IMTKU Emotional Dialogue System for Short Text Conversation at NTCIR-14 STC-3 (CECG) Task

Min-Yuh Day[1], Chi-Sheng Hung[1], Jhih-Yi Chen[1], Yi-Jun Xie[2], Yu-Ling Kuo[2] and Jian-Ting Lin[1]

Tamkang University, Taipei, Taiwan

myday@mail.tku.edu.tw,
{askahung0702, jj330418}@gmail.com
{emily983227, sylvia_1119, 606630274}@gms.tku.edu.tw,

Abstract. This paper describes the IMTKU (Information Management at Tamkang University) emotional dialogue system for Short Text Conversation at NTCIR-14 STC-3 Chinese Emotional Conversation Generation (CECG) Subtask. The IMTKU team proposed an emotional dialogue system that integrates retrieval-based model, generative-based model, and emotion classification model with deep learning approach for short text conversation focusing on Chinese emotional conversation generation subtask at NTCIR-14 STC-3 task. For the retrieval-based method, the Apache Solr search engine was used to retrieve the responses to a given post and obtain the most similar one by each emotion with a word2vec similarity ranking model. For the generative-based method, we adopted a sequence-to-sequence model for generating responses with emotion classifier to label the emotion of each response to a given post and obtain the most similar one by each emotion with a word2vec similarity ranking model. The official results show that the average score of IMTKU is 0.592 for the retrieval-based model and 0.06 for the generative-based model. The IMTKU self-evaluation indicates that the average score is 1.183 for retrieval-based model and 0.1the 6 for the generative-based model. The best accuracy score of the emotion classification model of IMTKU is 87.6% with bi-directional long short-term memory (Bi-LSTM).

Team Name. IMTKU

Subtasks. STC-3(CECG)

Keywords: artificial intelligence, deep learning, dialogue systems, encoder-de-coder, sequence-to-sequence, recurrent neural network, long short-term memory

2

1 Introduction

NTCIR (NII Testbeds and Community for Information Access Research) is a conference which is organized by Japan. NTCIR-14 STC-3 is a competition for short conversation published in 2018. Chinese Emotional Conversation Generation (CECG) is subtask of NTCIR-14 STC-3. The main task of this competition is to make the conversation emotional [1].

In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. We submit two runs with retrieval-based method and generative-based method for NTCIR-14 STC-3 CECG subtask. Each question will prompt a different response based on each emotion. For the retrieval-based method [2], we use the Apache Solr search engine to retrieve the best response. For the generative-based method [3, 4], we build a sequence-to-sequence neural network model and LTSM model with an attention mechanism to encode a post sentence into a sequence of 256-dimensional vectors, and decode them into a sequence of comment words with a weighted attention.

The experimental result shows that the average score of the retrieval-based model is better than generative-based model.

2 Background

With the development of Artificial Intelligence, Chat Bots are starting to be applied widely in many fields. However, most Chat Bots can only respond to particular questions, which makes users feel a lack of intimacy in their interaction. Consequently, how to make Chat Bot capable of interacting with people in an emotional way has been a challenge task recently. This research focuses on Chat Bots that can not only interact with initial basic communication skills but also respond with corresponding emotions. As a result, we apply different models for this research. Background information of these models is described as follows.

2.1 Retrieval-based Dialogue Model

The term “information retrieval” was first published by Calvin Mooers in 1950 [5]. In the beginning, we used certain equipment and methods to find the information we needed from documents, materials or data in a certain format. With the emergence of different types of information, the technology of information retrieval began to develop in various ways. At present, the conversation robots based on the retrieval model are commonly used by enterprises [6]. For example, Microsoft's Little Bing relies on the principle of collecting a large number of human conversational languages for indexing, and using the user's dialogue sentences to find the most appropriate response in the corpus.

2.2 Generative-based Dialogue Model

The generative learning model is an unsupervised learning model based on knowledge. After Wittrock first proposed the sub-model in 1974, a series of experiments and studies were conducted [3], and long short-term memory (LSTM) [7] is one of the most commonly used generation models.

