that are complex. Studies based on visualization have increased nowadays as an important research focus area that is offered to neural networks [7]. Errors in CNN are identified as well as the hierarchy in the class is learned [8]. Process of training in CNN is analyzed that helps in providing guidelines to design the architecture of deep neural networks [9]. The system introduced also helped in to tune Facebook neural networks [10]. TensorFlow focused on representing graphs related to computation in deep neural networks visually in hierarchical fashion [11]. Tools used in visualization also carried out the process of training in the models of neural networks [12].

Fewer studies were focused on NLM (Neural Language Processing Models) [13]. Advantages of visualization helped in models of NLP as well as for design of several applications in the future [14]. Significant patterns in the trained data were also identified in recurrent a network that works character-level. Also linguistic models were also analyzed [15][20]. Hidden state units have been found from the inputs given as text and the responses are found. With the help of information related to gradient also, visualization was performed and trained the data as well [16]. Also hidden states were represented in varying time slots also in RNN. Relationships based on semantics shall also be improved through projection in a linear manner in the visualization where embedding of the words takes place [17-19].

The existing system mainly focuses on the trained data as a static object. Our system uses

Bi-LSTM for interacting with the user where the environment is dynamic which leads to tremendous analysis.

3. Current NLP System

The system first performs the process of pre-processing of input data. Preprocessing is further divided into the following steps namely,

a) Upper case to lower case

b) Number values and punctuations removal

c) Unwanted extra spaces removal

d) Tokenization converted to words

e) Repeated words and end words removal

f) Lemmatization

i) Even though SVM is supervised learning model is applying text processing the RNN training model takes the steps by conversion of lower case, duplication of words,stop words. This step helps in text feature extraction simple and numbers of counts in words are also reduced.

ii) To standardize the text dependency the punctuation repetition is removed.

iii) Stop words such as'the','me','is','was','he','she','it' etc are process one at time every preprocessed method. For various languages all stop words might be processed and tagged in speech also similarly frequent words are removed.

iv) Either a phrase or paragraph or sentence or the whole document are divided into small units with individual words those are tokenization.

v)Similar like stemming transformed words are considered and those words are not used after their process such as repeated words.

3.1 Language Translation Using Bi-LSTM

Long short term memory (LSTM) is a recurrent neural network technique which is available in deep learning field. LSTM is a prediction algorithm which helps to find solution for complex problem such as texture translation, speech recognition language translation etc. there are sequence classification problems such as bidirectional LSTM where the process will be slow in single LSTM model. In first LSTM input is tokenizing the sequence and second RNN-LSTM. The second stage is to be system to store data as time of arbitrary label and to

reduce the noise and irrelevant data is removed. The estimated output is reasonable time by training the neural network.

Even though there are many prediction analysis techniques applicable in neural network data such as time series will be efficiently analyzed and performed using a long short term memory network. Data set considered here is travel information. Main purpose of using LSTM is, it is familiar in use of Recurrent neural network where like an image or other data can't be predicted easily when it is fetched as paragraph in any place. Like back propagations failure in updating the weight and shrinking of completed information time we are dealing with LSTM.

Original weight = Current weight - Learning rate * gradient

The above rule was applied in gradient descent where those layers might forget the lengthy sequence of data with its short memory.As depending on the data set the travel data need not be same distance and place so the prediction is depending on the features of the current table data such as the destination, source and distance of travelers. Also the expenses depend on the feature of previous travel location. When recurrent neural network is considered, fitting the model with the last dataset available will not be same as data are categorical.

3.2 A Comparison of LSTM in RNN short term dependencies

To understand the sentence using LSTM Techniques we consider the below statement

The old man reached India last_____

Above information is not easy to predict the context of the next statement

The general prediction is

The old man reached India last night.

Vanila RNN Or Peehole connections are not exactly suitable to predict the correct content. One of the layers where weight is updated along with the several multiple time of learning rate and the error rate is identified with the previous input of the layer. LSTM is applied to solve all over problems with the term called cell stated where the information passes to select the prediction data or forget things.

3.3 Long Short Term Memory Architecture

As LSTM is through out with travel dataset the attributes such as expenses, location, booking history, food and packages are considered. All three steps are utilized to predict the function and visualization. Let us assume that, a person location changes when he travel from one place to other booking history also gets updated, when immediately the last visited location is forgotten then the travel person will not be able to find the history of travel.

a) Remembrance State:

Input is given entirely new that also will cause to many steps, so all split data might be not sufficient to predict the location where he is going to visit.To summarize the overall information and produce the next location as output is task.

