Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230454

Prediction and processing of natural language

Yi Wang

School of Computer Science and Engineering, Tianjin University of Technology, 288 Xiuchuan Road, Xiqing District, Tianjin, China.

100877@yzpc.edu.cn

Abstract. With the rapid increase in the computing power of electronic computers and the dramatic decrease in the cost of manufacturing, researchers are refocusing on the challenging research field of natural language processing (NLP). Therefore, in a natural language training development as the main line, from the following several aspects work: first of all, on the basis of preliminary training technical update route, this paper introduces the traditional natural language training in advance technology and training neural network prediction technology, and analyse the related technical features, comparison, generalize the development and trends of natural language processing technology; Secondly, this paper introduces the improved natural language processing models based on BERT from two aspects, and summarizes these models from the aspects of pre-training mechanism, advantages and disadvantages, and performance. Furthermore, the main application fields of NLP are introduced, and the current challenges and corresponding solutions of NLP are described. Finally, this paper summarizes the work and predict the future development direction of NLP.

Keywords: natural language processing, semantic analysis, pre-training techniques.

1. Introduction

Natural language processing pre-training has different names at different times, but the essence is to use a large number of corpuses to predict the corresponding word or phrase, generating a semi-finished product to train the subsequent task. At present, there are relatively many reviews of pre-training techniques based on neural networks, but for most of the traditional pre-training techniques are not involved or mentioned in a single pass, there is an artificial separation of the development of natural language pre-training, which is not conducive to the development of natural language processing technology [1]. Neural network pre-training technology is a general term for the use of neural network models for pre-training in the pre-training stage, because the coupling between pre-training and subsequent tasks is not strong, it can become a model alone, so it is also called pre-training language model, which is different from the traditional pre-training technology.

Because natural language processing involves text, speech, video, image. And different types of corpora, the concept is relatively broad. This article introduces natural language processing pre-training technology. Specifically, first from the pass to illustrate the techniques of integrated pre-training and neural network pre-training. The basic methods of pre-training and the advantages and disadvantages of each method are introduced.

DOI: 10.54254/2755-2721/4/20230454

2. Traditional pre-training technique

As far as most of the natural language processing pre-training articles have been published so far, few articles have a more detailed introduction to traditional pre-training techniques, and the reasons for this may be caused by the following two points: First, in traditional natural language processing, pre-training techniques and models have strong coupling, and there is no independent and separable pre-training technology; Second, the development of neural networks, especially deep neural networks, has led researchers to pay insufficient attention to traditional natural language processing techniques, including traditional pre-training techniques and traditional models. However, traditional natural language processing technology (including traditional pre-training technology and traditional model) as a product of the historical stage of natural language processing, has played a major role in promoting the development of natural language, so it is necessary to introduce the traditional pre-training technology in more detail.

2.1. Vector space modeling techniques

The main idea is that all the corpora in a corpus are represented in the form of spatial vectors, and each feature word of the corpus corresponds to each dimension of the corpus vector. Specifically, the technique contains several main steps such as text preprocessing, feature selection and feature calculation, and algorithm to calculate the accuracy, etc. The text of the vector space model is represented as a bag-of-words model. Since this paper introduces the pre-training technique, this section focuses on the feature engineering (feature selection and feature computation) related techniques [2].

2.1.1. TF-IDF technology. TF-IDF uses unsupervised learning, taking into account both word frequency and freshness attributes, to filter some common words and retain important words that can provide more information. In practical applications such as search engines, this technique is the main means of information retrieval. However, using word frequency to measure the importance of a word is not comprehensive and this calculation does not reflect positional relationships; at the same time, it relies heavily on the level of word separation (especially for Chinese word separation).

2.1.2. Information Gain Technology. Theoretically, information gain should be the best feature selection method, but in practice, since many features with higher information gain tend to occur less frequently, when the number of features selected using information gain is relatively small, there is usually a data sparsity problem, and the model is less effective at that time. Therefore, generally, when the system is implemented, the information gain is first calculated for each word (in terms of words as features) appearing in the training corpus, and then a threshold value is specified to remove those words from the feature space whose information gain is lower than this threshold, or the number of features to be selected is specified, and the features are selected to form a feature vector in the order of the highest to lowest gain value. The information gain is suitable for examining the contribution of features to the overall model, but not to a specific category, which makes it suitable for "global" feature selection but not "local" feature selection. This technique is suitable for classification fields such as sentiment classification, intent recognition, and automatic spam processing.

