Convolutional Neural Network using a Threshold Predictor for Multi-label Speech Act Classification

for the 5th Dialogue State Tracking Challenge

Guanghao Xu*, Hyunjung Lee**, Myoung-Wan Koo*, Jungyun Seo*
Department of {Computer Science and Engineering*, English**}
Sogang University
Seoul, Korea
guanghao412@gmail.com, {hyunjlee, mwkoo, seojy}@sogang.ac.kr

***Abstract*—Regarding the spoken language understanding (SLU) pilot task of the Dialog State Tracking Challenge 5 (DSTC5), it is required to classify label sets of speech acts on human-to-human dialogues. In this paper, we propose a multi-label classification model with the assistance of algorithm adaptation method. To be specific, a Convolutional Neural Network (CNN) model on top of pre-trained word vectors is adapted for the multi-label classification task by utilizing a threshold learning mechanism. In order to evaluate the performance of our proposed model, comparative experiments on the DSTC5 dialogue datasets are conducted. Experimental results show that the proposed model outperforms most of the submitted model in the DSTC5 in terms of F1-score. Without any manually designed features, our model has advantage of handling the multi-label SLU task, using only publicly available pre-trained word vectors.**

Keywords—Multi-label; Convolutional Neural Network; Speech Act Classification; Algorithm Adaptation.

I. INTRODUCTION

The spoken language understanding (SLU) is one of the core components of an end-to-end dialogue system [1]. The SLU is aimed at extracting semantic meaning of user's utterances and building a concept structure which facilitates for a dialogue manager to decide what to say in the next turn. The pilot SLU task of the Dialog State Tracking Challenge 5 (DSTC5) is no more than challengeable due to the following points: *human-to-human dialogues, cross-linguistic data* and *multi-label classification task* [2]. To be more specific, while human-to-machine dialogues are provided in the previous challenge series, the dialogue corpus of the DSTC5 is constructed on the basis of human-to-human dialogues to secure diverse patterns of utterances. Data translated in English used in the evaluation process are syntactically defective, since the English data as test data is built by translating the Chinese dialogue corpus into English one with a Chinese-to-English machine translation system. Furthermore, a multi-label classification task significantly increases possible combinations of speech acts to be annotated to utterances.

As the DSTC5 builds its corpus by collecting human-to-human dialogues in a natural setting, it is more likely that utterances in a single turn contains much more various pragmatic elements than traditional human-to-machine dialogues, which are collected under the control. It means that more than one speech act can be required to be tagged to a single utterance in order to articulate the semantic meaning of user's utterances elaborately. Consequently, it is more appropriate to segment a single utterance into a sub-utterance having a speech act and incorporate them into a full utterance level, which causes each utterance to contain zero, one or more speech acts. Therefore, a multi-label classification task for polymorphous speech acts is to be tackled in the SLU task of the DSTC5.

However, to the best of our knowledge, there are no previous works that have explored the performance of Convolutional Neural Network (CNN) model on a multi-label speech act classification task, though Deep Neural Network models have achieved remarkable results in text classification task [3, 4]. In this paper, we propose a CNN classifier on top of pre-trained word vectors in conducting a multi-label speech act classification task. In addition, a threshold learning mechanism is engaged to enable our proposed model to produce an output of multiple speech acts. For the purpose of our research, we examine the performances of CNN models built on top of different word-embedding algorithms.

The rest of this paper is organized as follows. Section 2 gives a detailed description of the DSTC5 corpus and the SLU pilot task along with brief review of multi-label classification and some related works. In Section 3, we introduce the architecture of our proposed model and a threshold predictor. The section 4 describes how we set up the experiments for training data and evaluation process. In Section 5, we provide our experimental results to optimize the performance of our CNN classifier on the multi-label classification task. The Section 6 concludes and discusses the future research.

II. BACKGROUND

A. Data Characteristics and Task Description

The DSTC5 provides the TourSG corpus, which consists of dialogue sessions collected from Skype calls between tour guides and tourists focusing on offering touristic information of Singapore [2, 3]. For the SLU task, the system is given the utterances from both the tourist and the guide as its input, and the system subsequently tags the utterances spoken by both the speakers with appropriate speech acts categories and attributes.

This work was supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No. 10048448, Development of Conversational Q&A Search Framework Based on Linked Data).