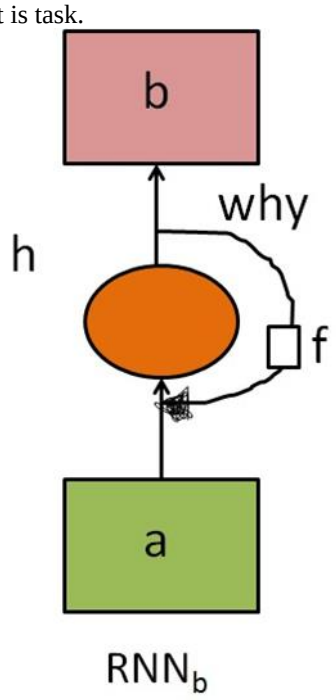

Figure 1. Remembrance State

b) Hidden State (forget) cell state:

Let's take an example statement,

Raj travels to England. Ram receives him on the airport.

Once full stop is detected or identified the forget get found that context may change from the next level of statement.

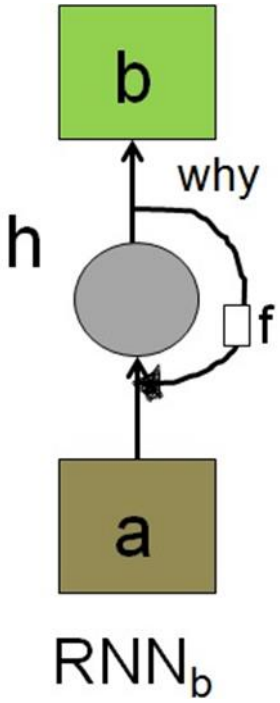

Figure 2. Hidden State

c) Current State (Output):

Output gate is the main step that helps to create a vector that can be applied on cell states. Cell state uses various activation functions such as tanh. When there is a proper noun adjective and full stop in the input, the filter applied here is tanh that has more improvement that sigmoid functions show the input dimension that matches based on the regularity of filter.

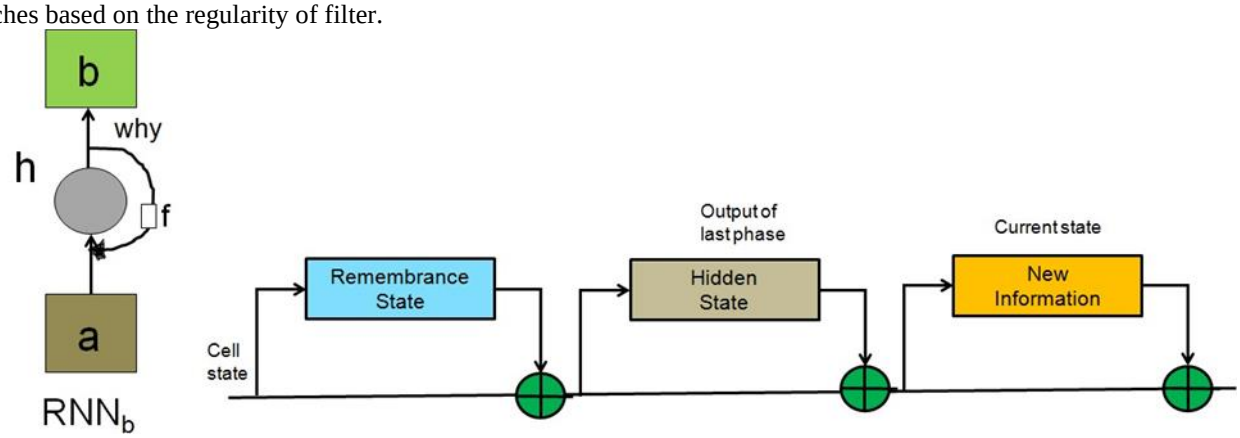

Figure 3. Bi-LSTM Architecture with RNN

The sample text when applied Bi-LSTM classifier is also shown in Fig 4.