2.2. Semantic analysis

2.2.1. Implicit semantic analysis. LSA does not need to determine the semantic encoding, but only relies on the connection of things in the text context and uses semantic relations to represent the text, simplifying the purpose of text vectors. LSA uses low-dimensional word and text vectors instead of original space vectors, which can effectively handle large-scale corpus and has fast and efficient features, and is suitable for generative natural language processing such as information filtering, text summarization and cross-language information retrieval such as machine translation. However, LSA ignores the grammatical information of words (even the order of words in a sentence) when performing information extraction, and the object of processing is the visible corpus, which cannot computationally

DOI: 10.54254/2755-2721/4/20230454

obtain the metaphorical and analogical inferential meanings of words, and requires a large number of documents and words to obtain accurate results, which has the disadvantage of low characterization efficiency.

2.2.2. Probabilistic implicit semantic analysis. The authors argue that each corpus contains a series of possible potential topics, and that each word in the corpus is not generated out of thin air, but by certain probability guided by these potential topics, which is the core idea of the generative model proposed by PLSA [3]

PLSA can interpret the model from a probabilistic perspective, making the model tolerant. It is easy to understand; at the same time, PLSA's EM (expectation maximization) algorithm has linear convergence speed compared to LSA's SVD method, which can make the likelihood function reach local optimum. However, the model is unable to generate new unknown documents, and the complexity of the model increases rapidly as the number of documents and words increases, leading to severe overfitting of the model.

3. Neural network pre-training techniques

Neural network natural language pre-training techniques have been improved, mainly by taking into account the contextual relationship between word order and the actual corpus. However, the paper has a large crossover in different classifications. However, the review is recent and narrowly focused, and does not cover the part on neural network pre-training techniques and traditional pre-training techniques (Table 1).

3.1. Word vector fixed representation

Word vector fixation representation is to take contextual related words of target words into account, which can better solve the problem of lexical isolation and incoherence. Common word vector fixed representations are neural network language model (NNLM), C&W (Collobert and Weston), Word2vec (word to vector) [4], etc.

3.1.1. Neural language model. Neural Language Model NNLM: The neural language model estimates the value of $P\left(w_i \mid w_{i-(n-1)}, w_{i-(n-2)}, \cdots, w_{i+1}\right)$ by modeling a meta-linguistic model. Unlike traditional techniques, NNLM does not perform probabilistic calculations of the target conditions by counting, but rather models the solution of the target by constructing a neural network structure. Figure 3 shows the NNLM model structure.

The NNLM model uses low-dimensional compact word vectors to represent the above, which solves the problems of data sparsity and semantic gaps brought by bag-of-words models. This technique is generally applied in the field of sentence-level natural language processing such as missing value interpolation, sentence slicing, recommendation systems, and text noise reduction. However, the model can only use the above information of the current corpus for normalization operation, and cannot adjust the word meaning in real time according to the context; meanwhile, the number of parameters of the model is significantly larger than other traditional models. The LBLM model has no activation function and uses word vectors directly from the hidden layer to the output layer, which makes the model more concise and accurate. However, theoretically LBLM needs to construct multiple matrices (several matrices are required for several words), and the realistic pressure to use approximation processing, thus there are deviations in accuracy; meanwhile, LBLM still cannot solve the problem of multiple meanings of words.

3.1.2. C&W Technology. The C&W technique is a modeling technique proposed by Collobert and Weston in 2008 that aims at generating word vectors (most previous models have generated word vectors as a by-product), which designs models and objective functions directly from the perspective of distributed hypotheses. the structure of the C&W model is shown in Figure 1.

DOI: 10.54254/2755-2721/4/20230454

Table 1. Summary of traditional pre-training techniques.

