Co-GAT: A Co-Interactive Graph Attention Network for Joint Dialog Act Recognition and Sentiment Classification

Libo Qin, Zhouyang Li, Wanxiang Che,* Minheng Ni, Ting Liu
Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
{lbqin, zhouyangli, car, mhni, tliu}@ir.hit.edu.cn

Abstract

In a dialog system, dialog act recognition and sentiment classification are two correlative tasks to capture speakers' intentions, where dialog act and sentiment can indicate the explicit and the implicit intentions separately. The dialog context information (*contextual information*) and the *mutual interaction information* are two key factors that contribute to the two related tasks. Unfortunately, none of the existing approaches consider the two important sources of information simultaneously. In this paper, we propose a **Co**-Interactive **G**raph **A**ttention Network (Co-GAT) to jointly perform the two tasks. The core module is a proposed co-interactive graph interaction layer where a *cross-utterances connection* and a *cross-tasks connection* are constructed and iteratively updated with each other, achieving to consider the two types of information simultaneously. Experimental results on two public datasets show that our model successfully captures the two sources of information and achieve the state-of-the-art performance. In addition, we find that the contributions from the contextual and mutual interaction information do not fully overlap with contextualized word representations (BERT, Roberta, XLNet).

Introduction

Dialog act recognition (DAR) and sentiment classification (SC) are two correlative tasks to correctly understand speakers' utterances in a dialog system (Cerisara et al. 2018; Lin, Xu, and Zhang 2020; Qin et al. 2020a). DAR aims to attach semantic labels to each utterance in a dialog which represent the underlying intentions. Meanwhile, SC can detect the sentiments in utterances which can help to capture speakers' implicit intentions. More specifically, DAR can be treated as a sequence classification task that maps the utterances sequence $(u_1, u_2..., u_N)$ to the corresponding utterances sequence DA label $(y_1^d, y_2^d, ..., y_N^d)$, where N is the number of utterances in dialog. Similarly, SC can be also seen as an utterance-level sequence classification problem to predict the corresponding utterances sentiment label $(y_1^s, y_2^s, ..., y_N^s)$.

Intuitively, there are two key factors that contribute to the dialog act recognition and sentiment prediction. One is *the mutual interaction information* across two tasks and the other is *the contextual information* across utterances in a dialogue.

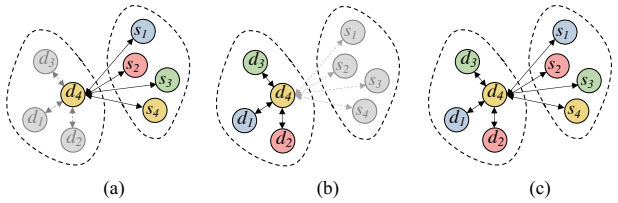


Figure 1: Methods for Joint DAR and SC. Previous work either incorporate *the mutual information* (a), or leverage *the contextual information* (b). Our co-interactive graph interaction method can leverage both the two sources of information (c). s denotes sentiment representation and d denotes dialog act representation.

As illustrated in Figure 2, to predict the sentiment label of `User B`, annotated with `Negative`, its mutual interaction information (i.e. `Agreement` DA label) and contextual information (i.e. sentiment label of `User A`) contribute a lot to the final prediction. The reason is `Agreement` means the `User B` agrees with previous `User A` and hence the `User B` sentiment label tends to be `Negative`, the same with the `User A` sentiment label (`Negative`). Similarly, knowing the mutual sentiment interaction information and the contextual information also contributes to the DA prediction. Thus, it's critical to take the two sources of information into account.

To this end, Cerisara et al. (2018) proposes a multi-task framework to jointly model the two tasks, which can implicitly extract the shared mutual interaction information, but fail to effectively capture the contextual information, which is shown in Figure 1(a). Kim and Kim (2018) explicitly leverage the previous act information to guide the current DA prediction, which captures the contextual information, which is shown in Figure 1(b). However, the model ignores the mutual interaction information, which can be used for promoting the two tasks. Recently, Qin et al. (2020a) propose a *pipeline* method (*DCR-Net*) to incorporate the two types of information. In *DCR-Net*, a hierarchical encoder is proposed to capture the contextual information, followed then by a relation layer to consider the mutual interaction information. Although DCR-Net has achieved good performance, we argue that the *pipeline* method suffers from one major issue:

*corresponding author.

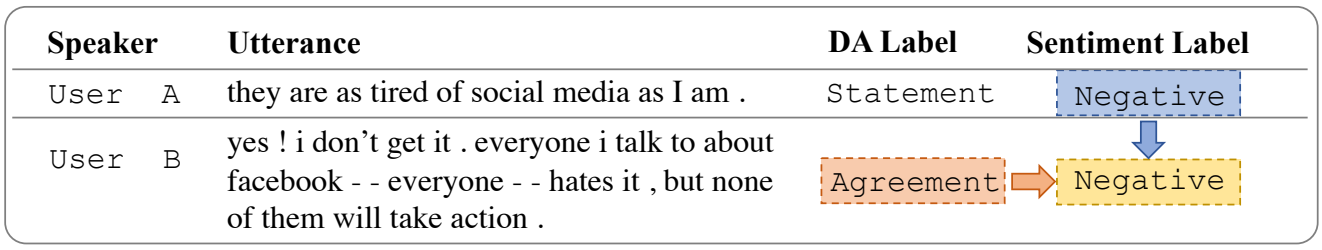


Figure 2: A snippet of a dialog sample from the Mastodon corpus and each utterance has a corresponding DA label and a sentiment label. (DA represents Dialog Act). The blue color segment represents *the contextual information* and the red segment denotes *the mutual interaction information* while the yellow segment represents the target label to predict.