TABLE I. SPEECH ACT CATEGORIES

Metrics	Descriptions
QST (Question)	Used to identify utterances that pose either a question or a request
RES (Response)	Used to identify utterances that answer to a previous question or a previous request
INI (Initiative)	Used to identify utterances that constitute new initiative in the dialogue
FOL (Follow)	A response to a previous utterance that is not either a question or a request

TABLE II. EXAMPLE TEST UTTERANCES AND SPEECH ACT INFORMATION

Speaker	Utterances	Speech Act Category (Attribute)
Guide	um, sentosa the universal studios in the matter. you see it, the whole family. (嗯，圣淘沙里面的环球影城啦，你看啦，一家大小。)	**FOL** (ACK) **FOL** (INFO)
Tourist	there are still in the place where i can recommend? (还有地方可以介绍的吗?)	**QST** (RECOMMEND) **QST** (WHERE)
Guide	yes, we have, um, the zoo. the daytime the zoo. (嗯，我们有一个动物园，那个日间动物园。)	**FOL** (RECOMMEND) **FOL** (WHERE)
Tourist	how big is the singapore? (新加坡有多大？)	**QST** (INFO)

Each sub-utterance belongs to one of the four basic speech act categories that denote general information of each utterance in the current dialogue flow. More specific speech act information can be annotated by the combination with the speech act attributes. Table 1 gives the list of speech act categories with their descriptions. Reference [3] gives complete list of the speech act attributes.

Table 2 shows Chinese test utterances and ones translated in English that annotated with their corresponding speech act categories and attributes.

B. Multi-label Classification

When it comes to classification methods, single-label classifications such as *binary* and *multiclass classification* are mostly treated, in which one instance $\boldsymbol{x}_i$ is associated with a single label l from a label set L, $|L| > 1$ [6]. On the contrary, each instance of data is labeled with a set of labels $Y_i \subseteq L$, the so-called relevant labels in a multi-label classification. To put it differently, a system predicts that an $|L|$-dimensional target vector $\boldsymbol{y} \in \{0,1\}^L$, where $y_j = 1$ is relevant for a given instance, whereas $y_j = 0$ indicates that it is irrelevant [7].

The existing methods for multi-label classification are divided into *problem transformation* and *algorithm adaptation* methods [6]. The most well-known approach of problem transformation is the binary relevance learning (BR); BR independently trains one binary classifier for each label l in L, which ignores the dependent relationship between the labels. For this reason, BR fails to obtain high predictive performance, because it does not consider label dependency, as it makes the strong assumption of label independency. The other problem transformation methods, such as pairwise decomposition (PW) [8], label power set approach (LP) [6] and classifier chains (CC) [9], have improved on the predictive performance by considering label dependencies during the transformation.

An alternative to problem transformation is *algorithm adaptation*, which deals with data in the whole without converting a set of labels into one label. [10]. Zhang and Zhou [11] introduced a back-propagation neural network adapted for multi-label classification (BP-MLL) by having multiple output nodes, one for each label. Similarly, Nam and Kim [7] proposed a simple neural network (NN) approach, which directly builds upon BP-MLL. Their model named NN was proved as a state-of-the-art model to classify multi-labeled text data. The gains were observed in that this model replaced BP-MLL's pairwise ranking loss with cross entropy and employed recent techniques of deep learning such as rectified linear units (ReLU), Dropout and AdaGrad [7].

C. Related Works on SLU Task of the DSTC

A simple baseline model for the SLU task is provided by the committee of the DSTC 5. It uses a BR approach and trains a set of linear support vector machines (SVM) for multi-label speech act classification. The baseline model utilizes traditional TF-IDF approach based on keywords that appeared in the utterances[1]. This approach, however, cannot detect semantic meanings of utterances effectively, since it only superficially depends on words on the surface level.

In the DSTC 4, Adobe-MIT proposed several classifiers to recognize the speech acts. The best performing model claimed by Adobe-MIT is also operated on the basis of a SVM classifier: the features are the 5000 most common unigrams, bigrams and trigrams. They transformed the multi-label task into a multiclass classification task [12]. Since it is assumed that each utterance belongs to exactly one speech act category and single speech act attribute, the system has shortcomings that it cannot produce multiple speech acts for each corresponding utterance.