2.3 Sentiment Analysis

Sentiment analysis [8] is also known as opinion exploration [9]. It is the use of natural language processing or biometrics technology through artificial intelligence technology. In recent years, many researchers have studied sentiment classification. General text sentiment analysis primarily analyzes articles or sentences as positive or negative emotions. But with more in-depth research, we can analyze more emotional states, such as “happy”, “sad”, “angry” and so on. Zhou et al. [10] added emotions to the dialogue system and classified emotions as “Like”, “Happy”, “Sad”, “Disgust” and “Angry”.

3 Research Method

This research will develop a system that can use collected dialogue information to process, analyze, and training data with systematization, and finally produce emotional dialogue. Here, we will introduce the

4

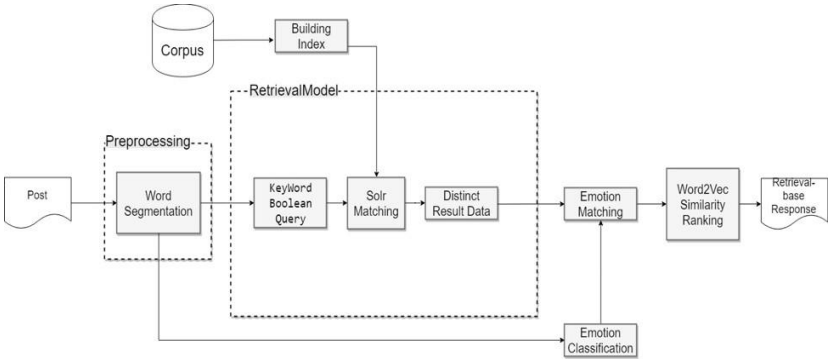

Fig. 1. The system architecture of IMTKU retrieval-based model for NTCIR-14 STC-3

research method and system architecture, and explain the procedure of the research.

3.1 Data Preprocessing

The data set we received from is unstructured. To make the data more effective, the data will be pre-processed. The source of the corpus is the messages posted by the users on the Weibo. There may be many meaningless symbols in the content. To make the data analysis more effective, this study also did 'data cleaning' to remove redundant points.

3.2 Retrieval-Based Model

This study used a retrieval-based model, which is one of the main conversation models. We collected 600,000 chatting pairs. The corpus also provided the connected emotion, so he retrieval-based model yielded a better performance.

We used Apache Solr as a searching engine for the retrieval-based model. Solr is an Open Source search engine platform provided by Apache. The main function is to do full-text searches, it's one of the most popular search engines. In this study, the data in the corpus is classified into fields, and the data is input into the Solr search engine through the Python suite provided by Solr as a dataset of the search model.

After the database for the search model is established, the post is segmented by Jieba and the semantics were analyzed.

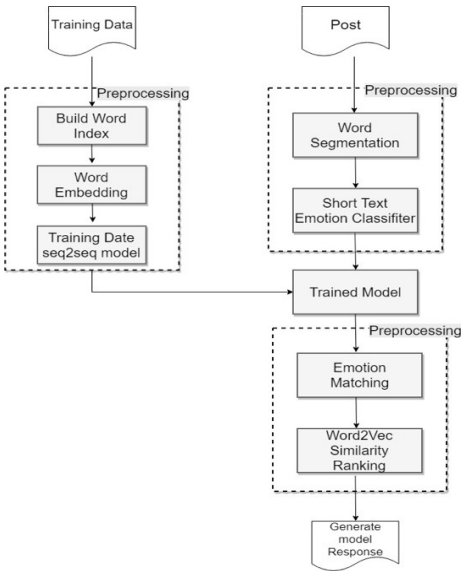

Fig. 2. The system architecture of IMTKU generative-based model for NTCIR-14 STC-3

We extracted the keywords and emotion categories in the post, and searched for relevant results according to different emotions using a boolean search. Finally, we found the most suitable one by Word2Vec similarity. Figure 1 presents the system architecture of IMTKU retrieval-based model for STC-3.

3.3 Generative-Based Model

This study also uses the generative model as the experimental framework of another dialogue robot. It uses a long short-term memory (LSTM) as a training model for generative dialogue. The generative model is more like a general model. It uses daily chat conversation, but lets the meaningful dialogue be generated naturally and is in line with the mood of coping, which many experts have focused on in recent years. Figure 2 presents a the system architecture of IMTKU generative-based model for STC-3.