two information are modeled separately. This means the updated process of two types of information are totally isolated, resulting in one type of information can not propagate another type of information in the updated process, which is not effective enough for leveraging knowledge across utterances and tasks. In general, the existing models either consider only one source of information, or employ the above two types of information with *pipeline* modeling method. This leaves us with a question: *Can we simultaneously model the mutual interaction and contextual information in a unified framework to fully incorporate them ?*

Motivated by this, we propose a **Co**-Interactive **G**raph **A**ttention Network (Co-GAT) for joint dialog act recognition and sentiment classification. The core module is a proposed *Co-Interactive* graph interaction layer, which achieves to fully use the aforementioned two sources of information simultaneously. In *Co-Interactive* graph, we perform a dual-connection interaction where a *cross-utterances connection* and a *cross-tasks connection* are constructed and iteratively updated with each other, which is shown in Figure 1(c). More specifically, the *cross-utterances connection*, where each utterance connects other utterances in the same dialog, is used for capturing the contextual information. The *cross-tasks connection*, where node in one task connects all nodes in another task, is used for making an explicit interaction to obtain the mutual interaction information. Further, the *cross-utterance connection* and *cross-task connection* are updated simultaneously and integrated into a unified graph architecture, achieving to answer the proposed question: *each utterance node can be updated simultaneously with the contextual information and mutual interaction information*.

We conduct experiments on two real-world benchmarks including Mastodon dataset (Cerisara et al. 2018) and Dailydialog dataset (Li et al. 2017). Experimental results show that our model achieves significant and consistent improvements as compared to all baseline models by successfully aggregating the mutual interaction information and contextual information. On Mastodon dataset, our model gains 3.0% and 1.9% improvement on F1 score on SC and DAR task, respectively. On Dailydialog dataset, we also obtain 5.6% and 0.3% improvement. In addition, we explore the pre-trained model (BERT, Roberta, XLNet) (Devlin et al. 2019; Liu et al. 2019; Yang et al. 2019) in our framework.

In summary, the main contributions of our work are concluded as follows:

- We make the first attempt to simultaneously incorporate *contextual information* and *mutual interaction information* for joint dialog act recognition and sentiment classification.
- We propose a co-interactive graph attention network where a *cross-tasks connection* and *cross-utterances connection* are constructed and iteratively updated with each other, achieving to model simultaneously incorporate contextual information and mutual interaction information.
- Experiments on two publicly available datasets show that our model obtains substantial improvement and achieves the state-of-the-art performance. In addition, our framework is also beneficial when combined with pre-trained models (BERT, Roberta, XLNet).

To make our experiments reproducible, we will make our code and data publicly available at https://github.com/RaleLee/Co-GAT.

Approach

In this section, we describe the architecture of our framework, as illustrated in Figure 3. It is mainly composed of three components: a shared hierarchical speaker-aware encoder, a stack of co-interactive graph layer to simultaneously incorporate the contextual information and mutual interaction information, and two separate decoders for dialog act and sentiment prediction. In the following paragraph, we first describe the vanilla graph attention network and then the details of other components of framework are given.

Vanilla Graph Attention Network A graph attention network (GAT) (Veličković et al. 2017) is a variant of graph neural network (Scarselli et al. 2009). It propagates features from other neighbourhood's information to the current node and has the advantage of automatically determining the importance and relevance between the current node with its neighbourhood.

In particular, for a given graph with N nodes, one-layer GAT take the initial node features $\tilde{\boldsymbol{H}} = \{\tilde{\boldsymbol{h}}_1, \ldots, \tilde{\boldsymbol{h}}_N\}, \tilde{\boldsymbol{h}}_n \in \mathbb{R}^F$ as input, aiming at producing more abstract representation, $\tilde{\boldsymbol{H}}^{'} = \{\tilde{\boldsymbol{h}}^{'}_1, \ldots, \tilde{\boldsymbol{h}}^{'}_N\}, \tilde{\boldsymbol{h}}^{'}_n \in \mathbb{R}^{F'}$, as its output. The graph attention operated on the node representation can be written as:

$$\tilde{\boldsymbol{h}}^{'}_i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}_h \tilde{\boldsymbol{h}}_j\Big), \quad (1)$$

