Chat Discrimination for Intelligent Conversational Agents with a Hybrid CNN-LMTGRU Network

Dennis Singh Moirangthem and Minho Lee
School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
{mdennissingh,mholee}@gmail.com

Abstract

Recently, intelligent dialog systems and smart assistants have attracted the attention of many, and development of novel dialogue agents have become a research challenge. Intelligent agents that can handle both domain-specific task-oriented and open-domain chit-chat dialogs are one of the major requirements in the current systems. In order to address this issue and to realize such smart hybrid dialogue systems, we develop a model to discriminate user utterance between task-oriented and chit-chat conversations. We introduce a hybrid of convolutional neural network (CNN) and a lateral multiple timescale gated recurrent units (LMTGRU) that can represent multiple temporal scale dependencies for the discrimination task. With the help of the combined slow and fast units of the LMTGRU, our model effectively determines whether a user will have a chit-chat conversation or a task-specific conversation with the system. We also show that the LMTGRU structure helps the model to perform well on longer text inputs. We address the lack of dataset by constructing a dataset using Twitter and Maluuba Frames data. The results of the experiments demonstrate that the proposed hybrid network outperforms the conventional models on the chat discrimination task as well as performed comparable to the baselines on various benchmark datasets.

1 Introduction

Dialogue systems can be classified as domain-specific task-oriented and open-domain chit-chat dialog systems (Williams and Young, 2007; Wallace, 2009). The task-oriented dialog systems help users complete tasks in specific domains. The chit-chat dialog systems enable users to have an open-ended chat conversations with the system. While most of the functionalities offered by the two types of systems are complementary to each other, there have been very little efforts made to combine these two type of systems. Therefore, the potential of chat agents have been limited.

Recently, intelligent assistants have become popular with the integration of such systems in smartphones and home appliances. These intelligent assistants typically perform various tasks including weather forecast alerts, alarm settings, web search, and so on. Moreover, such assistants need to have the ability to perform chit-chat conversation with the users. This has led to the need for the development of novel and hybrid multi-domain task-oriented agents and open-domain chit-chat agents.

In order to develop such hybrid agents, we have to determine whether a user will have a chit-chat with the system or the user is looking for a task completion. For example, if a user says "*Hi, how are you doing?*", then the user can be considered to have a chat with the system. Alternatively, if the user says "*I want a flight to Los Angeles*," then the user is looking for a completion of a specific task. We address this task as a binary classification problem and call this task as *chat discrimination*.

Chat discrimination has not been sufficiently investigated in recent times. This is mainly because there are not enough studies to develop hybrids of task-oriented and chit-chat agents. Although task-oriented and chit-chat agents have long research histories, they do not require chat discrimination. We usually assume that the users of task-oriented agents will have task-oriented conversations with the systems and the users of chit-chat

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 30–40
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

agents will always have non task-specific conversations with the systems. In a recent study, researchers in (Akasaki and Kaji, 2017) have tried chat detection using conventional classifiers with the help of a newly created dataset in Japanese language. But this dataset has not been released for further research or comparison.

In this work, we develop a hybrid network for chat discrimination by combining a convolutional neural network (CNN) and a gated recurrent unit (GRU). CNNs have been proven to be suitable for text classification problems (Kim, 2014; Johnson and Zhang, 2015a,b). Moreover, the temporal hierarchy concept with multiple timescale gated recurrent unit (MTGRU) (Kim et al., 2016) has also been proven to perform well in language modeling (Moirangthem and Lee, 2017; Moirangthem et al., 2017) and summarization (Kim et al., 2016) tasks. The MTGRU is known to handle long term dependency better with the help of the varying timescales to represent multiple compositionalities of language. The temporal hierarchy approach has also been shown to eliminate the need for complex structures and normalization techniques (Cooijmans et al., 2017; Krueger and Memisevic, 2016; Chung et al., 2017; Ha et al., 2017), and thereby increasing the computational efficiency of the model.