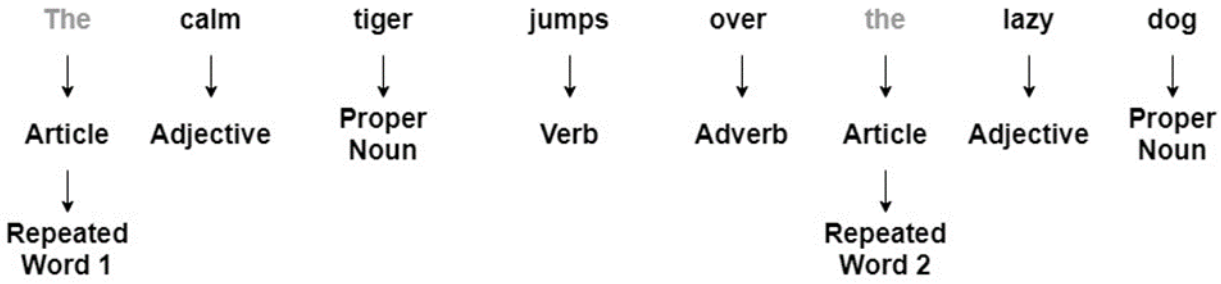

Figure.4. Bi-LSTM classifier applied text

4. Results and Discussion:

Jupyter notebook is used to implement the system which is a python IDE that plays a major role in visualization as well as analysis of data. Several features missed while the usage of python is being utilized in case of our implementation tool. The major advantage behind the usage is the efficacy that it offers in the workflow of programming.

The process starts as usual by training the data as well as their labels followed by representing data with term frequency-inverse document frequency that identifies the repetitive count of each word in the entire document given as input feed. Fig.8. shows the sample text along with its count being mapped correspondingly. Now the data is tested and trained followed by encoding using Multi label Binarizer for multiple label columns. After the process of encoding, the corresponding words in the text documents need to be turned as several vectors and the prediction process starts here.

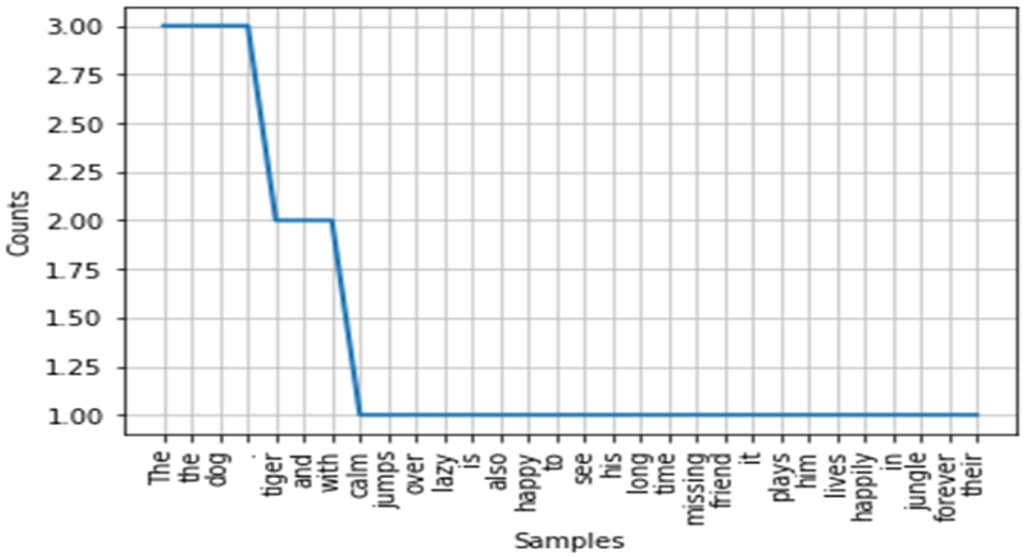

Figure.5 Sample Text Graph After identifying Count

LSTM plays an important role in prediction and method makes use of Bi-LSTM as a classifier for our input sequence. Fig.9 depicts the frequency distribution set by applying Bi-LSTM classifier to reduce the replication of data. The graph clearly depicts how the replication of data is handled in case of LSTM and enhances the efficiency of the system.