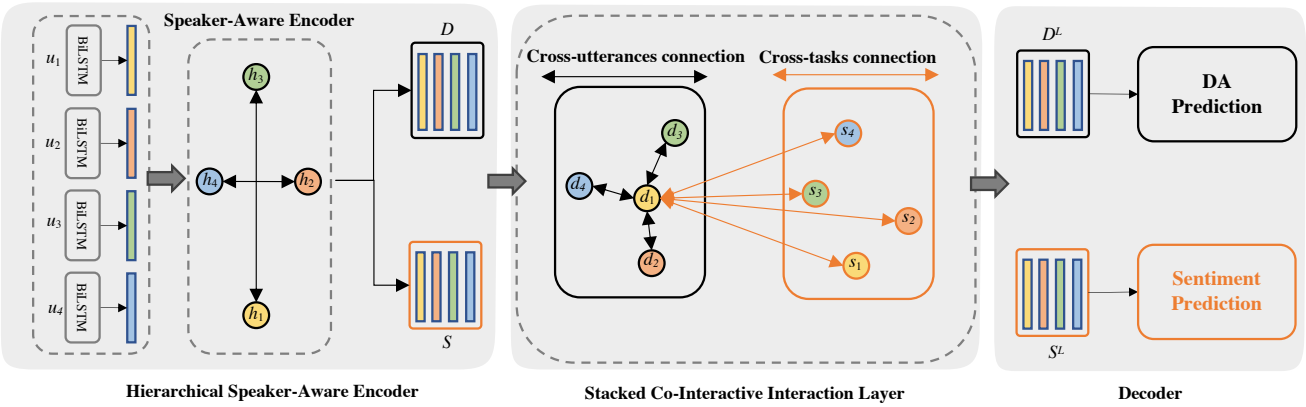


Figure 3: The illustration of our proposed framework, which consists of a hierarchical speaker-aware encoder, a stacked graph-based interaction layer and two separate decoders.

where $\mathcal{N}_i$ is the first-order neighbors of node i (including i) in the graph; F and F' are the input and output dimension; $\boldsymbol{W}_h \in \mathbb{R}^{F' \times F}$ is the trainable weight matrix and σ represents the nonlinearity activation function.

The weight α_{ij} in above equation is calculated via an attention process, which models the importance of each h_j to h_i:

$$\alpha_{ij} = \frac{\exp(\mathcal{F}(\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{h}}_j))}{\sum_{j' \in \mathcal{N}_i} \exp\left(\mathcal{F}(\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{h}}_{j'})\right)}, \tag{2}$$

where $\mathcal{F}$ is an attention function.

In our experiments, following Veličković et al. (2017), the attention function can be formulated as:

$$\mathcal{F}(\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{h}}_j) = \text{LeakyReLU}\left(\mathbf{a}^\top [\boldsymbol{W}_h \tilde{\boldsymbol{h}}_i \| \boldsymbol{W}_h \tilde{\boldsymbol{h}}_j]\right), \tag{3}$$

where $\mathbf{a} \in \mathbb{R}^{2F'}$ is the trainable weight matrix.

In addition, to stabilize the learning process of self-attention, GAT extend the above mechanism to employ *multi-head attention* from (Vaswani et al. 2017):

$$\tilde{\boldsymbol{h}}_i' = \Big\|_{k=1}^{K} \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \tilde{\boldsymbol{h}}_j\Big), \tag{4}$$

where K is the number of heads, α_{ij}^k is the normalized attention weight at k head and $\|$ is concatenation operation and K is the number of heads.

Finally, following Veličković et al. (2017), we employ averaging instead of concatenation to get the final prediction results.

Hierarchical Speaker-Aware Encoder

In our framework, a hierarchical speaker-aware encoder is shared across the dialog act recognition and sentiment classification to leverage the implicit shared knowledge. Specially, it consists of a bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997), which captures temporal relationships within the words, followed by a speaker-aware graph attention network (Veličković et al. 2017) to incorporate the speaker information.

Utterance Encoder with BiLSTM Given a dialog $C = (u_1, ..., u_N)$ consists of a sequence of N utterances and the t-th utterance $u_t = (w_1^t, ..., w_n^t)$ which consists of a sequence of n words, the encoder first maps the tokens in w_i^t to vectors with embedding function ϕ^{emb}. Then, BiLSTM reads it forwardly from w_1^t to w_n^t and backwardly from w_n^t to w_1^t to produce a series of context-sensitive hidden states $\boldsymbol{H} = \{\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, \ldots, \boldsymbol{h}_n^t\}$. Equations are as follows:

$$\overrightarrow{\boldsymbol{h}_i^t} = \overrightarrow{\text{LSTM}}(\phi^{\text{emb}}(w_i^t), \overrightarrow{\boldsymbol{h}_{i-1}^t}), t \in [1, n], \tag{5}$$

$$\overleftarrow{\boldsymbol{h}_i^t} = \overleftarrow{\text{LSTM}}(\phi^{\text{emb}}(w_i^t), \overleftarrow{\boldsymbol{h}_{i+1}^t}), t \in [n, 1], \tag{6}$$

$$\boldsymbol{h}_i^t = [\overrightarrow{\boldsymbol{h}_i^t}, \overleftarrow{\boldsymbol{h}_i^t}]. \tag{7}$$

Then, the last hidden state $\boldsymbol{h}_n^t$ can be seen as the utterance u_t representation $\boldsymbol{e}_t$ (i.e., $\boldsymbol{e}_t = \boldsymbol{h}_n^t$). Hence, the sequentially encoded feature of N utterances in C can be represented as $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N)$.

Speaker-Level Encoder We propose to use a speaker-aware graph attention network to leverage the speaker information, which enables the model to better understand how the emotion and act intention change within the same speaker (Ghosal et al. 2019). We build graphical structures over the input utterance sequences to explicitly incorporate the speaker information into the graph attention network, and construct the graph in the following way,

Vertices: Each utterance in the conversation is represented as a vertex. Each vertex is initialized with the corresponding sequentially encoded feature vector $\boldsymbol{e}_i$, for all $i \in [1, 2, ..., N]$. We denote this vector as the vertex feature. Hence, the first layer states vector for all nodes is $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N)$.

Edges: Since we aim to model the speaker information in a dialog explicitly, vertex i and vertex j should be connected if they belong to the same speaker. More specifically, $\boldsymbol{A}$ is an adjacent matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ with $\boldsymbol{A}_{ij} = 1$ if they're from the same speaker and $\boldsymbol{A}_{ij} = 0$ otherwise[1], By doing this, the

[1]In our paper, we only consider the first-order neighbors to alleviate the overfitting problem.

speaker features can be propagated from neighbour nodes to the current node. In particular, the aggregation process can be rewritten as:

$$\tilde{\boldsymbol{e}}_i' = \mathop{\Big\|}_{k=1}^{K} \sigma\big(\sum_{j\in\mathcal{S}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \tilde{\boldsymbol{e}}_j\big), \quad (8)$$

where $\mathcal{S}_i$ represents the nodes that belong to the same speaker with i node.

After stacking m layer, we obtain the speaker-aware encoding features $\boldsymbol{E}^m = (\boldsymbol{e}_1^m, \ldots, \boldsymbol{e}_N^m)$. Following Qin et al. (2020a), we first apply separate BiLSTM over act information and sentiment information separately to make them more task-specific, which can be written as $\boldsymbol{D}^0 = \text{BiLSTM}\ (\boldsymbol{E}^m)$ and $\boldsymbol{S}^0 = \text{BiLSTM}\ (\boldsymbol{E}^m)$. $\boldsymbol{D}^0 = (\boldsymbol{d}_1^0, \ldots, \boldsymbol{d}_N^0)$ and $\boldsymbol{S}^0 = (\boldsymbol{s}_1^0, \ldots, \boldsymbol{s}_N^0)$ can be seen as the initial shared representations of dialog act and sentiment.

Stacked Co-Interactive Graph Layer

One core advantage of our framework is modeling the contextual information and mutual interaction information into a unified graph interaction architecture and updating them simultaneously. Specially, we adopt a graph attention network (GAT) to model the interaction process with the *cross-tasks connection* and *cross-utterances connection*. Graph interaction structure has been shown effective on various of NLP tasks (Lu and Li 2020; Chai and Wan 2020; Qin et al. 2020c,b). We construct the graph in the following way,

Vertices: Since we model the interaction between the two tasks in graph architecture, we have $2N$ nodes in the graph where N nodes for sentiment classification task and the other N nodes for dialog act recognition task. We use the speaker-aware encoding representation $\boldsymbol{D}^0$ and $\boldsymbol{S}^0$ to initialize our corresponding sentiment and dialog act node vertices, respectively. Thus, we obtain the initialization node representation $\boldsymbol{H}^0 = [\boldsymbol{D}^0; \boldsymbol{S}^0] = [\boldsymbol{d}_1^0, \ldots, \boldsymbol{d}_N^0, \boldsymbol{s}_1^0, \ldots, \boldsymbol{s}_N^0] \in \mathbb{R}^{2N\times d}$, where d represents the dimension of vertice representation.

Edges: In the graph, there exist two types of edges.

cross-utterances connection: We construct the *cross-utterances connection* where node i should connect to its context utterance node to take the contextual information into account. More specifically, we denote the graph adjacent matrix as $\boldsymbol{A} \in \mathcal{R}^{2N\times 2N}$, where the $\boldsymbol{A}_{i,j}^I = 1$ if they are in the same dialogue.

cross-tasks connection: The *cross-tasks connection* is constructed where node i connects to all another task node to explicitly leverage the mutual interaction information, where $\boldsymbol{A}_{i,j}^I = 1$ when they belongs to the different tasks.

By doing this, we model the two sources of information in a unified graph interaction framework with *cross-utterances connection* and *cross-tasks connection*. In particular, we use $\boldsymbol{d}_i^{(l)}$ and $\boldsymbol{s}_z^{(l)}$ to represent dialog act representation of node i and sentiment representation of node z in the l-th layer of the graph, respectively. For $\boldsymbol{d}_i^{(l)}$, the graph interaction update process can be formulated as:

$$\boldsymbol{d}_i^{(l+1)} = \mathop{\Big\|}_{k=1}^{K} \sigma\big(\sum_{j\in\mathcal{D}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{d}_j^{(l)} + \sum_{j\in\mathcal{A}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{s}_j^{(l)}\big), \quad (9)$$

where $\sum_{j\in\mathcal{D}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{d}_j^{(l)}$ is the *cross-utterances connection* to integrate the contextual information while $\sum_{j\in\mathcal{A}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{s}_j^{(l)}$ denotes the *cross-tasks connection* for incorporating the mutual interaction information.

Similarly, the graph interaction update process for $\boldsymbol{s}_i^{(l)}$ can be formulated as:

$$\boldsymbol{s}_i^{(l+1)} = \mathop{\Big\|}_{k=1}^{K} \sigma\big(\sum_{j\in\mathcal{A}_i'} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{s}_j^{(l)} + \sum_{j\in\mathcal{D}_i'} \alpha_{ij}^k \boldsymbol{W}_h^k \boldsymbol{d}_j^{(l)}\big). \quad (10)$$

Decoder for Dialog Act Recognition and Sentiment Classification

In order to learn deep features, we apply a stacked graph attention network with multiple layers. After stacking L layer, we obtain a final updated feature representation $\boldsymbol{E}^L = [\boldsymbol{D}^L; \boldsymbol{S}^L]$ including: $\boldsymbol{D}^L = (\boldsymbol{d}_1^L, ..., \boldsymbol{d}_N^L)$ and $\boldsymbol{S}^L = (\boldsymbol{s}_1^L, ..., \boldsymbol{s}_N^L)$. Then, we perform linear transform and LSTM upon the $\boldsymbol{S}^L$ and $\boldsymbol{D}^L$ to make the representation more task-specific, where the $\boldsymbol{S}^{L'} = \text{Linear}\ (\boldsymbol{S}^L)$ and $\boldsymbol{D}^{L'} = \text{LSTM}\ (\boldsymbol{D}^L)$. We then adopt separate decoder to perform dialog act and sentiment prediction, which can be denoted as follows:

$$\boldsymbol{y}_t^d = \text{softmax}(\boldsymbol{W}^d \boldsymbol{d}_t^{L'} + \boldsymbol{b}_d), \quad (11)$$

$$\boldsymbol{y}_t^s = \text{softmax}(\boldsymbol{W}^s \boldsymbol{s}_t^{L'} + \boldsymbol{b}_s), \quad (12)$$

where $\boldsymbol{y}_t^d$ and $\boldsymbol{y}_t^s$ are the predicted distribution for dialog act and sentiment respectively; $\boldsymbol{W}^d$ and $\boldsymbol{W}^s$ are transformation matrices; $\boldsymbol{b}_d$ and $\boldsymbol{b}_s$ are bias vectors.

Joint Training

The dialog act recognition and sentiment classification objective are formulated as:

$$\mathcal{L}_1 \triangleq -\sum_{i=1}^{N}\sum_{j=1}^{N_S} \hat{y}_i^{(j,s)} \log\Big(y_i^{(j,s)}\Big), \quad (13)$$

$$\mathcal{L}_2 \triangleq -\sum_{i=1}^{N}\sum_{j=1}^{N_D} \hat{y}_i^{(j,d)} \log\Big(y_i^{(j,d)}\Big), \quad (14)$$

where $\hat{y}_i^d$ and $\hat{y}_i^s$ are gold act label and gold sentiment label separately; N_D is the number of dialog act labels; N_S is the number of sentiment labels and N is the number of utterances.

Following Qin et al. (2019), the dialog act recognition and sentiment classification can be considered jointly, the final joint objective is:

$$\mathcal{L}_\theta = \mathcal{L}_1 + \mathcal{L}_2. \quad (15)$$

Experiments

Datasets

We conduct experiments on the benchmark Dailydialog (Li et al. 2017) and Mastodon (Cerisara et al. 2018). On Dailydialogues dataset, we follow the same format and partition as in Li et al. (2017). The dataset contains 11,118 dialogues

Model	Mastodon						Dailydialog					
	SC			DAR			SC			DAR		
	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)
HEC (Kumar et al. 2018)	-	-	-	56.1	55.7	56.5	-	-	-	77.8	76.5	77.8
CRF-ASN (Chen et al. 2018)	-	-	-	55.1	53.9	56.5	-	-	-	76.0	75.6	78.2
CASA (Raheja and Tetreault 2019)	-	-	-	56.4	57.1	55.7	-	-	-	78.0	76.5	77.9
DialogueRNN (Majumder et al. 2019)	41.5	42.8	40.5	-	-	-	40.3	37.7	44.5	-	-	-
DialogueGCN (Ghosal et al. 2019)	42.4	43.4	41.4	-	-	-	43.1	44.5	41.8	-	-	-
JointDAS (Cerisara et al. 2018)	37.6	41.6	36.1	53.2	51.9	55.6	31.2	28.8	35.4	75.1	74.5	76.2
IIIM (Kim and Kim 2018)	39.4	40.1	38.7	54.3	52.2	56.3	33.0	28.5	38.9	75.7	74.9	76.5
DCR-Net + Co-Attention (Qin et al. 2020a)	45.1	47.3	43.2	58.6	56.9	60.3	45.4	40.1	56.0	79.1	79.0	79.1
Our model	**48.1***	**53.2***	**44.0***	**60.5***	**60.6***	**60.4**	**51.0***	**45.3***	**65.9***	**79.4**	78.1	**81.0***

Table 1: Comparison of our model with baselines on Mastodon and Dailydialog datasets. SC represents Sentiment Classification and DAR represents Dialog Act Recognition. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.05$ under t-test.

for training, 1,000 for validation and 1,000 for testing. On Mastodon dataset, it includes 239 dialogues for training, 266 dialogues for testing. We keep the train/test partition unchanged.[2]

Experimental Settings

In our experiment setting, dimensionality of all hidden units are 256. And the dimensionality of the embedding is 800 and 128 for Mastodon and Dailydialog, respectively. L2 regularization used on our model is 1×10^{-8}. In addition, we add a residual connection in graph attention network layer for reducing overfitting. We use Adam (Kingma and Ba 2014) to optimize the parameters in our model and adopt the suggested hyper-parameters for optimization. We set the stacked number of GAT as 2 on Mastdon dataset and 3 on Dailydialog dataset. For all experiments, we pick the model which works best on the dev set, and then evaluate it on the test set. All experiments are conducted at GeForce RTX 2080Ti. The epoch number is 300 and 100 for Mastodon and Dailydialog, respectively.

Baselines

We compare our model with several of state-of-the-art baselines including: 1) the separate dialog act recognition models: HEC, CRF-ASN and CASA. 2) the separate sentiment classification models: DialogueGCN and DialogueRNN. 3) the joint models including: JointDAS, IIIM and DCR-Net. We briefly describe these baseline models below: 1) **HEC** (Kumar et al. 2018): This work uses a hierarchical Bi-LSTM-CRF model for dialog act recognition, which capture both kinds of dependencies including word-level and utterance-level. 2) **CRF-ASN** (Chen et al. 2018): This model proposes a crf-attentive structured network for dialog act recognition, which dynamically separates the utterances into cliques. 3) **CASA** (Raheja and Tetreault 2019): This work leverages a context-aware self-attention mechanism coupled with a hierarchical deep neural network. 4) **DialogueRNN** (Majumder et al. 2019): This model proposes a RNN-based neural architecture for emotion detection in a conversation to keep track of the individual party states throughout the conversation and uses this information. 5) **DialogGCN** (Ghosal et al. 2019): This model proposes a dialogue graph convolutional network to leverage self and inter-speaker dependency of the interlocutors to model conversational context for emotion recognition 6) **JointDAS** (Cerisara et al. 2018): This model uses a multi-task modeling framework for joint dialog act recognition and sentiment classification. 7) **IIIM** (Kim and Kim 2018): This work proposes an integrated neural network model which simultaneously identifies speech acts, predicators, and sentiments of dialogue utterances. 8) **DCR-Net**: This model proposes a relation layer to explicitly model the interaction between the two tasks and achieves the state-of-the-art performance.

[2]The two datasets are available in http://yanran.li/dailydialog and https://github.com/cerisara/DialogSentimentMastodon

Model	Mastodon		Dailydialog	
	SC (F1)	DAR (F1)	SC (F1)	DAR (F1)
without cross-tasks connection	46.1	58.1	49.7	78.2
without cross-utterances connection	44.9	58.7	48.1	78.2
separate modeling	46.7	58.4	45.6	78.3
co-attention mechanism	46.5	59.4	46.2	79.1
without speaker information	46.4	59.0	47.6	79.2
Our model	**48.1**	**60.5**	**51.0**	**79.4**

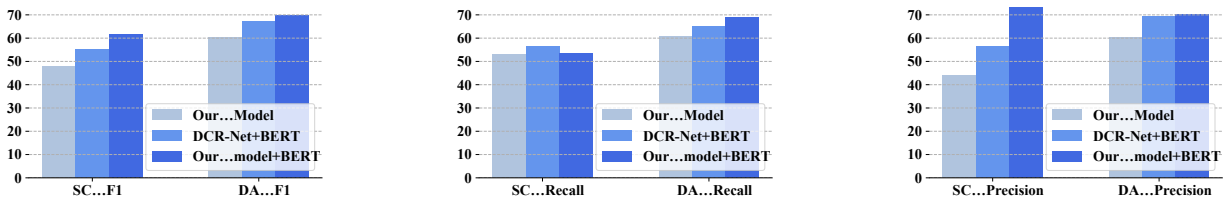
Table 2: Ablation study on Mastodon and Dailydialog test datasets.

Overall Results

Following Kim and Kim (2018); Cerisara et al. (2018); Qin et al. (2020a), we adopt macro-average Precision, Recall and F1 for both sentiment classification and dialog act recognition on Dailydialog dataset and we adopt the average of the dialog-act specific F1 scores weighted by the prevalence of each dialog act on Mastodon dataset.

The experimental results are shown in Table 3. The first block of table represents the separate model for dialog act recognition task while the second block denotes the separate model for sentiment classification task. The third block of table represents the state-of-the-art joint models for the two task. From the results, we can observe that:

1. Our framework outperforms the state-of-the-art dialog act recognition and sentiment classification models which trained in separate task in all metrics on two datasets. This shows that our proposed graph interaction model has incorporated the mutual interaction information between the two tasks which can be effectively utilized for promoting performance mutually.



(a) (b) (c)

Figure 4: The performance of BERT-based model on Mastodon.F1 scores are shown in (a). Recall results are shown in (b) and (c) shows the precision results.

Model	Mastodon					
	SC			DAR		
	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)
RoBERTa+Linear	55.7	54.4	59.7	61.6	61.8	61.4
Co-GAT+RoBERTa	61.3	58.8	64.3	66.1	64.8	67.5
XLNet+Linear	58.7	60.9	56.6	62.6	61.8	63.4
Co-GAT+XLNet	65.9	65.8	66.1	67.5	66.0	69.2

Table 3: Results on the pre-trained models.

2. We obtain large improvements compared with the state-of-the-art joint models. On Mastodon dataset, compared with *DCR-Net* model, our framework achieves 3.0% improvement on F1 score on sentiment classification task and 1.9% improvement on F1 score on dialog act recognition task. On Dailydialog dataset, the same trend has been observed. This demonstrates the effectiveness of simultaneously leveraging contextual information and the mutual interaction information with graph-interaction method, compared with *DCR-Net* which separate considers the two types of information.

Analysis

Although achieving good performance, we would like to know the reason for the improvement. In this section, we study our model from several directions. We first conduct several ablation experiments to analyze the effect of different components in our framework, including the effect of mutual interaction and contextual information. Then, we analyze the effect of simultaneous modeling method. Next, we incorporate and analyze the pre-trained model (BERT, RoBERTa, XLNet) in our framework.

Effectiveness of the Mutual Interaction Information In this setting, when constructing the graph architecture for graph interaction, we only consider the *cross-utterances connection* by removing the edges connecting from one task to another task, which can be seen as ignoring the mutual interaction information. We name it as *without cross-tasks connection* and the result is shown in Table 2. We can see 2.0% and 1.3% drop in terms of F1 scores in sentiment classification while 2.4% and 1.2% drop in dialog act recognition on Mastodon and Dailydialog dataset, respectively. We attribute it to the fact that explicitly modeling the interaction between two tasks with graph-interaction layer can encourage model to effectively utilize the information of one task for another task.

Effectiveness of the Contextual Information Similarly, when constructing the graph architecture for graph-interaction, we only consider the *cross-tasks connection* by removing the edges connecting from one node to its contextual node, which can be seen as ignoring the contextual information. We name it as *without cross-utterances connection* and the result is shown in Table 2. The results show a significant drop in performance, which verifies the effectiveness of contextual information. The reason is that contextual information help reduce ambiguity, which improves performance.

Simultaneous Modeling vs. Separate Modeling To verify the effectiveness of simultaneously modeling the two sources of information in a unified co-interactive graph interaction mechanism, we remove the co-interactive interaction layer and only use two separate sub GAT to represent the *cross-utterance connection* and *cross-task connection* to model the two tasks separately and adopt the sum operation based on the output of GAT to consider their interaction. We refer it as *separate modeling* and the result is shown in Table 2 and the results show a significant drop in performance. This indicates that modeling the two sources of information with a co-interactive graph interaction mechanism can better incorporate information simultaneously compared with model the two types of information separately.

In particular, *DCR-Net* can be seen as the SOTA *pipeline* method. To make a more fair comparison with *DCR-Net*, we replace the co-interactive interaction layer with co-attention mechanism in *DCR-Net* and we keep other components unchanged. We name it as *co-attention mechanism*. The results are shown in Table 2 and we can see that our framework outperforms the *co-attention mechanism* by a large margin. This again demonstrates that simultaneously modeling the contextual information and interaction information by proposed co-interactive graph interaction mechanism is effective than the *pipeline* model to incorporate two types of information in *DCR-Net*.

Effectiveness of Speaker Information In this settings, we remove the speaker-aware encoder and only keep the BiLSTM encoder as the same. We refer it as *without speaker information* and the result is shown in Table 2. From the result, we can see that 1.7% and 3.4% drop in terms of F1 scores in sentiment classification while 1.5% and 0.2% drop in dialog act recognition on two datasets. On Dailydialog dataset, we can also observe the same trends that the F1 score drops a lot. This demonstrates that properly modeling the

speaker information can help model to capture the sentiment and act flow in a dialog, which can enhance their performance. It is noticeable that even without the speaker-aware encoder, our framework still performs the state-of-the-art *DCR-Net* model , which again demonstrates the effectiveness and robustness of our framework.

Effectiveness of Pre-trained Model Finally, following Qin et al. (2020a), we also explore the pre-trained model, BERT (Devlin et al. 2019) in our framework. In this section, we replace the hierarchical speaker-aware encoder by BERT base model [3] and keep other components as same with our framework. We conduct experiments on Mastodon dataset and the results are shown in Figure 4. From the results, we can observe: 1) the BERT-based model performs remarkably well and achieves a new state-of-the-art performances. This indicates that the performance can be further improved a lot with the pre-trained model and our framework works orthogonally with BERT. We attribute this to the fact that pre-trained models can provide rich semantic features, which can improve the performance on both two tasks. 2) Our BERT-based model outperforms the baseline (*DCR-Net* + BERT), which again verifies the effectiveness of our proposed co-interactive graph interaction framework.

In addition, to further verify the contribution from our proposed model is still effective over the strong pre-trained model, we perform experiments with Roberta and XLNet. To further verify that our contribution from Co-GAT does not fully overlap with contextualized word representations (Roberta, XLNet), we have conducted the following experiments on Mastodon dataset:

1) Roberta/XLNet+Linear. In this setting, we adopt the Roberta/XLNet model as the shared encoder and add two different linear decoders for SC and DAR task.

2) Co-GAT + Roberta/XLNet. Here, we replace the hierarchical speaker-aware encoder by Roberta/XLNet model and keep other components as same with our framework. The Roberta/XLNet is fine-tuned in our experiment.

Results are shown in Table 3. From the results, we find that the integration of Co-GAT and Roberta/XLNet can further improve the performance, demonstrating that contributions from the two are complementary.

Related Work

Dialog Act Recognition

Kalchbrenner and Blunsom (2013) propose a hierarchical CNN to model the context information for DAR. Lee and Dernoncourt (2016) propose a model which combine the advantages of CNNs and RNNs and incorporated the previous utterance as context to classify the current for DAR. Ji, Haffari, and Eisenstein (2016) use a hybrid architecture, combining an RNN language model with a latent variable model. Furthermore, many work (Liu et al. 2017; Kumar et al. 2018; Chen et al. 2018) explore different architectures to better incorporate the context information for DAR. Raheja and Tetreault (2019) propose the context-aware self-attention mechanism for DAR and achieve the promising performance.

[3]The BERT is fine-tuned in our experiment.