6

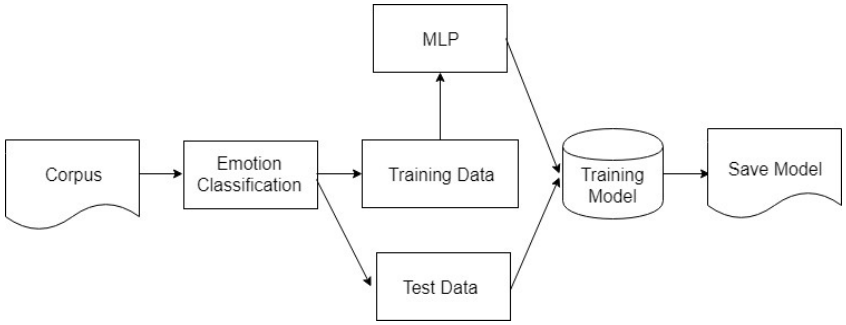

Fig. 3 The system architecture of IMTKU emotional prediction

3.4 Emotion Classification Model

We proposed a Multilayer Perceptron (MLP) [11] as our emotional prediction model. Figure 3 presents a structural diagram of our emotional prediction model. We used the long short-term memory (LSTM) and Bi-Directional LSTM [12] to train other two sentiment prediction models. We compare these two models with the original ones.

4 Experimental Results and Discussions

The results of every model in this experiment will be demonstrated here, including the presentation of the experiment and the methods of evaluation.

4.1 Retrieval-Based Model

The main purpose of the information retrieval dialogue model is to use a pre-defined corpus to find a suitable response with logical operations. This experiment inputs the posts, responses and corresponding emotions in the experimental corpus into Solr to establish a search engine. In the information retrieval phase, the dialogue is first used to break the words with Jieba; after the word segmentation, the related responses are retrieved through boolean search and grouped by emotion. Finally, we used the Word2Vec model to obtain the similarity of the posts and responses. The Word2Vec similarity ranking model performs the weight of the similarity, and the response with the

POST	Response Emotion	Response
为什么你们都不陪我看电影？ Why are you not accompanying me to the movies?	喜欢 Like	最近有什么电影好看？ What movie is great recently?
	难过 Sadness	呵呵～为什么呀？ Hehe~Why?
	厌恶 Disgust	大过年的，请不要在这自攻自受好吗… The New Year, please don’t take it for yourself.
	生气 Anger	为什么啦？ Why?
	幸福 Happiness	啥电影啊？[偷笑] [偷笑] [偷笑] What movie?[smirking]

Fig. 4 The example of the conversation with the retrieval-based model

Table 1. The hyperparameters for training generative-based model

Parameter	Value
Epoch	500000
Batch_Size	64
Drop_Out	0.7
Dense	1
Optimizer	adam
Max_Len	15

highest weight in each emotion group is used as the output dialogue. Figure 4 shows the example of the conversation with the retrieval-based model.

4.2 Generative-Based Model

This experiment uses long short-term memory (LSTM) with attention mechanism as a training model for generative dialogue. Before training the LSTM dialogue model, it is necessary to first segment the training data and establish the word index and the word vector. To increase the effectiveness of the generated dialogue, this study padded all training data the same length. As shown in Table1, the setting of the parameter for training generative-based model. After generating the dialogue model, this experiment integrates the dialogue model with the sentiment prediction model and the similarity model. The post will generate a total of 700 responses for each emotion through our system. We can obtain

8

POST	Response Emotion	Response
为什么你们都不陪我看电影？ Why are you not accompanying me to the movies?	喜欢 Like	我都不知道啊！ I don’t know!
	难过 Sadness	快过一节面试…上课几天啦？ Going through an interview…How many days in class?
	厌恶 Disgust	今天有五节课吗？ Are there five classes today?
	生气 Anger	多少号玩手机 How many numbers to play with the phone
	幸福 Happiness	很久没看过了啊 I haven’t seen it for a long time.

Fig.5 The example of the conversation with the generative-based model.

Table 2. The hyperparameters of the three emotion classification model

Hyperparameter	Value
Epoch	20
Batch_Size	64
Drop_Out	0.7
Dense	1

the best response of each emotion by ranking model. Figure 5 shows the example of the conversation with the generative-based model.

4.3 Emotion Classification Model

In this study, a deep learning model of multi-layer perceptron is used to train the sentiment analysis and prediction model in the sentiment analysis experiment. After the competition, we used long short-term memory (LSTM) and Bi-Directional LSTM to train two sentiment prediction models. We compared these two models with the original ones. The effectiveness evaluation method of the sentiment analysis model is based on the accuracy of the model prediction.

All training data must be converted into a vector before training each model. After the conversion, the multi-class sentiment predictive model can be trained. The sentiment classification of this study is divided into multiple categories. In the multi-class sentiment prediction model, the activation function used is the Softmax, whereas the loss function is Categorical-Crossentropy [13]. Table 2 shows the parameters setting of the three models of this experiment.

9

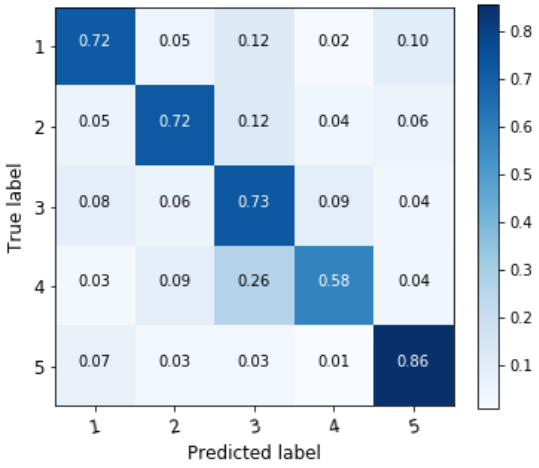

Fig. 6 The confusion matrix of emotion classification model with multi-layer perceptron (MLP)

The multi-layer perceptron sentiment prediction model obtained the prediction accuracy of 73.91% after training 20 Epoch. Figure 6 provides The confusion matrix of emotion classification model with multi-layer perceptron (MLP).

After training 20 Epoch, the long short-term memory sentiment prediction model obtained 84.4% accuracy through test data prediction. Figure 7 shows the confusion matrix of the emotion classification model with long short-term memory (LSTM).

10

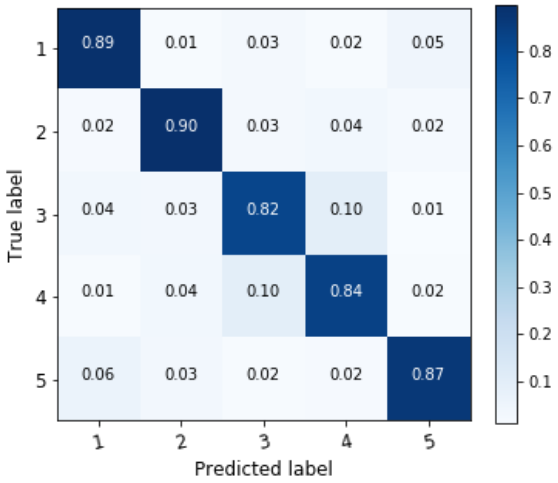

Fig. 7 The confusion matrix of the emotion classification model with long short-term memory (LSTM).

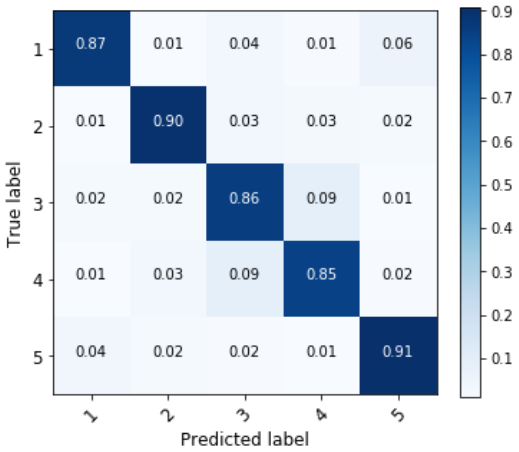

Fig. 8 The confusion matrix of the emotion classification model with bi-directional long short-term memory (BiLSTM).

The bi-directional long shor-term memory sentiment prediction model achieve an accuracy of 87.6% after training 20 Epoch. Figure 8 shows the confusion matrix of the emotion classification model with bi-directional long short-term memory (BiLSTM).