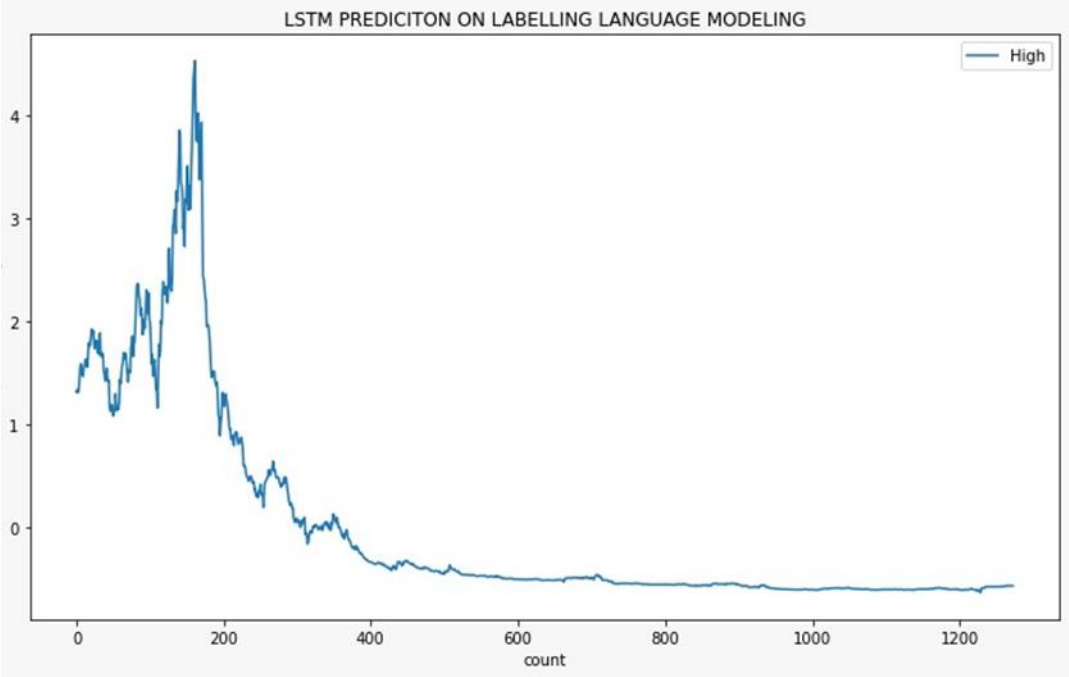

Figure 6. Frequency distribution set by applying Bi-LSTM classifier

4.1 Comparison with existing system

While text processing takes the general steps such as defining vectors from the corpus and computing word2vec by applying dimensionality reduction with the neighborhood training data sets the Bi-LSTM techniques as learning order of inputs from RNN helps to overcome the gradient descent facing problems in classification techniques. Also the predications of Sentence from the input are independent hidden states that have backward and forward cell states information.Bi-LSTM not only better remembrance but also connecting the backward cells by forward propagation multi times with its activation function.

Fig.7 shows the comparison of our work with several other existing system using various classifiers such as SVM and so on. Our performance keeps on increasing when the data set size increases.

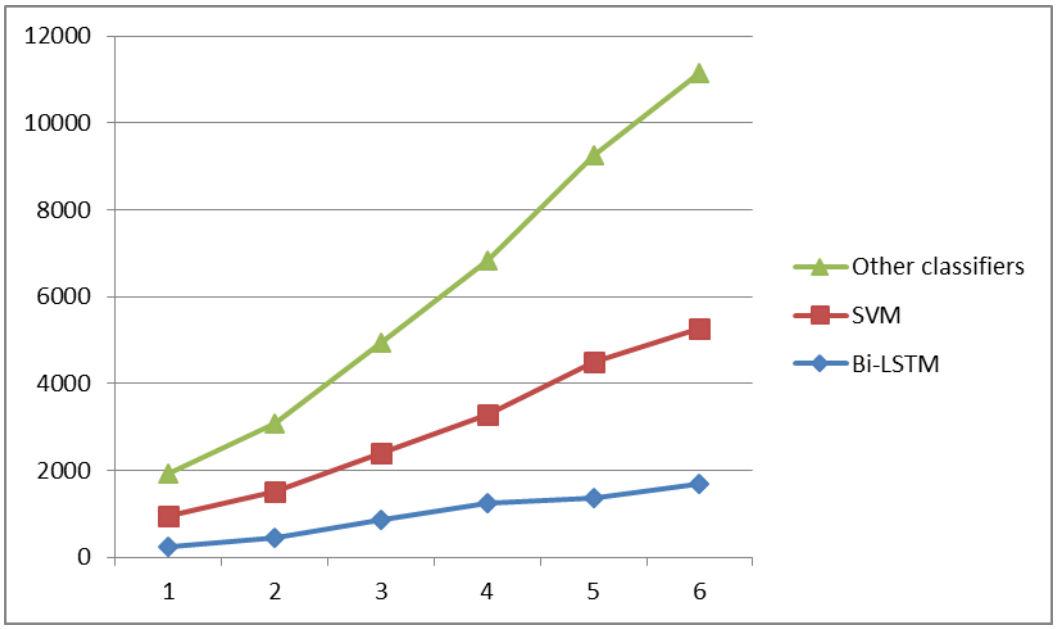

Figure 7. Comparison Chart

5. Conclusion

Automatic manipulation is another term coined for natural language processing including speech as well as text. The main purpose of natural language processing to convert the text data to machine readable form and in understandable form. When replication of data stands as a long term unsolvable problem in NLP, Bi-LSTM comes to our rescue as the classifier used mainly focus on vectorization of data thereby reducing the stop words in the document given as input. The output document after classification showcases the necessity the classifier in the field of natural language processing. In this model can also be used to benefit semi-supervised systems because it helps to carry huge information from a large text input than the individual data word. When Bi-LSTM is used for the travel dataset the advantage is to identify the future and past state of position without depending on the state