Sentiment Classification

Sentiment classification in dialog system can be seen as the sentence-level sequence classification problem. One series of works are based on CNN (Zhang, Zhao, and LeCun 2015; Conneau et al. 2017; Johnson and Zhang 2017) to capture the local correlation and position-invariance. Another series of works adopt RNN based models (Tang, Qin, and Liu 2015; Yang et al. 2016; Xu et al. 2016) to capture temporal features for sentiment classification. Besides, Some works (Xiao and Cho 2016; Shi et al. 2016; Wang 2018) combine the advantages of CNN and RNN. Recently, Majumder et al. (2019) introduce a party state and global state based recurrent model for modeling the emotional dynamics. Majumder et al. (2019) propose a dialogGCN which leverages self and inter-speaker dependency of the interlocutors to model context and achieve the state-of-the-art performance.

Joint Model

Considering the correlation between dialog act recognition and sentiment classification, many joint models are proposed to consider the interaction between the two tasks. Cerisara et al. (2018) explore the multi-task framework to model the correlation between the two tasks. Kim and Kim (2018) propose an integrated neural network for identifying dialog act, predicators, and sentiments of dialogue utterances. Compared with their work, our framework simultaneously considers the contextual information and mutual interaction information into a unified graph interaction architecture. In contrast, their models only consider on type of information (contextual information or mutual interaction information). More recently, Qin et al. (2020a) propose a DCR-Net which adopts a relation layer to model the relationship and achieve the state-of-the-art performance. This model can be regarded as the *pipeline* method to model the contextual and mutual interaction information, which ignores the contextual information when performing interaction between the two tasks. In contrast, we propose a co-interactive graph attention network where *cross-utterances connection* and *cross-tasks connection* are constructed and iteratively updated with each other to simultaneously model the contextual information and the mutual interaction information into a unified graph structure. To the best of our knowledge, we are the first to simultaneously model the mutual information and contextual information in a unified graph interaction architecture

Conclusion

In this paper, we propose a co-interactive graph framework where a *cross-utterances connection* and a *cross-tasks connection* are constructed and iteratively updated with each other, achieving to simultaneously model the contextual information and mutual interaction information in a unified architecture. Experiments on two datasets show the effectiveness of the proposed models and our model achieves state-of-the-art performance. In addition, we analyze the effect of incorporating strong pre-trained model in our joint model and find that our framework is also beneficial when combined with pre-trained models (BERT, Roberta, XLNet).

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 61772153. This work was also supported by the Zhejiang Lab's International Talent Fund for Young Professionals.

References

Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; and Le, H. T. 2018. Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, 745–754. Santa Fe, New Mexico, USA: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1063.

Chai, Z.; and Wan, X. 2020. Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 225–237. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.21. URL https://www.aclweb.org/anthology/2020.acl-main.21.

Chen, Z.; Yang, R.; Zhao, Z.; Cai, D.; and He, X. 2018. Dialogue act recognition via crf-attentive structured network. In *Proc. of SIGIR*.

Conneau, A.; Schwenk, H.; Barrault, L.; and Lecun, Y. 2017. Very Deep Convolutional Networks for Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1107–1116. Valencia, Spain: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1104.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1015. URL https://www.aclweb.org/anthology/D19-1015.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* .

Ji, Y.; Haffari, G.; and Eisenstein, J. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913* .

Johnson, R.; and Zhang, T. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 562–570. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-1052. URL https://www.aclweb.org/anthology/P17-1052.

Kalchbrenner, N.; and Blunsom, P. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584* .

Kim, M.; and Kim, H. 2018. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. *Pattern Recognition Letters* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kumar, H.; Agarwal, A.; Dasgupta, R.; and Joshi, S. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proc. of AAAI*.

Lee, J. Y.; and Dernoncourt, F. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827* .

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1099.

Lin, T.-E.; Xu, H.; and Zhang, H. 2020. Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Liu, Y.; Han, K.; Tan, Z.; and Lei, Y. 2017. Using Context Information for Dialog Act Classification in DNN Framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2170–2178. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/D17-1231. URL https://www.aclweb.org/anthology/D17-1231.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 505–514. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.48. URL https://www.aclweb.org/anthology/2020.acl-main.48.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qin, L.; Che, W.; Li, Y.; Ni, M.; and Liu, T. 2020a. DCR-Net: A Deep Co-Interactive Relation Network for Joint Dialog Act Recognition and Sentiment Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qin, L.; Che, W.; Li, Y.; Wen, H.; and Liu, T. 2019. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proc. of EMNLP*.

Qin, L.; Ni, M.; Zhang, Y.; Che, W.; Li, Y.; and Liu, T. 2020b. Multi-Domain Spoken Language Understanding Using Domain-and Task-Aware Parameterization. *arXiv preprint arXiv:2004.14871* .

Qin, L.; Xu, X.; Che, W.; and Liu, T. 2020c. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1807–1816. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.163. URL https://www.aclweb.org/anthology/2020.findings-emnlp.163.

Raheja, V.; and Tetreault, J. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proc. of NAACL*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1): 61–80.

Shi, Y.; Yao, K.; Tian, L.; and Jiang, D. 2016. Deep LSTM based Feature Mapping for Query Classification. In *Proc. of NAACL*.

Tang, D.; Qin, B.; and Liu, T. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proc. of ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Proc. of NIPS*. Curran Associates, Inc.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Wang, B. 2018. Disconnected Recurrent Neural Networks for Text Categorization. In *Proc. of ACL*.

Xiao, Y.; and Cho, K. 2016. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367* .

Xu, J.; Chen, D.; Qiu, X.; and Huang, X. 2016. Cached Long Short-Term Memory Neural Networks for Document-Level Sentiment Classification. In *Proc. of EMNLP*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proc. of NAACL*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Proc. of NIPS*, 649–657.