Large category of models	Specific models	Technical Features	Advantages	Disadvantages	Application conditions and scope
Vector space model	Unique Hot Code	Small extension of text tables to European space for easy calculation and comparison	Expanded features, simple and effective, easy to understand	Too high dimensionality, semantic gaps and inability to reflect the proximity of words	Suitable for models based on parameters and distances, such as SVM, NN, KNN, etc.
Vector space model	TF-IDF	Calculation of word importance based on word frequency and inverse document frequency	Unsupervised learning that filters common words and retains important words: information	Does not reflect position relations and relies heavily on participles	Applicable to the field of question-and-answer search, such as search engines, query systems, etc.
Textrank Technical wood	Information Gain	Difference in information entropy of feature information before and after appearance	Theoretically it should be the best feature selection method, theoretically perfect	Less frequent words with higher information gain, thus producing data sparsity	Suitable for classification fields, such as spam filtering, sentiment classification, etc.
Semantic Analysis	Implicit Semantic Analysis	Use low-dimensional word and text vectors instead of original space vectors	Fast, efficient and easy to understand models	Ignore the grammatical information of words, cannot calculate the allusion meaning and analogical inference meaning of words, need a lot of documents to get accurate results and low characterization efficiency	For generative natural language processing domains such as information filtering, text summarization, and cross-language information retrieval such as machine translation
Semantic Analysis	Probabilistic implicit semantic analysis	EM method is used instead of singular value decomposition SVD	Fast, efficient and easy to understand models	Ignore the grammatical information of words, cannot calculate the allusion meaning and analogical inference meaning of words, need a lot of documents to get accurate results and low characterization efficiency	For generative natural language processing domains such as information filtering, text summarization, and cross-language information retrieval such as machine translation
	Bayesian	A directed acyclic graph of -God from prior probability to posterior probability	Short, fast and less complex	Physical meaning is insufficient and does not correspond to reality	Suitable for word-level natural language processing domains.

DOI: 10.54254/2755-2721/4/20230454

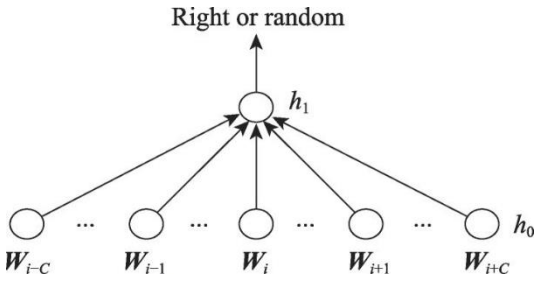

Figure 1. C&W model.

The C&W model differs from the NNLM mainly in that C&W places the target words in the input layer, while the output layer is changed from |V| nodes in the neurolinguistic language model to one node, and the value of this node represents the score of the n-tuple phrases, which is only high or low, without the probability property, so no normalization operation is required [5-6]. The C&W model uses this approach to reduce the |V|×|h| operations of the NNLM model at the last level to |h| operations, which greatly reduces the time complexity of the model. However, the C&W model only utilizes local context, which cannot solve the problem of multiple meanings of words; at the same time, the context information cannot be too long, and there is information loss if it is too long.

3.1.3. Word2vec technology. Word2vec is a word embedding tool open sourced by Google in 2013. embedding is essentially a low-dimensional vector representation of the corpus text, where vectors with similar distances correspond to objects with similar meanings. word2vec tool contains two main models: continuous bag of words (CBOW) and skip-gram of words (CBOW figure 2) and the skip-gram model. figure 3) Two efficient training methods: negative sampling (negative sampling) and hierarchical Softmax (hierarchical Softmax). Since this paper introduces pre-training techniques, this subsection only introduces two models: continuous bag of words and skip-gram.

The CBOW model input is a unique thermal code; the hidden layer has no activation function, i.e., it is a linear unit; the output layer dimension is the same as the input layer dimension, using Softmax regression. The subsequent task processes the new task with the parameters learned by the training model (e.g., the weight matrix of the hidden layer), rather than with the trained model.

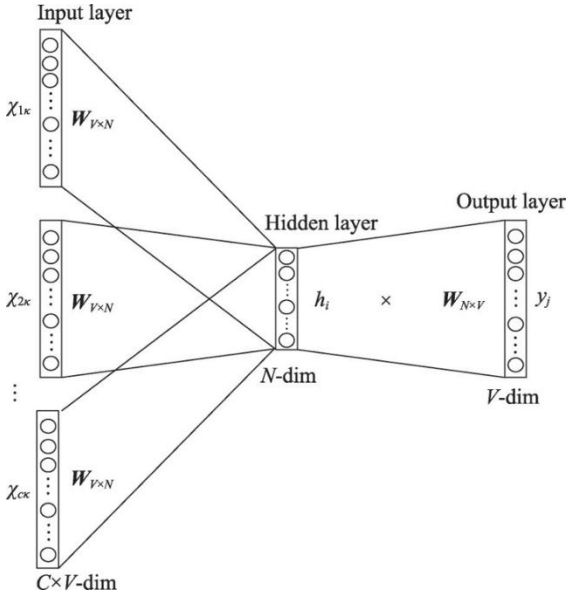

Figure 2. CBOW model.

DOI: 10.54254/2755-2721/4/20230454

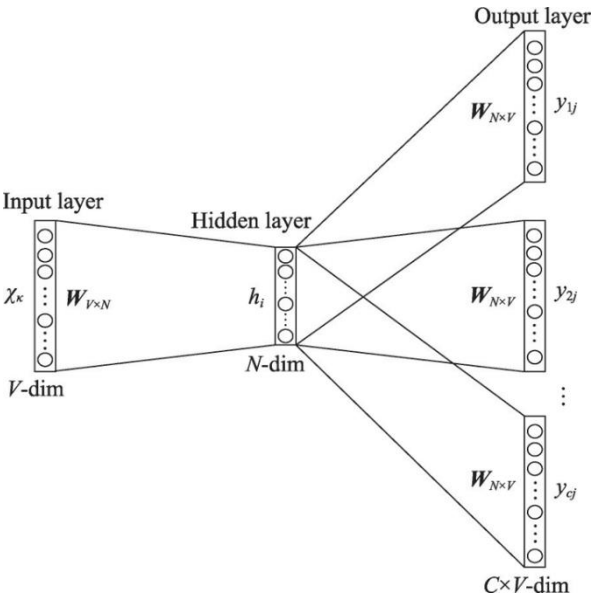

Figure 3. Skip-gram model.

Since Word2vec considers contextual relationships, compared with traditional Embedding, the embedding has fewer dimensions, is faster, more general and more effective, and can be applied in a variety of natural language processing tasks, such as common text similarity detection, text classification, sentiment analysis, recommendation systems and question and answer systems, and other sentence-level and chapter-level natural language processing fields [7]. However, it cannot solve the problem of multiple meanings of words because of the one-to-one relationship with vectors. Also, Word2vec is a static method that cannot be dynamically optimized for a specific task, and its relevant context cannot be too long.

3.2. Word vector dynamic representation

Word vector dynamic representation takes into account the contextual related words of the target word in the pre-training stage, and the context of the target word when it comes to specific utterances, which can better solve the problems of lexical isolation and multiple meanings of words. Then various dynamic representation techniques were born, such as Elmo (embeddings from language models), GPT (generative pre-training) and BERT models.

3.2.1. GPT model. The training process is simple: n word vectors of the sentence are input to the Transformer with positional encoding, and n outputs are predicted for the next word at that position. Figure 10 shows the one-way Transformer structure of GPT and the model structure of GPT.

In general, GPT is divided into two phases: unsupervised pre-training and supervised fitting, with a subsequent fitting phase after the first phase of pre-training. The model is similar to Elmo, with two main differences: first, a Transformer is used as the feature extractor instead of an LSTM; second, GPT uses a one-way language model as the target task.

The GPT model uses Transformer as a feature extractor, which can effectively extract corpus features compared to LSTM. Although its application area is relatively wide, its most prominent area is text generation. However, the one-way Transformer technique used will lose more key information. GPT-2 still follows the GPT one-way Transformer model, but with some improvements on GPT. First, instead of fine-tuning the modeling for different layers separately, the model does not define the specific tasks of this model, and the model will automatically identify what tasks are needed; second, the complexity of the corpus and the network is increased; further, the layer normalization of each layer is put before

DOI: 10.54254/2755-2721/4/20230454

each Sub-block, and the last Self-attention After the last Self-attention, another layer normalization operation is added.

Compared with the GPT model, GPT-2 is more capable of extracting information and has particularly superior performance in text generation. However, the disadvantage of this model is the same as GPT in that more critical information is lost by using a one-way language model [8]. In the field of general-purpose NLP, GPT-3 has the highest performance so far, but its performance on some economic and political problems is less satisfactory (caused by the quality of the pre-trained corpus); at the same time, the model is so huge that most scholars can only look at it from a distance, and it is still far away from the real practical stage.

3.2.2. *BERT model.* BERT uses exactly the same two-stage model as GPT, firstly, language model pre-training, and secondly, fitting training for subsequent tasks. The main difference from GPT is the use of Elmo-like bi-directional language modeling technique, MLM (mask language model) technique and NSP (next sentence prediction) mechanism in the pre-training phase. Figure 4 shows the BERT model.