References

1. M.Abadi,P.Barham,J.Chen,Z.Chen,A.Davis,J.Dean,M.Devin,S.Ghemawat,G.Irving,M.Isard,etal.Tensorflow:A system for large-scale machine learning. In OSDI,vol.16, pp.265–283, 2016.
2. “Deep Learning in Natural Language Processing: A State-of-the-Art Survey”, Junyi Chai Beijing Normal University, Hong Kong Baptist University United International College,Division of Business and Management,Zhuhai,China; Anming Li, “ IEEE Xplore: 06 January 2020.
3. Natural language processing future,"2013 International Conference on Optical Imaging Sensor and Security (ICOSS), "Natural language processing future", IEEE EXplore, 06 January 2020.
4. An Improved LSTM Structure for Natural Language Processing, IEEE Xplore: 15 April 2019, Lirong Yao Qingdao No.2 Middle School, Qingdao, China; Yazhuo Guan, April 2019.
5. Raja, Dr S. Kanaga Suba, C. Viswatnathan, Dr D. Sivakumar, and M. Vivekanandan. "Secured Smart Home Energy System (SSHEMS) Using Raspberry Pi." Journal of Theoretical and Applied Information Technology 10: 305-314.
6. "Named entity recognition with bidirectional LSTM-CNNs", J.P. Chiu, E. Nichols, arXiv preprint arXiv: 1511.08308, 2015.
7. "A Domain Knowledge-Enhanced LSTM-CRF Model for Disease Named Entity Recognition",Yuan Ling, PhD, Sadid A. Hasan, PhD, OladimejiFarri, MBBS, PhD, Zheng Chen, MSc, Rob van Ommering, PhD, Charles Yee, PhD, and NevenkaDimitrova, PhD, AMIA Jt Summits TranslSci Proc. 2019; 2019: 761–770.
8. A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), TarekKanan ; OdaiSadaqa ; AmalAldajeh ; HanadiAlshwabka ; Wassan AL-dolime ; ShadiAlZu’bi ;May 2019
9. "An Automatic Reference Aid for Improving EFL Learners’ Formulaic Expressions in Productive Language Use", Mei-Hua Chen, Chung-Chi Huang, Shih-Ting Huang, Jason S. Chang, and Hsien-Chin Liou, IEEE Transactions On Learning Technologies, Vol. 7, No. 1, January-March 2014.
10. Natural Language-based User Interface for Mobile Devices with Limited Resources, So-Young Park, Member, IEEE, JeunghyunByun, Hae-Chang Rim, Do-Gil Lee, and Heuiseok Lim, IEEE Transactions on Consumer Electronics, Vol. 56, No. 4, November 2010.
11. ”An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation Roberto Navigli and MirellaLapata", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 4, April 2010.
12. "NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models", Shusen Liu, Zhimin Li, Tao Li, VivekSrikumar, Valerio Pascucci and Peer-Timo Bremer, IEEE Transactions On Visualization And Computer Graphics, Vol. 25, No. 1, January 2019.

13. Data science in light of natural language processing: An overview",ImadZerouala* and AbdelhakLakhouajaa, Faculty of Sciences, Mohamed First University, Av Med VI BP 717, Oujda 60000, Morocco.elsevier, science-direct.
14. Raja, S. Kanaga Suba, and T. Jebarajan. "Reliable and secured data transmission in wireless body area networks (WBAN)." European Journal of Scientific Research 82, no. 2 (2012): 173-184.
15. "The Text Encoding Initiative: 30 years of accumulated wisdom and its potential for a bright future.",Romary, L.: In: Language Technologies & Digital Humanities 2016 (2016).
16. "Developing typewritten Arabic corpus with multi-fonts In: Proceedings of the International Workshop on Multilingual OCR. ACM (2009), Khorsheed, M.S., Alhazmi, K.M., Asiri, A.M.: