RECURRENT CONVOLUTIONAL NEURAL NETWORKS FOR STRUCTURED SPEECH ACT TAGGING

Takashi Ushio, Hongjie Shi, Mitsuru Endo, Katsuyoshi Yamagami and Noriaki Horii

Interactive AI Research Group, Panasonic Corporation, Osaka, Japan

ABSTRACT

Spoken language understanding (SLU) is one of the important problem in natural language processing, and especially in dialog system. Fifth Dialog State Tracking Challenge (DSTC5) introduced a SLU challenge task, which is automatic tagging to speech utterances by two speaker roles with speech acts tag and semantic slots tag. In this paper, we focus on speech acts tagging. We propose local coactivate multi-task learning model for capturing structured speech acts, based on sentence features by recurrent convolutional neural networks. An experiment result, shows that our model outperformed all other submitted entries, and were able to capture coactivated local features of category and attribute, which are parts of speech act.

Index Terms— spoken language understanding, speech act tagging, text classification, multi-task learning, neural networks

1. INTRODUCTION

Speech act tagging has remained a fundamental problem for automatic agent response and text classification. Several proposals have been made on speech act definitions for describing a variety of human speech utterances in target domain datasets. Switchboard-DAMSL[1], ICSI-MRDA[2], AMI[3], DIT++[4], and some ISO standards[5] are examples of such proposals.

The Switchboard Dialog Act corpus (SwDA) is the most widely used scheme for studying human-to-human dialogs with two speaker. Stolcke et al.'s[1] was one of the first studies to apply a machine learning strategy (HMM) to automatic segmentation of speech units and identification of dialog acts in the SwDA. Yang et al[6], proposed a semi-supervised learning approach using sentence similarity based on word embeddings for classification of ten simplified dialog acts. They represent a dialog utterance as a vector form using word embeddings, then by using a seeded k-means clustering, they classify the utterances into several categories that correspond to particular dialog acts. Their experiment showed classification accuracy to be enhanced by 14% over the conventional method based on a Support Vector Machine. However, their method, since it lacks contextual information, cannot capture dependencies between utterances in a dialog (e.g., between answers and questions), making it unable to fully classify fine-tuned dialog acts, due to their innate ambiguity.

Most previous works treat discourse compositionality as a relation classification problem, and instead aim to capture the semantic aspects of paragraphs or longer texts using bag-of-n-grams or sentence vector averaging strategies[7, 8]. Kalchbrenner et al.[9] have, however, proposed a sentence model based on convolutional neural networks(CNN) and a discourse model based on recurrent neural networks that is conditioned both on the current sentence and on the current speaker. Their experimental results for dialog act classification show that the discourse model is able to capture both the sequentiality of sentences and the interaction between different speakers. The Fifth Dialog State Tracking Challenge (DSTC5)[10] provides the TourSG corpus as a training and evaluation target. This consists of dialog sessions on touristic information for Singapore, collected from Skype calls between tour guides and tourists. Unlike for SwDA, the speech act tagging task of DSTC5 requires the prediction of multiple labels at the utterance level: these comprise two types of information. One is a category for the control of multi turn dialog and the other is an attribute for representation of intent. Definition of categories and attributes, shown in Tables 1 and 2. The utterance of one speaker is not processed by category level segmentation, and both category and attribute contain ambiguity. Consequently, the task represents a more complex and difficult problem in multi-label tagging than for the SwDA.

Various approaches to multi-label problems have been suggested[11, 12, 13]. Cerri et al.[11] have proposed hierarchical multi-label classification using neural networks to predict a label combination, that is hierarchically composed of two types of label. Their method incrementally trains a multi-layer perceptron for each level of the classification hierarchy, that is, the outputs from pre-trained neural networks at a given level are used as inputs to other neural networks responsible for prediction at the next level. However, this method transfers one knowledge into the label combination classifications. Hence, it does not clearly capture both two types of label features. To our knowledge, there are few reports on capturing these for speech act tagging.

In this paper, we describe a local coactivate multi-task

Table 1. Definition of catego

Category	Definition
QST	a question or a request
RES	an answer to a previous ous request
INI	new initiative in the dia constitute either a ques or follow up action to a
FOL	a response to a previous either a question or a re

Table 2. Definition of attribu

Attribute	Definition
ACK	acknowledgment
CANCEL	cancelation
CLOSING	closing remarks
COMMIT	commitment
CONFIRM	confirmation
ENOUGH	no more informa
EXPLAIN	explanation of a
HOW_MUCH	money or time a
HOW_TO	specific instructi
INFO	information requ
NEGATIVE	negative respons
OPENING	opening remarks
POSITIVE	positive response
PREFERENCE	preferences
RECOMMEND	recommendations
THANK	thank you remarks
WHAT	concept related utterances
WHEN	time related utterances
WHERE	location related utterances
WHICH	entity related utterances
WHO	person related utterances and questions

learning model for predicting structured speech acts in DSTC5, and investigate the effects of capturing coactivated local features of both category and attributes. We proceed as follows. In Section 2 we give our motivation and the definitions for the CNN sentence model. In Section 3 we do the same for the RCNN sequence model and proposed model. In Section 4 we describe the speech act tagging experiment and the training procedure. In Section 5 we report the experiment results and discuss with them. Finally we conclude in Section 6.

2. SENTENCE MODEL

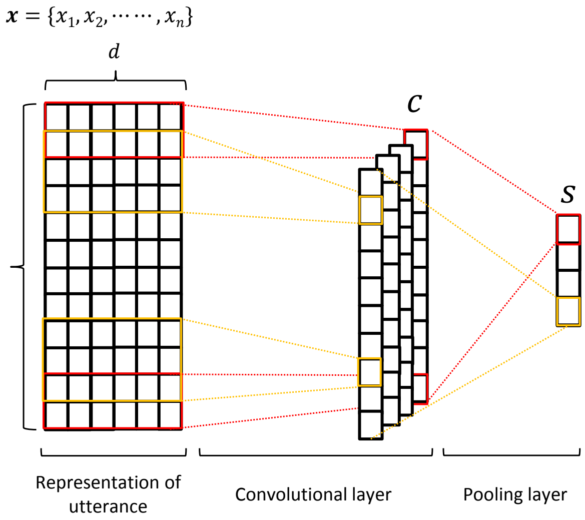
$x = \{x_1, x_2, \cdots\cdots, x_n\}$
d
n
C
S
Representation of utterance
Convolutional layer
Pooling layer

Fig. 1. CNN for Sentence Embedding [Kim, 2014]

A sentence model gives a single vector representation for a sequence of words contained in a sentence. In the discourse model[9] described in Section 1, the sentence model comprises hierarchical convolution layers and max pooling layers and dynamically changes the window size for each layer and sentence length. The sentences that appear in the DSTC5 corpus are unsegmented by speech act category level, and contain multiple intents. The model cannot successfully handle this sentence situation. We use shallower convolutional neural networks[14], whose convolution and pooling result by each filter directly correspond to a dimension of the vector representation. It also has been shown that this CNN model is good performance, including sentiment analysis and question classification[14, 15], despite its simple architecture which requires little tuning.

Our model architecture, shown in Figure 1, is the single-channel CNN architecture of Kim[14]. Let $x_i \in \mathbb{R}^d$ be the d-dimensional word vector corresponding to the i-th word in the sentence. A sentence of length n is given by $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. A zero vector is padded as necessary.

In a convolutional layer, a filter $\mathbf{m} \in \mathbb{R}^{ld}$ is applied to the word vectors in l, where l is the size of window. Created feature c_i is given by:

$$c_i = \sigma(\mathbf{m} \cdot \mathbf{x}_{i:i+l\quad 1} + b) \tag{1}$$

Here $b \in \mathbb{R}$ is a bias term. This filter is used from the beginning of a sentence to $n - l + 1$ to produce a feature map $\mathbf{c} = \{c_1, c_2, \ldots, c_{n\quad l+1}\}$ with $c \in \mathbb{R}^{n\quad l+1}$. Then, a max pooling operation is applied to this feature map, the maximum value $\tilde{c} = max(\mathbf{c})$ is given as the most important feature. We perform the same processes using multiple filters with different window sizes. A vector representation $\mathbf{S} \in \mathbb{R}^k$ for the sentence is given, where k is the number of filters.

Table 3. Example of multiple speech acts for an utterance

Speaker	Transcription	Speech acts
Guide	Let's try okay?	
Tourist	Okay.	
Guide	It's InnCrowd Bacl per person only twe fifty nine dollars .	
Tourist	Um. Wow, that's gc	

3. PROPOSED

Our model outputs multiple spee
when given a sequence of utteran
An example of multiple speech act
in Table 3. We describe the sequ
both the sequentiality of sentence
tween different speakers.

3.1. Sequence Model

Our sequence model is based on RI
the output sequence of a sentence :
resentation $\mathbf{S}_t$ for the t-th utteran
shared CNN at each time step.

Long short-term memory (LST
rent neural network that has a vecto
and a set of element-wise multiplic
information is stored, forgotten, an
As various connectivity designs f
posed, the architecture used in this
following equations.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t\quad 1} + \mathbf{b}_i) \tag{2}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t\quad 1} + \mathbf{b}_f) \tag{3}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t\quad 1} + \mathbf{b}_o) \tag{4}$$

$$\tilde{\mathbf{c}}_t = tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t\quad 1} + \mathbf{b}_c) \tag{5}$$

$$\mathbf{c}_t = \mathbf{i}_t \odot \tilde{\mathbf{c}}_t + \mathbf{f}_t \odot \mathbf{c}_{t\quad 1} \tag{6}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t) \tag{7}$$

where, $\mathbf{U} = \{\mathbf{U}_i, \mathbf{U}_f, \mathbf{U}_o, \mathbf{U}_c\}$, is a set of weights of connection from the previous state to a current state. This is replaced with $\mathbf{U}^{a_t} = \{\mathbf{U}_i^{a_t}, \mathbf{U}_f^{a_t}, \mathbf{U}_o^{a_t}, \mathbf{U}_c^{a_t}\}$, which is a set of weights depending on a binary a_t, which identifies ”Tourist” or ”Guide” in DSTC5. In our preliminary study, this replacement reduced labeling error, especially for category(FOL, RES, INI, QST), which has great importance for capturing the interaction between different speakers compared to the case with attributes.

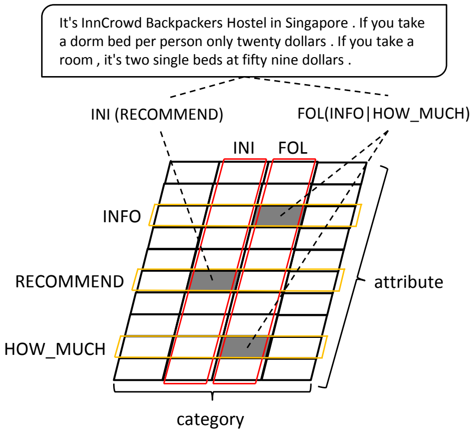
It's InnCrowd Backpackers Hostel in Singapore . If you take a dorm bed per person only twenty dollars . If you take a room , it's two single beds at fifty nine dollars .
INI (RECOMMEND)
FOL(INFO|HOW_MUCH)
INI
FOL
INFO
RECOMMEND
HOW_MUCH
attribute
category

Fig. 2. Example of speech act representation

3.2. Local Coactivate Multi-Task Learning Model

The aim of our model is to capture the local features of both category and attributes. An example of speech act representation is shown in Figure 2.

Speech acts can be often predicted by considering whether they have local features that correspond to category and attribute. For example, it is easy to predict ”Yes ?” as ”QST(ACK)” using its words. Moreover, ”% Uh it's quite a simple clothing.” can be predicted as ”FOL(INFO)” by using its phrases or words.

As an approach for capturing the local features, Multi-Task-Learning(MTL) is a easy and strong learning method for neural networks. It coactivates neurons in the hidden layer between multi tasks, and has specific neurons in the output layer for each tasks. It also has shown that MTL is good performance, including text classification and POS tagging[17, 18].

However, if we capture category and attribute local features with MTL, it often results in the prediction of labels that does not consider its combination since it has the damage of over-fitting for the easier task. For example, a guide response to a tourist question, ”% Uh you can take the train from City Hall Station .”, is incorrectly predicted as both ”FOL(INFO)” and ”RES(INFO)”.

Therefore, our model has two LSTMs and two sentence

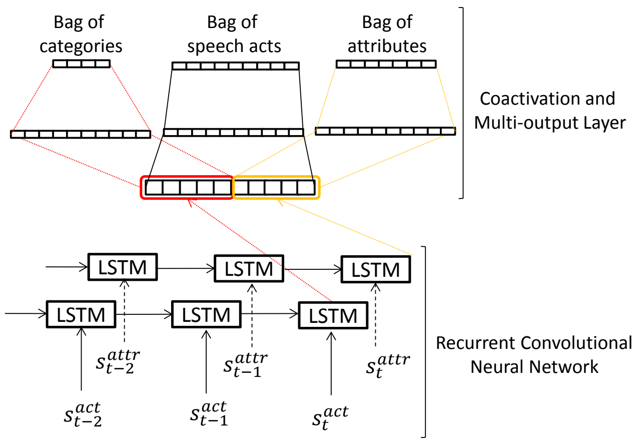

Fig. 3. The architecture of the LC-MTL

embeddings, and locally coactivates neurons in those hidden layer between sub task(categories/attributes) and main task(speech act). The architecture of the Local Coactivate Multi-Task Learning model(LC-MTL) is shown in Figure 3.