III. Model Architecture

In this section, we propose a multi-label classification model based on Convolutional Neural Network (CNN). Our model consists of two modules: a CNN with multiple output nodes that produces scores for each label, and a multi-label threshold predictor that generates a reference point using the scores of the labels. The threshold is then used to for the system to decide whether each label is as relevant or irrelevant.

A. CNN Architecture

Coming up with the architecture of CNN, proposed by Kim [4], we propose a CNN classifier consisting of a single convolutional layer and a single channel, as illustrated in the Fig. 1. Formally, let $\boldsymbol{w_i} \in \mathbb{R}^k$ be the k-dimensional word embedding vector corresponding to i-th word in a given utterance. An utterance $\boldsymbol{x}$ of length n can then be represented as a $n \times k$ matrix:

[1] https://github.com/seokhwankim/dstc.

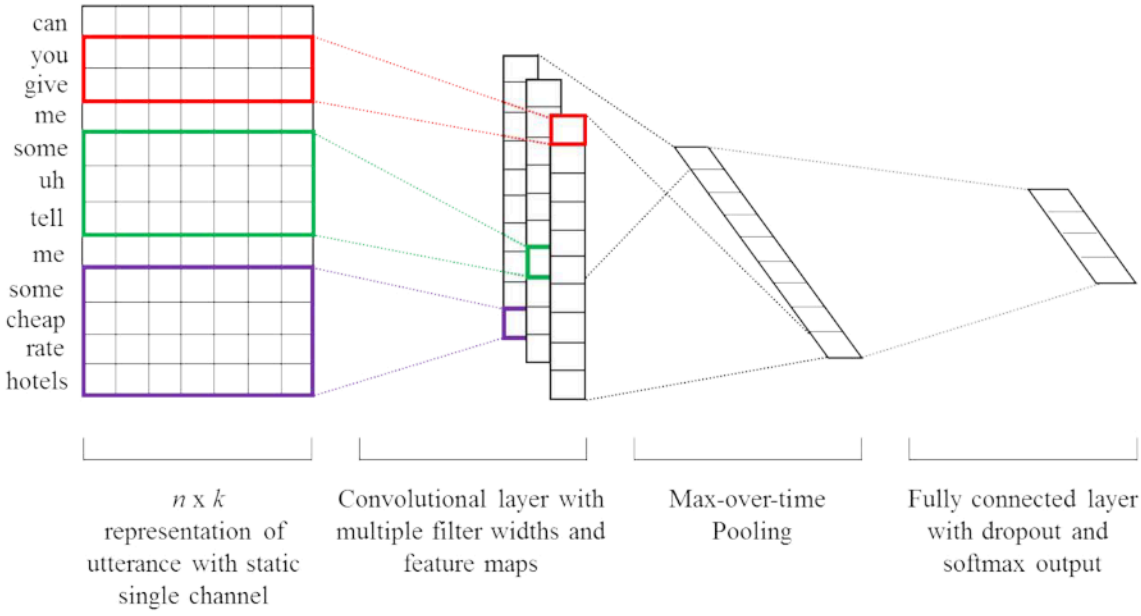

Fig. 1. CNN model architecture with single channel for an example utterance.

$$\boldsymbol{x} = \boldsymbol{w}_1 \oplus \boldsymbol{w}_2 \oplus \ldots \oplus \boldsymbol{w}_n \,, \tag{1}$$

where $\oplus$is a concatenation operator. A convolutional operation involves a filter $\mathbf{m} \in \mathbb{R}^{h \times k}$, which is applied to a window of h words. A feature c_i is generated by

$$c_i = f(\boldsymbol{m} \cdot \boldsymbol{w}_{i:i+h-1} + b) \tag{2}$$

where f is a hyperbolic tangent function and $b \in \mathbb{R}$ is a bias term. The filter is applied to every possible window of words in the utterance to produce a feature map:

$$\boldsymbol{c} = [c_1, c_2, \ldots, c_{n-h+1}] \tag{3}$$

A max-over-time pooling is then operated to take the maximum value $\hat{c} = \max\{\boldsymbol{c}\}$ as a representative feature for this filter.

In our model, multiple filters with varying region size h are integrally engaged into obtain multiple adjacent features. These features are then concatenated into a fixed-length and 'top-level' feature vector, which is passed to a fully connected softmax layer whose output is the probability distribution over all the labels. At the penultimate layer, we apply a Dropout [13] as a means of regularization.