Table 3. The experimental results of emotion classification model

Emotion Classification Model	**Loss**	**Accuracy**
multi-layer perceptron (MLP)	0.788	73.9%
Long Short Term Memory(LSTM)	0.365	86.4%
Bi-directional long short term memory (BiLSTM)	**0.334**	**87.6%**

Table 3 presents the summary of the experimental results of emotional classification model. We can see that this experiment has found that the bi-directional long-term memory (BiLSTM) sentiment prediction model works best.

4.4 Ranking

This study used the Word2Vec approach to train the similarity model [14, 15]. The corpus used a total of 309,602 articles from Wikipedia, which are used as training data after the word segmentation and stop words are removed through the Jieba. By using Word2Vec skip-gram algorithm, we finally use the trained Word2Vec similarity model to calculate the weight of similarity, and use the response with the highest weight in each emotion group as the output dialogue. The formula for calculating the weight is as follows (1):

$$x = (Sp*\alpha+Sr*\beta)/2 \quad (1)$$

The *Sp* in the formula is the similarity index of the input sentence and the post, *Sr* is the similarity index of the input sentence and the corpus response, α is the weight number 0.2, and β is the weight number 0.8.

4.5 Result

4.5.1 Self-evaluation

12

```
IF Coherence and Fluency
    IF Emotion Consistency
        LABEL 2
    ELSE
        LABEL 1
ELSE
    LABEL 0
```

Fig9. Official scoring standard

Table 4. Emotional conversation robot consistency evaluation form (self-evaluation)

	All agree	Two agree	All disagree
Quantity	1366	603	31
Percentage(%)	68.3%	30.2%	1.5%

Our team used the official scoring standard (Figure 9) to re-evaluate to the response to the statement “whether it fits the question” (coherence and fluency) and “return whether it corresponds to emotion” (emotion consistency) as the scoring standard. The maximum score of each reply is two points, which is one point for one, but if the reply is not in line with the question, even if there is a corresponding emotion in the reply, it will not be given.

Through this research method, an emotional conversation robot consistency evaluation table is compiled (see Table 4). From Table 4, we can see that 98.5% of the responses are due to the principle of majority decision, which is a very good result in terms of evaluation consistency. In addition, 1.5% of the evaluation results are different. This part is judged and scored by the researchers. Table 5 represents the self-evaluation score we reassessed.

In this study, the retrieval-based and generative-based models are finally evaluated, and the scores are summed up by Equation 2 and the average score is calculated by Equation 3. In addition, this study uses Equation 4 to calculate an index between 0 and 1. If the index is closer to 1, it means that the effect of analyzing the emotion and reply emotion is better. Otherwise, the closer to 0, the worse the effect of analyzing the emotion and reply emotion is. In addition to total score and average score, we

Table 5. A comparison table of emotional dialogue robot scores (self-evaluation)

	Label 0	Label 1	Label 2	Total Score	Average Score	ACR Index
Retrieval-based Model	304	209	487	1183	1.183	0.591
Generated-based Model	875	90	35	160	0.16	0.08

proposed the Affective Conversational Robot Index (ACR Index) for a better evaluation of an emotional dialogue system. Table5 represents a comparison table of emotional dialogue robot scores (self-evaluation). The total score, average score, and ACR index is defined as follows:

$$\text{Total Score} = \sum_{i=0}^{2} i * num_i \tag{2}$$

where i is the score corresponding to Label, num_i is the total number of topics marked with Label.

$$\text{Average Score} = \frac{\sum_{i=0}^{2} i * num_i}{\text{Nt}} \tag{3}$$

where i is the score corresponding to Label, num_i is the total number of questions marked with Label, and Nt is the total number of all questions.

$$\text{ACR Index} = \frac{\sum_{i=0}^{2} i * num_i}{\text{Nt} * \max(i)} \tag{4}$$

where i is the score corresponding to Label, num_i is the total number of questions labeled Label, Nt is the total number of all questions, and $\max(i)$ is the maximum value of i.

From Table 5, we can see that the emotional dialogue robot with the retrieval-based model is better than the generative-based model. This study also calculates the scores of various emotions. In the scores of different emotions, the effect of happiness prediction is better than that of other emotions, and the prediction of angry emotions is not good enough.