Let both $\mathbf{h}_t^{cat}$ and $\mathbf{h}_t^{attr}$ be the h-dimensional LSTM state vector corresponding to the t-th utterance in a dialog. Here, each vector is given by different weightings in the sentence model and sequence model.

$$\mathbf{h}_t^{cat} = LSTM(\mathbf{s}_t^{cat}, \mathbf{h}_{t\;1}^{cat}) \tag{8}$$

$$\mathbf{h}_t^{attr} = LSTM(\mathbf{s}_t^{attr}, \mathbf{h}_{t\;1}^{attr}) \tag{9}$$

Next, concatenation operation is applied to this vector set, which is given by $\tilde{\mathbf{h}}_t = \mathbf{h}_t^{cat} \oplus \mathbf{h}_t^{attr}$. Logistic regression with the hidden layer for this vector is used to predict speech acts. It is given by:

$$\mathbf{o}_t = \sigma(\mathbf{W}^{(2)}\{\sigma(\mathbf{W}^{(1)}\tilde{\mathbf{h}}_t + \mathbf{b}^{(1)})\} + \mathbf{b}^{(2)}) \tag{10}$$

For multiple speech act classification of class $y_t \in \{1, 0\}^J$, where J is the number of categories and attribute combination. We use cross-entropy loss as the objective:

$$L(\theta) = -\sum_{j=1}^{j=J}\{y_t^j \log o_t^j + (1 - y_t^j)\log(1 - o_t^j)\} \tag{11}$$

where y_t^j is the binary corresponding to the j-th class. The loss is summed over all samples in the mini-batch. In the same way, a local feature, $\mathbf{h}_t^{cat}$ and $\mathbf{h}_t^{attr}$, is used and learned to predict a part of speech acts. Finally, our model optimizes the sum of all task objectives, that is, speech acts, categories and attributes.

4. EXPERIMENT

We conducted experiments with the prediction of speech acts in DSTC5[10] and DSTC4[19] to evaluate the effects and generality of the proposed model.

4.1. Details of DSTC4 and DSTC5

The DSTC4 is a prior task of the DSTC5, which is performed on the TourSG corpus. For speech act tagging, all utterances in a dialog are annotated by 84 speech acts consisting of four categories and 21 attributes. The purposes of the DSTC4 and the DSTC5 differ as follows.

- **DSTC4**: Single-language problem in English. In the development phase, we used a training set of 14 English dialogs and six development sets of English dialogs with manual annotations over frame structures. In the test phase, we evaluated the results generated for a test set of nine unlabeled English dialogs.
- **DSTC5**: Cross-language transfer problem based on the DSTC4 corpus, which contains utterances of machine translation between English and Chinese. In the development phase, we used a training set of 35 English dialogs and two developments sets of Chinese dialogs with manual annotations over frame structures. In the test phase, we evaluated the results generated for a test set of eight unlabeled Chinese dialogs.

4.2. Training Procedure

We minimized the loss of predicted and true distributions by back-propagation and included L2 regularization parameters. The input of a Chinese utterance in a dialog is split into units of character for the DSTC5, and into units of English word for the DSTC4.

In CNN, the window size of the filter is unigram, bigram, and trigram, and the number of filters is 100 for each window size. Word vectors are initialized to random vectors of length 100 and the pre-training procedure is performed by word2vec[20] using a large-scale Wikipedia corpus. The RCNN is truncated to a depth d = 5 so that the prediction of a dialog act depends on the previous five utterances since adopting depths > 5 has not yielded improvements in our preliminary study. We minimize the objective using AdaDelta in a mini-batch mode of size 32 by early stopping in a develop set.

5. RESULTS AND DISCUSSION

We trained a couple of speaker role dependent model. In the test phase, thresholds were applied to the sigmoid output in each speech act, which were used as predictions of a speech act. Thresholds were adjusted by the score of the develop set. First, we evaluated whether our model could robustly improve performance, measured as precision/recall/micro-f1 for each speaker role.