For our classification model, we develop a lateral multiple timescale structure. Our proposed lateral multiple timescale gated recurrent unit (LMTGRU) is significantly different from the conventional hierarchical MTGRU structure. The conventional MTGRU is most effective for handling long term dependencies in very long text inputs for applications such as summarization but performs comparable to vanilla GRU with shorter text inputs. Unlike the hierarchical architecture, the lateral connections in an LMTGRU will enable encoding of rich features that have different temporal dependencies from the input utterances in order to help classify the information correctly. LMTGRU follows a lateral (branch or root) architecture where the slow and fast units are directly connected to the inputs and the final output of the units are combined to form the final representation. This structure enables all the layers with different timescales to capture relevant features directly from the inputs unlike hierarchical multi-layer structures. Since the data consist of utterances as input, and the input to the RNN is represented as higher order features from the CNN, LMTGRU proves to be more suitable for this task.

Our major contributions are as follows:

- We introduce a hybrid CNN-LMTGRU structure to build rich features from input texts to classify utterances correctly.
- The LMTGRU architecture enables our model to perform well on longer text sequences with the help of the slow layer as well as maintain comparable performance on shorter sequences.
- To address the lack of dataset, we create a dataset using Twitter data (Microsoft Research Social Media Conversation Corpus) (Sordoni et al., 2015) for chit-chat conversations and Maluuba Frames data (El Asri et al., 2017) for task-oriented conversations.
- In order to demonstrate that the proposed model performs well on other text classification tasks and to compare it to the existing baselines, we report the performance on various sentence classification benchmark datasets. The results of our experiments demonstrate that the proposed model performs well on the benchmark datasets as well.

2 Related Work

Although there have been enough studies for task-oriented and chit-chat agents independently, developing hybrid models of the two types of agents has not been explored enough. Therefore, few attempts have been made to develop a chat discrimination model.

Niculescu and Banchs (2015) tried to combine task-oriented agents and chit-chat agents, but the authors did not have a clear way to automatically determine when to switch back to the chit-chat agent. Lee et al. (2009) proposed to combine task-oriented and chit-chat agents with the help of an example-based dialogue manager, but it is difficult to integrate the current state-of-the-art deep learning model based classifiers as a component in such a framework.

Wang et al. (2014) and Sarikaya (2017) proposed to combine a multi-domain task-oriented agents and chit-chat agents using machine-learning-based frameworks. Robichaud et al. (2014); Sarikaya et al. (2016) approached domain classification as ranking between alternate ”dialog experts”. In a recent study, Akasaki and Kaji (2017) tried chat detection using conventional classifiers with the help of a newly created dataset in Japanese language. They used concatenated features from multiple feature extractors for the classification. An end-to-end model was not explored. Moreover, the dataset has not been released for further research or comparison.

Deep learning based models have achieved great success in many NLP tasks, including learning distributed word, sentence and document representation (Mikolov et al., 2013; Le and Mikolov, 2014), parsing (Socher et al., 2013), statistical machine translation (Cho et al., 2014), sentiment classification (Kim, 2014), etc. Learning distributed sentence representation through neural network models requires little external domain knowledge and can reach satisfactory results in related tasks like sentiment classification, text categorization etc.

In recent sentence representation learning works, neural network models are constructed upon either the input word sequences or the transformed syntactic parse tree. Among them, convolutional neural network (CNN) and recurrent neural network (RNN) are two popular ones. The capability of capturing local correlations along with extracting higher-level correlations through pooling empowers CNN to model sentences naturally from consecutive context windows. Kim (2014) proposed a CNN architecture with multiple filters and multiple channels for text classification.

RNNs are able to deal with variable-length input sequences and discover long-term dependencies. Various variants of RNNs have been proposed to better store and access memories. The most popular variants are long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014). Recently proposed MTGRU (Kim et al., 2016), inspired by the concept of temporal hierarchy found in the human brain (Botvinick, 2007; Meunier et al., 2010), demonstrates the ability to capture multiple compositionalities similar to the findings of Ding et al. (2016). This better representation learning capability enhances the ability of the network to model longer sequences of text.

In this paper, we develop a hybrid of CNN and LMTGRU in a unified architecture for semantic sequence modeling. We apply CNN to text data and feed the features directly to the LMTGRU, and hence our architecture enables the network to learn multiple temporal scale dependencies from higher-order features. We hypothesize that the combination of slow and fast features will be beneficial for the chat discrimination task.

3 Proposed Model

We formulate chat discrimination as a binary classification problem. In this section, we explain the proposed hybrid classifier model shown in Figure 1.

3.1 The Convolutional Neural Network Layer

The CNN layer shown in Figure 1 is implemented using a single convolution and max-pooling layer and use a rectified linear unit (ReLU) as the nonlinear activation function following Kim (2014). Let $\mathbf{x}_i \in \mathbf{R}^d$ be the word vector of dimension d corresponding to the i-th word in the input utterance. An utterance of length n, which are padded if necessary, can be represented as

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n, \qquad (1)$$

where $\oplus$ is the concatenation operator. Let $\mathbf{x}_{i:i+j}$ be to the concatenation of words $\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+j}$. A convolution operation involves a *filter* $\mathbf{w} \in \mathbf{R}^{hd}$, which is applied to a window of h words to produce a new feature. For example, a feature c_i is generated from a window of words $\mathbf{x}_{i:i+h-1}$ by

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b). \qquad (2)$$

Here $b \in \mathbf{R}$ is a bias term and f is a non-linear function. This filter is applied to each possible window of words in the sentence $\{\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \ldots, \mathbf{x}_{n-h+1:n}\}$ to produce a feature map

$$\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}], \qquad (3)$$

with $\mathbf{c} \in \mathbf{R}^{n-h+1}$. A max pooling operation (Collobert et al., 2011) over the feature map is applied, which takes the maximum value $\hat{c} = \max\{\mathbf{c}\}$ as the feature corresponding to this particular filter. The idea is to capture only the most important features.

The processes described above is for *one* feature being extracted from *one* filter. The proposed CNN model includes a number of filters with multiple window sizes to obtain various features. These features are then split into

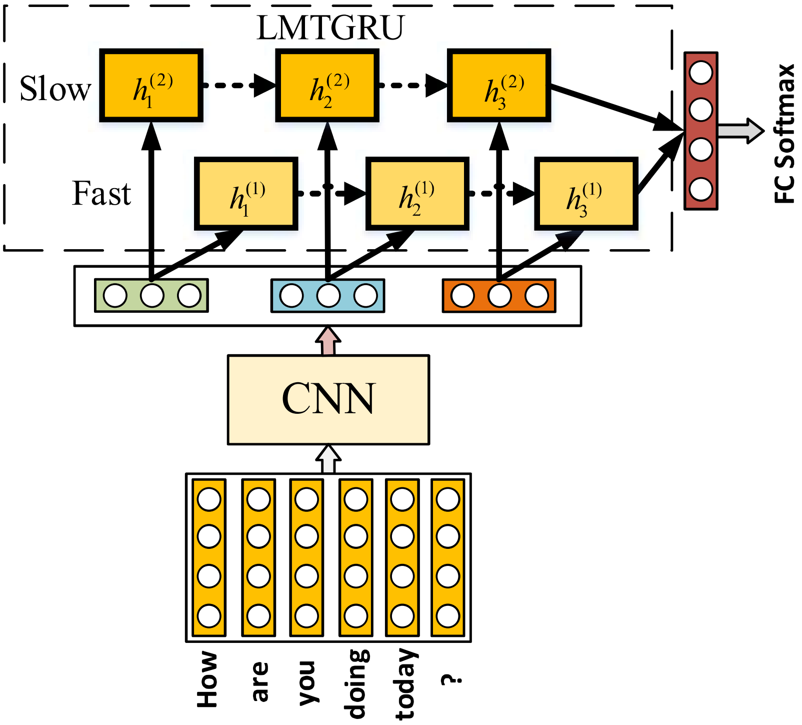

Figure 1: The proposed CNN-LMTGRU classifier. The input to the model is "How are you doing today?"

n/max_pool_size outputs and are passed on to the LMTGRU layer.

3.2 The Lateral MTGRU Layer

For this classification task, we implement a lateral multiple timescale architecture where half of the MTGRU units are fast and the remaining half are slow as shown in Figure 1. The fast and slow units can capture different temporal dependencies from the input sequence. The fast timescale layer can capture fast changing features (e.g. character or word) whereas slower timescales can represent phrase or sentence level features (Moirangthem et al., 2017). The proposed LMTGRU structure follows a lateral (branch or root) architecture where the slow and fast units are directly connected to the inputs. This lateral architecture is different from the conventional MTGRU with a hierarchical layer architecture, since the LMTGRU does not follow a multilayer structure. The LMTGRU structure is implemented using multiple single layer MTGRU networks whose timescales are different and the input to each layer comes directly from the input features. And the final output representation features of each layer are combined to form the penultimate representation of the input sequence that includes both fast and slow features.

The multiple timescales in an MTGRU network is implemented by applying a timescale variable at the end of a conventional GRU unit, essentially adding another gating unit that modulates the mixture of the past and current hidden states. In an MTGRU, each step takes as input $\mathbf{x}_t, \mathbf{h}_{t-1}$ and produces the hidden $\mathbf{h_t}$. The timescale τ added to the activation $\mathbf{h}_t$ of the MTGRU is shown in Eq. (4). τ is used to control the timescale of each GRU cell. Larger τ results in slower cell outputs but it makes the cell focus on the slow features and vice-versa. The timescale variable τ is scalar and one τ controls the slow cells and another τ controls the fast cells. We initialize the τ for each group of cells, e.g. larger τ for slow cells and smaller τ for fast cells. The τ is made as a trainable variable like any other weight of the network and is optimized during the training based on the final loss. An MTGRU cell is illustrated in Figure 2.

$$\begin{aligned}
r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1})\\
z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1})\\
u_t &= \tanh(W_{xu}x_t + W_{hu}(r_t \odot h_{t-1}))\\
\tilde{h}_t &= z_t h_{t-1} + (1-z_t)u_t\\
h_t &= \tilde{h}_t\frac{1}{\tau} + (1-\frac{1}{\tau})h_{t-1}
\end{aligned} \tag{4}$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and tangent hyperbolic activation functions, $\odot$ denotes

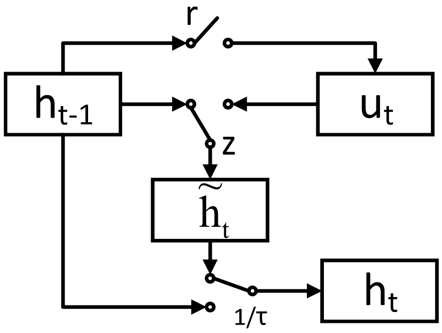

Figure 2: A Multiple Timescale Gated Recurrent Unit. The τ parameter is set for each layer and it controls the timescale of the layer.

the element-wise multiplication operator, and $\mathbf{r}_t$, $\mathbf{z}_t$ are referred to as *reset*, *update* gates respectively. $\mathbf{u}_t$ and $\tilde{\mathbf{h}}_t$ are the candidate activation and candidate hidden state of the MTGRU.

The proposed CNN-LMTGRU hybrid network consists of a CNN layer followed by a fast and a slow LMTGRU layer. The fast units as well as the slow units are directly connected to the CNN features. Finally the combined last hidden representation of the LMTGRU is passed to a fully connected softmax layer whose output is the probability distribution over the labels.

4 Chat Discrimination Dataset

Chat discrimination task requires a chat dataset like the one shown in Table 1. We address the lack of such a dataset by using the Microsoft Research Social Media Conversation Corpus[1] and Maluuba Frames[2] datasets. Microsoft Research Social Media Conversation Corpus is a collection of conversational snippets extracted from Twitter logs. The advantage of using this dataset is that it has been evaluated by crowd sourced annotators measuring quality of the response. These data are suitable for detecting open-domain non-task oriented chats. On the other hand, we use the Maluuba Frames dataset for the domain task-specific conversations. This corpus is for the travel agent domain where the users can inquire the agent and ask for booking of hotels and flights. The dialogs were recorded using 12 participants over a period of 20 days. We process the data to utilize only the user utterances in our chat discrimination dataset. Finally, we have 20,532 utterances with 10,266 in each class. We divide the data into 10% for validation, 10% for test, and the remaining for train.

5 Experiments and Results

We evaluate the performance of the proposed method and compare it to the conventional models using our chat discrimination dataset. In order to demonstrate that the proposed model performs well on other text classification datasets and to compare it to the existing baselines, we report the performance on various sentence classification benchmark datasets as well.

5.1 Experiment settings

We trained the proposed CNN-LMTGRU model in an end-to-end fashion, where we do not use any pre-trained word embedding. An embedding of size 300 was used for the model and was trained with the model. We used 128 filters of sizes $\{3, 4, 5\}$ for the CNN.

We used 300 units of MTGRU where half of the units are fast and the remaining are slow units to construct the LMTGRU structure. The τ for the fast units and the slow units were initialized to 1.0 and 1.25, respectively. We follow Moirangthem et al. (2017) to initialize the timescale parameter. In order to control the τ during training, we set the lower bound to 1.0 using clip by value. This is done as the fastest layer should have a τ of 1.0, however there is no upper bound for the slow layers. After training, the final τ values are 1.16 and 1.37 for the fast and the slow layers, respectively. The learning rate to update the τ, which is different from the global learning rate, is set to 0.00001 in order to avoid large changes in the timescale.

We used the RMSprop Optimizer (Tieleman and Hinton, 2012) to perform stochastic gradient descent with the decay set to 0.9 and the global learning rate to 0.001. For regularization we employ dropout of 0.5 on the final CNN output as well as in the LMTGRU layers to avoid overfitting. We utilized the validation performance for early stopping of the training for better generalization.

5.2 Baseline Models

The baseline models implemented for the comparison using our chat discrimination dataset are described as follows:

CNN We used the same parameters as before except the number of filters were increased to

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52375

[2] https://datasets.maluuba.com/Frames

Type	Example
Chit-Chat	Let’s meet at the coffee place and talk about you. What is your hobby? I will visit my parents for the vacation. I like pop music. Do you like soccer? I don’t know you, but you seem to be a serious person.
Task-oriented	Hello, I am looking to book a trip for 2 adults and 6 children. We are departing from Kochi for Denver. When would I be leaving for each of them? I would like to spend as much time in Denver as my budget will allow. Do these packages have different departure dates? Ok, I would like to purchase the trip with the 4-star hotel.

Table 1: Example utterances of the two kinds of conversations.

256. We followed (Kim, 2014) and used a fully connected softmax layer for the binary classification.

LSTM/GRU The same parameters were used as before except the number of hidden units is increased to 500. The LSTM/GRU takes every word vector in a sequence as input and the final representation is passed to a softmax layer for classification.

LMTGRU This LMTGRU model consists of a fast and a slow layer with 250 hidden units in each layer. The remaining settings are the same as the LSTM/GRU model.

CNN-LSTM/GRU This structure is almost identical to the proposed model, but instead of the LMTGRU, LSTM/GRU is used for comparison. The parameters remain the same.

5.3 Evaluation on Benchmark Datasets

Following Kim (2014), we test our model on various benchmarks. Summary statistics of the datasets are given below.

- **MR**: Movie reviews with one sentence per review. This binary classification task involves detecting positive/negative reviews (Pang and Lee, 2005). The average sequence length is 20 and the dataset size is $10,662$.
- **SST-1**: This is the Stanford Sentiment Treebank is an extension of MR with multiple labels (very positive, positive, neutral, negative, very negative) (Socher et al., 2013). The average sequence length is 18 and the dataset size is $11,855$.
- **SST-2**: This is similar to SST-1 but with binary labels. The average sequence length is 19 and the dataset size is $9,613$.
- **Subj**: Subjectivity dataset consists of sentences with binary labels (subjective or objective). The average sequence length is 23 and the dataset size is $10,000$ (Pang and Lee, 2004).
- **TREC**: The TREC task is a classification task to classify 6 types of question (questions about person, location, numeric information, etc.). The average sequence length is 10 and the dataset size is $5,952$ (Li and Roth, 2002).
- **CR**: Customer reviews of various products with positive/negative labels. The average sequence length is 19 and the dataset size is $3,775$ (Hu and Liu, 2004).
- **MPQA**: Opinion polarity detection is a subtask of the MPQA dataset with 2 classes. The average sequence length is 3 and the dataset size is $10,606$ (Wiebe et al., 2005).

For the evaluation on the benchmark datasets, we implemented a CNN-LMTGRU model that is identical to the one described in Section 5.1. The data for train, validation, and test for the benchmark datasets follow the previous works (Kim, 2014; Kalchbrenner et al., 2014).

5.4 Results

Table 2 illustrates the classification performance of the various models. The performance is given in accuracy and the results show that the proposed hybrid CNN-LMTGRU model outperforms

Model	Accuracy (%)
CNN	91.12
LSTM	89.67
GRU	90.56
LMTGRU	90.64
CNN-LSTM	92.31
CNN-GRU	93.01
Proposed CNN-LMTGRU	**94.69**

Table 2: Chat classification results on the test set.

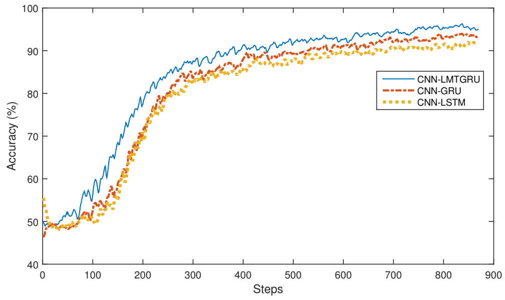

Figure 3: Classification accuracy curve on the validation set of the proposed method and the hybrid baseline models.

the baseline models. The performance curve of the hybrid models is shown in Figure 3, respectively.

In order to differentiate the performance of the proposed CNN-LMTGRU model and the CNN-GRU model, we divide the test data of the dialog classification dataset according to the length of the texts. Figure 4 shows the comparison of the performance accuracy on different lengths of test data. It can be seen that the LMTGRU structure enables the model to outperform GRU on longer text inputs and there is no significant performance degradation with the increase in input length. Whereas, the performance of GRU drops significantly with longer text inputs.

Table 3 shows the result of the comparison of our model with various other models using publicly available sentence classification datasets. These results illustrate that our proposed model either performed comparable to or outperformed existing models.

6 Discussion

When we look at the results illustrated in Table 2, the performance of the proposed CNN-LMTGRU increased significantly compared to CNN-GRU. As shown in Eq. (4), we know that if τ is close to 1, which is the case of a fast LMTGRU layer, the model becomes a vanilla GRU. Therefore, a vanilla GRU is considered as a fast layer and hence, a CNN-GRU network can be considered as a network with only fast units. The difference in performance when we have all the RNN units as fast, i.e. CNN-GRU, and when we have a combination of slow and fast units, i.e. CNN-LMTGRU, show the effectiveness of the multiple timescale approach. The results in Figure 4 also show the significance of the features from slow and fast layers, where the fast features helps maintain the performance with shorter text inputs and the slow features enable the model to perform significantly better with longer text inputs. This confirms our hypotheses that the proposed LMTGRU with the help of both slow and fast units can help encode different dynamic features in order to help classify the sentences and utterances correctly. The results indicate that the LMTGRU architecture increases the capability of the model to learn multiple temporal dependencies better for the discrimination task. The results also demonstrate that our hybrid CNN-LMTGRU network performs significantly better than the existing hybrid models.