B. Multi-label Threshold Predictor

When a trained CNN is used in prediction for a given utterance $\boldsymbol{x}$, the output probability distribution $p(o|\boldsymbol{x})$ from the softmax layer is used for multi-label prediction. A relevant label set Y for an utterance $\boldsymbol{x}$ is determined by a threshold t as follows:

$$Y = \{j | p_j > t,\ j \in L\}. \tag{4}.$$

The threshold learning mechanism used in the literature [7, 14] is adopted, which models t with a linear regression model. The learning procedure is described as follows: For each training example $(\boldsymbol{x}_m,\ Y_m)$, we set the target values t_m as

$$t_m = \text{argmin}_t(|\{k | k \in Y_m, p_k^m \leq t\}| + |\{l | l \in \overline{Y_m}, p_l^m \geq t\}|) \tag{5}$$

where p_k^m is the output probability of label k associated with utterance $\boldsymbol{x}_m$. The target threshold values t_m is used in learning the parameter θ of the threshold predictor $T(\boldsymbol{x}; \theta)$:

$$E(\theta) = \frac{1}{2}\sum_{m=1}^{M}(T(\boldsymbol{x}_m; \theta) - t_m)^2 + \frac{\lambda}{2}\|\theta\|^2, \tag{6}$$

where λ is the regularization parameter[2]. At the test time, the learned threshold value of a test utterance $\boldsymbol{x}_k$ is used to choose the relevant labels Y, as illustrated in (4).

IV. Experimental Setup

A. Statistics of DSTC5 Datasets

The summary statistics of the SLU datasets for the both speakers of the DSTC5 after tokenization are given in Table 3. For the case of *Guide*, one interesting point to note is that the size of label sets in the train set is smaller than that in the test set, which means that there is no way for the classifier to learn cases of certain labels assigned to utterances during the training and predict correct speech acts in the test set of *Guide*. Additionally, the average number of assigned labels per utterance for the both speakers is very close to 1.0, which indicates that most of the utterances in the datasets are annotated with only one label.[3]

[2] The "sklearn.linear_model.Ridge" package is utilized to learn linear regression model with *l*2-regularization ($\lambda = 1.0$).

[3] Although it sounds peculiar that the most of the datasets are assigned a single label, it is very challenging to predict a few numbers of multi-label speech acts.

TABLE III. STATISTICS OF DSTC5 DATASETS

Datasets (Speaker)	*M*	*D*	*L*	*C*	*W*
Train (Tourist)	14226	3327	74/88	1.19	6.94
Test (Tourist)	4085	1543	61/88	1.16	5.87
Train (Guide)	19916	5462	69/88	1.24	10.66
Test (Guide)	8555	2776	71/88	1.21	7.47

M: Number of utterances. *D*: Size of vocabulary. *L*: Size of label set (size/total). *C*: Average number of labels per utterance. *W*: Average length of utterance (in words).

B. Hyperparameters and Training

In our experiments, we use: Rectified Linear Units (ReLU) [15], filter windows (h) of 2, 3, 4 with 200 feature maps, dropout rate of 0.5 and a batch size of 60. We randomly select 20% of the training data for the validation set. Those values are chosen by adopting a rough grid search. The model undergoes training through stochastic gradient descent over shuffled mini-batches with RMSprop update rule. The model stops the iterant processes of learning by an early stopping mechanism. The CNN models are implemented in Keras [4] framework with Theano [16] backend.

C. Pre-trained Word Vectors

GloVe [17] and *word2vec* [18] are the two most popular word embedding algorithms aiming at mapping semantic meaning of words in a geometric space. In the experiments, we initialize our models with two publicly available pre-trained word vectors; *GloVe* that are trained on 6 billion words from Wikipedia 2014 and Gigaword5[5] and *word2vec* that are trained on 100 billion words from Google News[6]. Both word vectors have dimensionality of 300.

D. Model Variations

We evaluate three models with different word-vectors initialized.

- CNN-rand: All word vectors are randomly initialized and its weights are fine-tuned during training.
- CNN-word2vec: All word vectors are initialized from the pre-trained *word2vec* and its weights are kept static during training.
- CNN-glove: All word vectors are initialized from the pre-trained *GloVe* and its weights are kept static during training.

E. Evaluation Metrics

In the SLU task, a system is required to match relevant speech acts for a given unlabeled utterance spoken by the target role speaker. The following evaluation metrics are used in DSTC5 [1, 3]:

- Precision: Fraction of speech act labels that are correctly predicted.
- Recall: Fraction of speech act labels in the gold standard that are correctly predicted.
- F-measure: The harmonic mean between precision and recall.

V. RESULTS AND DISCUSSION

We evaluate each model in two settings: *CNN-models without a threshold predictor* and *CNN-model-threshold*. The CNN models without a threshold predictor produce only one label of speech act for a given test utterance, since they utilize maximum output probability only. In contrast, the CNN model with a threshold predictor handles the multi-label classification task without any transforming the data with the assistance of the threshold that predicts the probability distribution over all labels. All our models are evaluated 50 times, and the mean scores with standard deviations of each evaluation metric are calculated. Table 4 and Table 5 summarize the comparative results of our models for classifying speech acts of guide and tourist, respectively.

It is observed that even without utilizing threshold learning mechanism, simple CNN models which do not initialize any pre-trained word vectors (CNN-rand) significantly outperforms most of the models submitted in the DSTC5 for both speakers in terms of F1-score. These results suggest that CNN is effective in extracting semantic meanings from the utterances due to the location invariance property of CNN, which makes the system retrieve semantic values independent of the word order. In addition, when the models are built on top of pre-trained word vectors, word2vec and Glove, F1-score are additionally improved. It points that word embedding algorithms are of avail to capture more genuine semantic meanings, since both two models learn words in terms of the semantic relationship based on their context (i.e. co-occurrence) information.

We observe that all the models with the assistance of a threshold predictor show their effectiveness in multi-label speech act classification task, in terms of Recall score. The reason is that the single-label classification models have no chance to predict correct label sets consisting of multiple speech acts. The multi-label classification models, however, have the ability of making precise predictions for the utterance annotated with more than one label. In terms of F1-score, the overall performance of the CNN models is improved with the assistance of thresholds. It is advisable to apply a threshold predictor to the corpus where an average number of labels per utterance is much larger than 1.

Therefore, we propose the *CNN-GloVe-threshold* model, which shows the best performance among all the submitted models in the DSTC5 including Team 2's[7], except for the case of guide. For the case of tourist, *CNN-GloVe-threshold* slightly outperforms the Team 2's model in terms of F1-score. Unfortunately, or the case of guide, *CNN-GloVe-threshold* is behind that of Team 2's model. Nevertheless, it is worthy of noting the fact that our model tackles the multi-label SLU task,

[4] https://keras.io/.

[5] http://nlp.stanford.edu/projects/glove/.

[6] https://code.google.com/archive/p/word2vec/.

[7] One of the anonymous reviewers commented that to consolidate the content, the approach taken by Team 2 should be included in revision of this paper. Unfortunately, by the time of this writing, Team 2's model is not published.

TABLE IV. COMPARATIVE RESULTS FOR GUIDE

Models	Precision	Recall	F1-measure
Baseline (SVM)	0.4588	0.2480	0.3219
Team 2	0.5127	0.4251	**0.4648**
Team 3	0.4340	0.3635	0.3956
Team 5	0.4639	0.3820	0.4190
Team 7	0.5007	0.2976	0.3733
CNN-**rand** (single-label)	0.4562±0.001	0.3757±0.001	0.4121±0.001
CNN-**rand** (+*thresholding*)	0.4070±0.001	0.4287±0.001	0.4175±0.001
CNN-**word2vec** (single-label)	0.4768±0.004	0.3927±0.003	0.4307±0.004
CNN-**word2vec** (+*thresholding*)	0.4239±0.008	0.4295±0.002	0.4266±0.003
CNN-**GloVe** (single-label)	0.4635±0.010	0.3817±0.008	0.4187±0.009
CNN-**GloVe** (+*thresholding*)	0.4183±0.005	0.4320±0.008	0.4250±0.009