Table 4 summarizes the scores for speech act tagging in the DSTC4, comparing the following classifiers. In the same way, Table 5 summarizes our results(under the Model ”team2”) in DSTC5, including the other entry teams.

Table 4. Results for DSTC4

SLU	Guide Speech Act			Tourist Speech Act		
Model	Precision	Recall	F1	Precision	Recall	F1
SVM	0.745	0.615	0.674	0.562	0.467	0.510
Logistic Regression	0.676	0.558	0.611	0.573	0.476	**0.520**
RCNN	0.745	0.594	0.661	0.473	0.516	0.493
MTL	0.707	0.625	0.664	0.447	0.589	0.508
LC-MTL	0.758	0.614	**0.678**	0.487	0.526	0.506

Table 5. Results for DSTC5

SLU	Guide Speech Act			Tourist Speech Act		
Model	Precision	Recall	F1	Precision	Recall	F1
Team0	0.458	0.247	0.321	0.369	0.183	0.245
Team3	0.433	0.363	0.395	0.459	0.424	0.441
Team5	0.463	0.382	0.419	0.503	0.448	0.474
Team7	0.500	0.297	0.373	0.508	0.416	0.457
Team2-RCNN	0.545	0.391	0.455	0.500	0.550	0.524
Team2-MTL	0.530	0.396	0.454	0.533	0.526	**0.530**
Team2-LC-MTL	0.512	0.425	**0.464**	0.560	0.500	0.528

- **SVM**: SVM classifier(System 3) by the DSTC4 team3 (top team in DSTC4 main-task)[21], with unigram, bigram and trigram of current utterance and previous utterance, the binary information indicating whether the current speaker is equal to the previous speaker. Two SVM classifiers were trained: one for each speaker. The kernel function as well as the penalty parameter of the error term were both optimized with 5-fold cross-validation.
- **Logistic Regression**: an Logistic Regression classifier(System 5)[21], with features same as the SVM. However, it uses a single speaker-independent model.
- **RCNN**: our RCNN model with single-task learning for a speech act. It has a LSTM and a sentence embedding.
- **MTL**: our proposed model with multi-task learning for a speech act, category, and attribute. A sentence model and sequence model are shared among all tasks. It has a LSTM and a sentence embedding.
- **LC-MTL**: our proposed model with local coactivate multi-task learning(Figure 3). It has two LSTMs and two sentence embeddings.

The speech act tagging experiment on DSTC5 shows that our model outperforms all the other submitted entries. Moreover, for the DSTC4, without the feature selection, the result from our model is comparable to that of the SVM and Logistic Regression classifier.

5.1. Local Coactivate Multi-Task Learning Model

Both MTL and LC-MTL have the objectives for act and attribute, and are able to capture local features. For example, a guide response to a tourist question, ”% Uh you can take the train from City Hall Station,” RCNN predicted it as non-label, because no strong feature could be captured. MTL incorrectly predicted it as “FOL INFO” and “RES INFO” because no label combination could be clearly captured. In contrast, LC-MTL correctly predicted it to be “RES INFO,” and was able to capture local features coactivated by a label combination.

5.2. Cross Language Transfer Problem

As can be seen from Tables 4 and 5, there are several performance differences between the DSTC4 and DSTC5. The tourist performance was improved by about 2 %, while the guide performance fell by about 20 %. To identify the cause, we investigated the performance in frequent speech acts that are used by both guides and tourists. Figure 4 and Table 6 summarizes the results in DSTC4 and DSTC5.

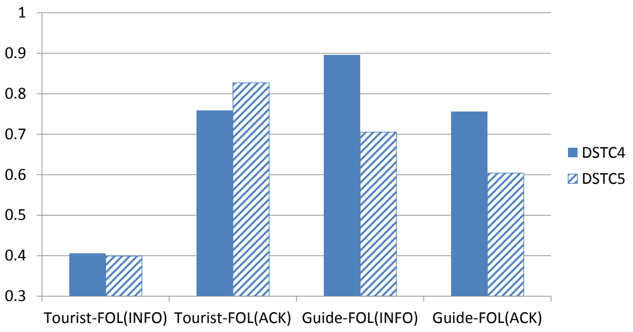

Fig. 4. F1-scores in frequent speech acts

It was clearly observed that the tourist performance was

Table 6. Frequent speech act labels with examples and frequencies in DSTC4 and DSTC5