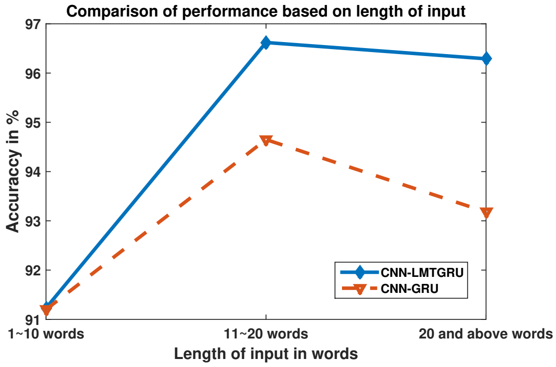

Figure 4: Classification performance comparison based on the length of input.

The results in Table 3 shows that our model performed fairly comparable to the baseline models. The enhanced performance of the proposed model in both SST-2 (average length of 19 words) and MPQA (average length of 3 words) over the baseline models also confirms our hypothesis that the rich features of the slow and fast layers help in the discrimination task even with diverse sequence lengths. However, for some of the datasets such as TREC, our end-to-end learning model cannot outperform the conventional models like SVM due to the limited size of the dataset.

The increased ability of the proposed model to

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-static (Kim, 2014)	81.0	45.5	86.8	93.0	92.8	84.7	89.6
CNN-non-static (Kim, 2014)	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel (Kim, 2014)	81.1	47.4	88.1	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2015)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM$_S$ (Silva et al., 2011)	–	–	–	–	**95.0**	–	–
Proposed CNN-LMTGRU	80.9	48.4	**89.4**	93.4	93.8	84.8	**90.8**

Table 3: Results of our CNN-LMTGRU model against other methods on various sentence classification benchmark datasets.

discriminate between open-domain chit-chat conversations and domain-specific task-oriented utterances will definitely help in the development of hybrid intelligent dialog systems that can handle both types of conversation. Moreover, with the help of this kind of classifier, the chat agents can dynamically switch between utterances in order to conduct a more natural and intelligent conversation with the users.

7 Conclusion and Future Work

This paper addressed the issue of discriminating conversations for combining domain-specific task-oriented agents and open-domain chit-chat agents. We developed a hybrid model consisting of a CNN and an LMTRGU network to classify the conversations. The proposed LMTGRU was able to effectively determine the type of conversation that a user will have with a dialog system. Moreover, we addressed the lack of dataset by constructing a dataset with chit-chat conversations and a task-oriented conversation corpus. We also evaluated the performance of the proposed hybrid model on various benchmark sentence classification datasets in order to compare to several existing models. The results of our experiments illustrated that the proposed end-to-end learning hybrid network with multiple timescales not only performed significantly better in case of longer texts inputs but also maintained good performance in case of shorter texts.

In the future, we plan to develop a more sophisticated dialog discrimination model to handle user utterances that are ambiguous in nature. It will be difficult for the standard classifiers to determine the actual type of conversation in such cases. One of the possible solution is to instruct the chat agent to follow up with clarification questions in case of ambiguity (Schlöder and Fernández, 2015). Another solution is to utilize contextual information by using previous dialogs from the system (Xu and Sarikaya, 2014). We plan to integrate features from the previous utterances for classification. This can be achieved by integrating the lateral architecture of an LMTGRU and the hierarchical organization of MTGRU along with the CNN features from the current and previous utterances to make the decision.

Although the studies on conversational agents have made significant progress in the recent years, it is still difficult for the systems to have a fluent conversation with the users (Higashinaka et al., 2015). We further plan to utilize the chat discrimination model to develop a hybrid system in order to improve such dialog agents. This will also allow us to evaluate the effectiveness of our model for this application.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2017M3C1B6071400)(40%), by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding)(30%), and by the Technology Innovation Program: Industrial Strategic Technology Development Program (10073162) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)(30%).