14

Table 6. The official assessment of IMTKU team at NTCIR-14 STC3

	The Result		Like		Sad	
	Overall Score	Average Score	Overall Score	Average Score	Overall Score	Average Score
IMTKU-1	592	0.592	127	0.635	120	0.6
IMTKU-2	60	0.06	8	0.04	17	0.085
	Disgust		Anger		Happy	
	Overall Score	Average Score	Overall Score	Average Score	Overall Score	Average Score
IMTKU-1	97	0.485	88	0.44	160	0.8
IMTKU-2	7	0.035	11	0.055	17	0.085

4.5.2 Official assessment

The organizers of NTCIR-14 STC-3 provided the official evaluation results of our submitted two runs for the retrieval-based model and generative-based model. IMTKU_1 represents our retrieval-based model; IMTKU_2 represents our generative-based model. Table1 6 shows the score of our team, including overall score , average score and scores of each emotions (Like, Sad, Disgust, Anger, Happy). From Table 6, we can see that the result of emotional robot with the retrieval-based model is better than the generative-based model. We can also find the emotion prediction of Happy is the best one, and the worst one is Anger.

5 Conclusion

In this paper, we report our emotional dialogue system for short text conversation at NTCIR-14 STC-3 (CECG) subtask. We proposed two different models for short text conversation , including retrieval-based method and generative-based method.

This study used a consistent value assessment in the evaluation of the retrieval-based method and the generative-based model. The evaluation shows that the retrieval-based model is superior to the generation-based model in the dialogue. Although the response of the retrieval-based method is fixed, there is a large number of corpus and responses with different emotions, and there are diverse responses. In this work, in the

evaluation of individual emotions, the effect of "Happy" emotional prediction and response performs the best, and "Anger" has the worst effect of emotional prediction and response. This may be related to the distribution and prediction accuracy of various emotions provided by the corpus of the emotional training model. This is worth observing and refinement.

We have implemented an emotional dialogue system prototype and provided the system architecture and the development approaches for building emotional dialogue system. The system implements the training of emotional prediction model and generative dialogue model, in which the dialogue model and the sentiment analysis model is integrated with the similarity model and automatically produces a variety of emotional responses.

Our future work will focus on the development of better deep learning language models for sentence generation. For instance, integrating bi-directional LSTM with attention mechanism, convolutional neural network (CNN), and generative adversarial networks (GAN) to improve the generative-based model emotional dialogue model.

6 Acknowledgement

This research was supported in part of TKU research grant and Ministry of Science and Technology. We would like to thank the support of IASL, IIS, Academia Sinica, Taiwan.

References

1. Zhang, Y., & Huang, M. Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge. Proceedings of NTCIR-14 (2019)
2. Lowe, R., Pow, N., Serban, I., & Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. Proceedings of the SIGDIAL 2015 Conference. (2015).
3. Jaakkola, T. and D. Haussler. Exploiting generative models in discriminative classifiers. Advances in neural information processing systems. (1999)
4. Alex Graves. Generating Sequences with Recurrent Neural Networks. Preprint arXiv: 1308.0850. (2013).
5. Mooers, C. N. Information retrieval viewed as temporal signaling. Proceedings of the International Congress of Mathematicians (Vol. 1, pp. 572-573). (1950)
6. Wu, Y., et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2017)

16

7. Hochreiter, S. and J. Schmidhuber. "Long short-term memory." Neural computation 9(8): 1735-1780. (1997)
8. Nasukawa, T. and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture, ACM. (2003)
9. Liu, B.. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies **5**(1): 1-167. (2012)
10. Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
11. Rummelhart, D. E. "Learning internal representations by error propagation." Parallel Distributed Processing: I. Foundations: 318-362. (1986)
12. Graves, A., et al.. Hybrid speech recognition with deep bidirectional LSTM. Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, IEEE. (2013)
13. Developers, G. "Machine Learning Glossary." from https://developers.google.com/machine-learning/glossary/. (2019)
14. Mikolov, T., et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv: 1301.3781. (2013)
15. Jurafsky, D., Martin, J.H.: Speech and language processing. Volume 3. Pearson London: (2014)