Speaker	Speech act	Transcript / Translation	Train[%]
Tourist	FOL(INFO)	Well um I'm planning to bi visit Singapore in the near future. / 毕嗯嗯我打算在不久的将来访问 新加坡。	5.5 / 11.4
Tourist	FOL(ACK)	Ah okay. / 啊好的。	27.0 / 25.8
Guide	FOL(INFO)	So almost all the shops in Singapore are having good discounts uh for their shopping. / 所以在新加坡有良好的优惠的那个他们几乎所有的商店买东西。	26.5 / 35.9
Guide	FOL(ACK)	Okay. / 好的。	12.9 / 11.2

improved in for "FOL(ACK)," and the guide performance fell for "FOL(INFO)." This suggests that machine translations are influenced by the length and variety of words in an utterance, because "FOL(INFO)" contains a greater variety of words than "FOL(ACK)." However, the guide performance also fell in with "FOL(ACK)." In many cases, the utterances of "FOL(ACK)" incorrectly were predicted as "None," indicating those labels to be empty. Approximately 30% of guide utterances and 7% of tourist utterances were predicted as "None" in the incorrect predictions made with "FOL(ACK)." For example, it is difficult to distinguish "嗯" and "啊" between back-channel and acknowledgement. Two examples of utterances in English, "% Uh huh" and "Yes" were also translated to "嗯". We therefore found that cross language increases ambiguity in speech act tagging.

6. CONCLUSION

In this paper, we describe a local coactivate multi-task learning model for predicting structured speech acts in the DSTC5, and investigate the effects of capturing local features coactivated by a label combination. The results of our speech act tagging experiment with the DSTC5 showed our model to outperform all other submitted entries. For the DSTC4, the result of our model is comparable to that of the SVM model. We show that our model can effectively capture the local features coactivated by a label combination in speech act tagging. We also found that cross language increases ambiguity in speech act tagging. Our future plans are to share knowledge between users and agents and to investigate the speech act segmentation problem.

7. REFERENCES

[1] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[2] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey, "The ICSI meeting recorder dialog act MRDA corpus," Tech. Rep., DTIC Document, 2004.

[3] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[4] Harry Bunt, "The DIT++ taxonomy for functional dialogue markup," in *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, 2009, pp. 13–24.

[5] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al., "Towards an ISO standard for dialogue act annotation," in *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[6] Xiaohao Yang, Jia Liu, Zhenfeng Chen, and Weilan Wu, "Semi-supervised learning of dialogue acts using sentence similarity based on word embeddings," in *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*. IEEE, 2014, pp. 882–886.

[7] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1201–1211.

[8] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein, "A latent variable recurrent neural network for discourse relation language models," *arXiv preprint arXiv:1603.01913*, 2016.

[9] Nal Kalchbrenner and Phil Blunsom, "Recurrent convolutional neural networks for discourse compositionality," *arXiv preprint arXiv:1306.3584*, 2013.

[10] Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams, Matthew Henderson, and

Koichiro Yoshino, “The Fifth Dialog State Tracking Challenge,” in *Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

[11] Ricardo Cerri, Rodrigo C Barros, and André CPLF De Carvalho, “Hierarchical multi-label classification using local neural networks,” *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014.

[12] Wei Bi and James T Kwok, “Multi-label classification on tree-and dag-structured hierarchies,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 17–24.

[13] Min-Ling Zhang and Zhi-Hua Zhou, “ML-KNN: A lazy learning approach to multi-label learning,” *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[14] Yoon Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.

[15] Mingbo Ma, Liang Huang, Bing Xiang, and Bowen Zhou, “Dependency-based convolutional neural networks for sentence embedding,” .

[16] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, “Representation learning using multi-task deep neural networks for semantic classification and information retrieval,” in *Proc. NAACL*, 2015.

[18] Ronan Collobert and Jason Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[19] Seokhwan Kim, Luis Fernando D’Haro, Rafael E. Banchs, Jason Williams, and Matthew Henderson, “The Fourth Dialog State Tracking Challenge,” in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS)*, 2016.

[20] T Mikolov and J Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, 2013.

[21] Franck Dernoncourt, Ji Young Lee, Trung H Bui, and Hung H Bui, “Adobe-MIT submission to the DSTC 4 spoken language understanding pilot task,” *arXiv preprint arXiv:1605.02129*, 2016.