TABLE V. COMPARATIVE RESULTS FOR TOURIST

Models	Precision	Recall	F1-measure
Baseline (SVM)	0.3694	0.1828	0.2446
Team 2	0.5331	0.5263	0.5297
Team 3	0.4591	0.4241	0.4409
Team 5	0.5026	0.4484	0.4739
Team 7	0.5079	0.4156	0.4571
CNN-**rand** (single-label)	0.5388±0.005	0.4807±0.005	0.5081±0.005
CNN-**rand** (+*thresholding*)	0.4837±0.014	0.5448±0.004	0.5122±0.006
CNN-**word2vec** (single-label)	0.5462±0.005	0.4873±0.005	0.5151±0.005
CNN-**word2vec** (+*thresholding*)	0.4806±0.005	0.5659±0.008	0.5198±0.005
CNN-**GloVe** (single-label)	0.5537±0.001	0.4940±0.002	0.5221±0.002
CNN-**GloVe** (+*thresholding*)	0.5010±0.002	0.5624±0.002	**0.5299±0.002**

using only publicly available pre-trained word vectors, without any manually designed features.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we propose the CNN classifier on top of pre-trained word vectors in conducting the multi-label speech act classification task. In addition, the threshold learning mechanism is engaged to enable our proposed model to produce an output of multiple speech acts. Experimental results show that the proposed model is highly comparable to the best model submitted in DSTC5. The proposed model is more economical on account of no need to use sophisticated features on the multi-label SLU task.

There is still room for improvement in our model. Although dialogue utterances are dependent in previous dialogue utterances, our model lacks of such ability to track context of the previous dialogue. A well designed recurrent neural network models like long short term memory (LSTM) is expected to complement the tracking ability of our classification model. Those researches are left for our future work.

ACKNOWLEDGMENT

The authors would like thank the associate editor and the anonymous reviewers. Due to their valuable comments and suggestions this paper has been improved and enriched.

REFERENCES

[1] K. Jokinen, and G. Wilcock, "Dialogues with Social Robots, " *Springer*, vol. 999, 2017.

[2] S. Kim, L. F. D'Haro, R. E. Banchs, J. Williams, M. Henderson, and K. Yoshino, "The fifth Dialog State Tracking Challenge," In *Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

[3] S. Kim, L. F. D'Haro, R. E. Banchs, M. Henderson, J. Williams, and K. Yoshino, "Dialog State Tracking Challenge 5 handbook v3.0," 2016.

[4] Y. Kim, "Convolutional Neural Networks for sentence classification," In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746-1751.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, and P. Kuksa, "Natural language processing (Almost) from Scratch," *Journal of Machine learning Research 12,* 2011, pp. 2493-2537.

[6] G. Tsoumakas, and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining,* 2007, pp. 1-13.

[7] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Furnkranz, "Large-scale multi-label text classification – Revisiting Neural Networks," *Machine Learning and Knowledge Discovery in Databases*, Springer, 2014, pp. 437-452.

[8] J. Fürnkranz, E. Hüllermeier, E. L. Mencía and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, November 2008, vol. 73(2), pp. 133-153.

[9] J. Read, B. Pfahringer, G. Holmes and E. Frank, "Classifier chains for Multi-label classification," *Machine Learning*, 2011, vol. 85(3), pp. 333-359.

[10] J. Read, and F. Perezcruz, "Deep learning for Multi-label classification," *Machine Learning,* 2014, vol. 85(3) pp. 333-359.

[11] M. L. Zhang, Z. H. Zhou, "Multi-label Neural Networks with applications to functional genomics and text categorization," *IEEE T. Knowl. Data En.* 2006, vol. 18(10), pp. 1338-1351.

[12] F. Dernoncourt, J. Y. Lee, T. H. Bui, and H. H. Bui, "AdobeMIT submission to the DSTC 4 Spoken Language Understanding pilot task," In *7th International Workshop on Spoken Dialogue Systems (IWSDS),* 2016.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research 15,* 2014, pp. 1929-1958.

[14] A. Elisseeff, J. Weston, "A kernel method for Multi-labelled classification," In *Advances in Neural Information Processing Systems,* 2001, vol. 14, pp. 681-687.

[15] V. Nair and G. E. Hinton, "Rectified Linear Units improve Restricted Boltzmann Machines," In *Proceedings of the 27th International Conference of Machine Learning,* 2010.

[16] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.* 2012.

[17] P. Jeffery, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP),* 2014, vol 12.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In *Advances in Neural Information Processing Systems,* 2013, pp. 3111-3119.