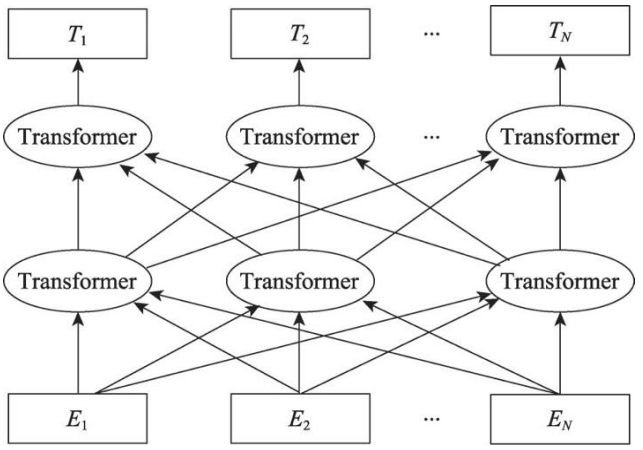

Figure 4. BERT pre-training model.

BERT uses a bi-directional Transformer technique to train word vectors more accurately, which in turn has led to a major earthquake in natural language processing. At this stage, most of the commonly used natural language processing techniques are based on BERT and its improvement techniques. From this stage, BERT has a wide range of application areas, from hotspot areas such as text classification and reading comprehension in natural language understanding to automatic digesting and text writing in natural language generation. However, the model has disadvantages such as huge number of participants and practical application difficulties [9].

As the most widely used modeling technique in the field of natural language, BERT model has spread to various fields of natural language processing and achieved great development. However, scholars have studied that BERT still has more obvious defects. Firstly, the NSP pre-training technique used in BERT leads to topic prediction, which is simpler than the actual prediction, and thus the effect is biased; secondly, the use of random Mask partial words instead of continuous phrases also leads to the discounted effect of BERT; finally, BERT has a relatively large number of parameters compared to other models, which is difficult to deploy on performance-constrained edge devices. on performance-constrained edge devices.

4. Challenges and Solutions

Machine translation was introduced in the 1840s, and natural language processing technology was born. After decades of development, natural language processing technology has evolved in a tortuous manner. For the present, it still faces great challenges, specifically in the following aspects.

DOI: 10.54254/2755-2721/4/20230454

4.1. Corpus

The corpus suffers from irregularity, ambiguity and infinity problems. First, the establishment of a large corpus inevitably requires automated or semi-automated tools to collect and organize the corpus, in which some problematic corpus may be collected, thus causing some impact on the performance of the model. Secondly, due to the characteristics of the corpus itself, there are ambiguities in the semantics, especially in some everyday words. Finally, the corpus itself is infinite, and it is impossible to create an infinitely large corpus.

To address the three problems of the corpus, the following aspects should be addressed. First, when collecting the corpus, it should choose the corpus with formal sources and greater influence to collect and organize; meanwhile, researchers should not focus all the attention on the size and performance of the model only, and developing more intelligent, fast and convenient tools for collecting and organizing the corpus is also one of the focuses in the next stage. Second, due to the ambiguity of the corpus itself, model development should be increased to make the models more intelligent; meanwhile, researchers can borrow some traditional techniques, such as word formation, to make the ambiguous corpus semantically homogeneous. Third, in daily natural language processing, the collection of dedicated corpora should be increased, and at the same time, under the condition of pre-training on large-scale unsupervised corpus, it is a necessary measure to adopt zero or small sample wood learning for subsequent tasks.

4.2. Models

Each development of natural language processing models from rule-based to statistics-based to neural network-based results in a more substantial improvement in their accuracy. The hottest neural networks at this stage have the problems of opaque, simple and crude model processes with large parameters. Specifically, the intermediate process of neural network models, especially deep neural network models, resembles a black box, and researchers have weak control over it, which does not facilitate optimal design. At the same time, the current stage of neural network models is designed in a simpler way compared to traditional refined design models, and most neural network models rely on large computational volumes for training and prediction, thus making the models appear less dexterous. Finally, the models are large in magnitude, and current mainstream models need to consume a large number of resources for training [10]. Although there is a large amount of work on light weighting models, the general lightweight models have scenarios that are limited or still difficult to deploy on edge devices.

In view of the above problems of the model, researchers should start to solve them from the following aspects. Firstly, researchers should increase the research on the intermediate process of the model to make the "black box" transparent and controllable; meanwhile, they should conduct further research on the design of the model to design a lightweight and simple model with strong generalization ability; finally, to address the problem that a large number of lightweight models cannot be applied in production life, they should conduct Finally, to address the problem that a large number of lightweight models cannot be applied to production life, the size of the models should be reduced by secondary light weighting or even multiple light weighting, and the size of the models should be reduced by circular iteration.

4.3. Application Scenarios

For most of the current landed technologies, the scenarios are generally independent and unambiguous, but the natural language processing application scenarios are scattered and complex, and difficult to be applied to a specific domain independently. Researchers should standardize a scenario that meets the public perception and is independent, which is of great practical significance for better application of natural language models to specific domains. Second, at this stage, the fields of automatic digesting and machine translation are in full swing, thus reflecting the strong momentum in the field of natural language generation, and researchers should increase their research in this area (Table 2).

DOI: 10.54254/2755-2721/4/20230454

Table 2. Challenges and solutions.

Major Issues	Difficulties	Technical limitations	Research trends and solutions
corpus	The existence of irregularity and ambiguity of the corpus collection tools of non-intelligence, the corpus itself to choose a reliable and formal source of the corpus, the development of intelligent and fast corpus collection workers and the problem of infinity	The existence of ambiguity models can not handle	It will increase the development of models and the collection of special corpus by drawing on traditional techniques.
Models	Model process is not transparent, simple and brutal at this stage researchers can not figure out the model within the model to increase the model intermediate process research, design a light and easy and generalization ability and the model is huge	Operating Mechanism of the Department	model, multiple light-weighting of the current stage of the model
Application Scenarios	The application field of natural language processing technology is difficult to statute a scenario that meets the public perception should be borrowed from other fields to statute a scenario that meets the public perception and the scene is scattered and complex	and independent scenes	Stand-alone application scenarios
Performance Evaluation	From the existing corpus to divide a currently not standardized and unified performance evaluation should draw on the idea of software engineering, all-round assessment of the model	Estimation Method	The capabilities of each item should include minimum functional testing, invariance testing and orientation

5. Conclusion

Natural language processing has made great progress and has been applied industrially in many fields, and has shown some market value and potential. However, there are still many bottlenecks in natural language processing technology, such as severe performance limitations on complex corpus and difficulty in understanding sentence meaning at the semantic level.

Natural language processing should be combined with other related fields: With the development of neural networks, especially the rise of deep learning, the connection between natural language processing and other disciplines has been further strengthened, and a large number of cross-cutting technologies have been created, such as combining natural language processing with speech to improve speech recognition performance, and combining natural language processing with images to produce interpretable images. In the next research work, the scope of integration with other fields should be increased, so that the results of natural language processing technology can benefit a larger area and accelerate its own development.

Natural language processing techniques should be combined with other techniques: natural language processing techniques involve data mining, probability theory, pattern recognition and other related knowledge. Related technologies can be borrowed and migrated to natural language processing, to avoid

DOI: 10.54254/2755-2721/4/20230454

closed doors to a certain extent. Of course, the development of natural language processing technology and other related technologies is a mutual promotion process.

References

[1] Liang J, Chen J H, Zhang Q, et al. Anomaly detection based on one-hot encoding and convolutional neural network. 2019 *J. Tsinghua Univ.*, 59(7): 523-529.

[2] Bacchi S, Gluck S, Tan Y, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. 2020, *Int. Emer. Med.*, 15(4).

[3] Tahir M, Hayat M , Gul S , et al. An intelligent computational model for prediction of promoters and their strength via natural language processing. 2020 *Chem. Intel. Lab. Sys.*, 202:104034.

[4] Yu H Q. Experimental Disease Prediction Research on Combining Natural Language Processing and Machine Learning. 2019 *Int. Conf. Com. Sci. Net. Tech*.020.

[5] Dominey P F, Inui T, Hoen M. Neural network processing of natural language: II. Towards a unified model of corticosteroid function in learning sentence comprehension and non-linguistic sequencing. 2009, *Brain Lang*, 109(2-3):80-92.

[6] Samsonova M, Pisarev A , Blagov M . Processing of natural language queries to a relational database. 2003 Bioinformatics, 241.

[7] Sitaula C , Shahi T B . Natural language processing for Nepali text: a review. Artificial Intelligence, 2022 *Int. Sci. Eng. J.,* (55-4).

[8] Jeon S , Colburn Z , Sakai J , et al. Application of natural language processing and machine learning to radiology reports. 2021 *Int. Conf. Bio., Com. Bio. Heal. Inf*. 129.

[9] Forestiero A , Papuzzo G . Natural language processing approach for distributed health data management. 2020 *Int. Emer. Med.*871.

[10] Zhang Y , Bogard B , Zhang C . Development of Natural Language Processing Algorithm for Dental Charting 2021, *Chem. Intel. Lab. Sys* 32.