References

Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. *arXiv preprint arXiv:1705.00746* .

Matthew M Botvinick. 2007. Multilevel structure in behaviour and in the brain: a model of fuster's hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1485):1615–26.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR* abs/1406.1078.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. 2017. Hierarchical multiscale recurrent neural networks. In *Proceeding of the International Conference on Learning Representations*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. 2017. Recurrent batch normalization. In *Proceeding of the International Conference on Learning Representations*.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience* 19(1):158–164.

Li Dong, Furu Wei, Shujie Liu, Ming Zhou, and Ke Xu. 2015. A statistical parsing framework for sentiment classification. *Computational Linguistics* 41(2):293–336.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057* .

David Ha, Andrew Dai, and Quoc V Le. 2017. Hypernetworks. In *Proceeding of the International Conference on Learning Representations*.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 894–904.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *SIGDIAL Conference*. pages 87–95.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 168–177.

Rie Johnson and Tong Zhang. 2015a. Effective use of word order for text categorization with convolutional neural networks pages 103–112.

Rie Johnson and Tong Zhang. 2015b. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in neural information processing systems*. pages 919–927.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 655–665.

Minsoo Kim, Moirangthem Dennis Singh, and Minho Lee. 2016. Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization. In *1st Rep4NLP*. Association for Computational Linguistics, pages 70–77.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*. Association for Computational Linguistics, pages 1746–1751.

David Krueger and Roland Memisevic. 2016. Regularizing rnns by stabilizing activations. In *Proceeding of the International Conference on Learning Representations*.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. volume 14, pages 1188–1196.

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. 2009. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication* 51(5):466–484.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 1–7.

D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, and E. T. Bullmore. 2010. Hierarchical modularity in human brain functional networks. *ArXiv e-prints* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Dennis Singh Moirangthem and Minho Lee. 2017. Temporal hierarchies in multilayer gated recurrent neural networks for language models. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, pages 2152–2157.

Dennis Singh Moirangthem, Jegyung Son, and Minho Lee. 2017. Representing compositionality based on multiple timescales gated recurrent neural networks with adaptive temporal hierarchy for character-level language models. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pages 131–138.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 786–794.

Andreea I Niculescu and Rafael E Banchs. 2015. Strategies to cope with errors in human-machine spoken interactions: using chatbots as back-off mechanism for task-oriented dialogues. *Proceedings of ERRARE, Sinaia, Romania* .

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 271.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 115–124.

Jean-Philippe Robichaud, Paul A Crook, Puyang Xu, Omar Zia Khan, and Ruhi Sarikaya. 2014. Hypotheses ranking for robust domain classification and tracking in dialogue systems. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34(1):67–81.

Ruhi Sarikaya, Paul A Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Celikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, et al. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, pages 391–397.

Julian J Schlöder and Raquel Fernández. 2015. Clarifying intentions in dialogue: A corpus study. In *IWCS*. pages 46–51.

Joao Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. 2011. From symbolic to subsymbolic information in question classification. *Artificial Intelligence Review* 35(2):137–154.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. 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, pages 1201–1211.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 151–161.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. pages 1631–1642.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*. Association for Computational Linguistics.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2):26–31.

Richard S Wallace. 2009. The anatomy of alice. *Parsing the Turing Test* pages 181–210.

Sida Wang and Christopher Manning. 2013. Fast dropout training. In *international conference on machine learning*. pages 118–126.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 90–94.

Zhuoran Wang, Hongliang Chen, Guanchun Wang, Hao Tian, Hua Wu, and Haifeng Wang. 2014. Policy learning for domain selection in an extensible multi-domain spoken dialogue system. In *EMNLP*. Association for Computational Linguistics, pages 57–67.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2-3):165–210.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language* 21(2):393–422.

Puyang Xu and Ruhi Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 136–140.

Bishan Yang and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 325–335.