Integrating Tree Structures and Graph Structures with Neural Networks to Classify Discussion Discourse Acts

Yasuhide Miura[†,‡]
yasuhide.miura@fujixerox.co.jp

Ryuji Kano[†]
kano.ryuji@fujixerox.co.jp

Motoki Taniguchi[†]
motoki.taniguchi@fujixerox.co.jp

Tomoki Taniguchi[†]
taniguchi.tomoki@fujixerox.co.jp

Shotaro Misawa[†]
misawa.shotaro@fujixerox.co.jp

Tomoko Ohkuma[†]
ohkuma.tomoko@fujixerox.co.jp

[†]Fuji Xerox Co., Ltd.
[‡]Tokyo Institute of Technology

Abstract

We proposed a model that integrates discussion structures with neural networks to classify discourse acts. Several attempts have been made in earlier works to analyze texts that are used in various discussions. The importance of discussion structures has been explored in those works but their methods required a sophisticated design to combine structural features with a classifier. Our model introduces tree learning approaches and a graph learning approach to directly capture discussion structures without structural features. In an evaluation to classify discussion discourse acts in Reddit, the model achieved improvements of 1.5% in accuracy and 2.2 in F_1 score compared to the previous best model. We further analyzed the model using an attention mechanism to inspect interactions among different learning approaches.

1 Introduction

With the recent growth of social news sites and debate portals, people today discuss various topics online. These discussions include valuable public opinions of crowds. However, automated analyses of them are often difficult, requiring understanding of textual contents and discussion structures. By extending approaches taken in the analysis of spoken dialogue acts (Stolcke et al., 2000; Bunt et al., 2010), there have been a number of attempts to analyze discussions with discourse acts. Discussions in emails (Cohen et al., 2004; Carvalho and Cohen, 2005; Carvalho and Cohen, 2006; Hu et al., 2009; Omuya et al., 2013), newsgroups (Wang et al., 2007), technical forums (Kim et al., 2010b; Wang et al., 2011; Bhatia et al., 2012; Liu et al., 2017) and social news (Zhang et al., 2017) are targeted in earlier studies. The automatic classification of these discourse acts can improve tasks like information access and summarization for discussions.

The importance of discussion structures for analyzing discourse acts has already been recognized in earlier studies. Relational features (Carvalho and Cohen, 2005), link features (Hu et al., 2009), and structural features (Wang et al., 2007; Kim et al., 2010a; Kim et al., 2010b; Wang et al., 2011; Bhatia et al., 2012; Zhang et al., 2017; Liu et al., 2017) are combined with a probabilistic graphical model, a structured prediction model, or sequential classification models. These approaches have achieved promising results in classifying discourse acts, but they require a sophisticated design to integrate structural features into a classification model. In this paper, we propose a model that integrates discussion structures with neural networks to classify discourse acts. Recently, neural networks have shown their effectiveness at capturing tree structures (Socher et al., 2011; Socher et al., 2014; Tai et al., 2015) and graph structures (Defferrard et al., 2016; Kipf and Welling, 2017). Our model introduces tree learning approaches and a graph learning approach to directly capture discussion structures without structural features to classify discourse acts.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3806–3818
Santa Fe, New Mexico, USA, August 20-26, 2018.

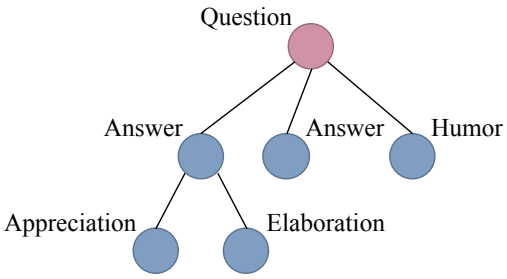

Figure 1: An example of a Reddit thread with discourse acts. The circles represent comments. An initial question is replied by answers and a humor. One of the answer is further replied by an appreciation and an elaboration.

Among the various targets that have been analyzed in earlier studies, we apply our models to the analysis of Reddit[1] discussions. Reddit is among the largest social news sites, hosting discussions of various topics. We aim to classify comments into 9 discourse acts of Zhang et al. (2017), namely *Answer*, *Elaboration*, *Question*, *Appreciation*, *Agreement*, *Disagreement*, *Humor*, *Announcement*, and *Negative Reaction*. Figure 1 illustrates an example of a Reddit thread with discourse acts. We chose Reddit discussions for two reasons. First, threads that comprise Reddit discussions have an explicit tree structure with reply relations as edges. Secondly, the publicly available dataset is larger than other datasets, which enables neural networks to learn tree structures and graph structures. We demonstrate that our model successfully learns discussion structures with tree learning approaches and a graph learning approach of neural networks.

The following are the contributions of this paper:

1. We propose a neural model with tree learning components and a graph learning component that can efficiently learn discussion structures.
2. We show that the proposed model outperforms earlier sequential learning approaches in a discourse act classification.
3. We analyze some components of the proposed model to ascertain the learned effective discussion structures.

In subsequent sections of the paper, we present related works in two criteria in Section 2. Section 3 describes the proposed model. Details of an experiment are reported in Section 4, with discussions in Section 5. Finally, Section 6 concludes the paper with a presentation of some future directions.

2 Related Works

2.1 Discourse Act Classification

The analysis of discussions attracted researchers to propose automated approaches to classify discourse acts. In earlier works, emails and forums were popularly targeted by several works. Cohen et al. (2004) annotated emails with speech acts and built a classifier with textual features for them. This work was extended by Carvalho and Cohen (2005) to combine relational features with a dependency-network-based collective classification model and by Carvalho and Cohen (2006) with an exploitation of n-gram features. Hu et al. (2009) also annotated emails with speech acts and trained a structured prediction classifier. Omuya et al. (2013) extended this work using per-class feature optimization and a cascade of classifiers. Kim et al. (2010b) tagged technical forums with dialogue acts and conducted an experiment on them using structural features and a sequential learner. Later, these dialogue acts were tackled by Wang et al. (2011) with a joint classification approach of discourse acts and link relations, and by Liu et al. (2017) with a sequential learner including an external memory. Bhatia et al. (2012) assigned dialogue acts to technical forums and evaluated classifiers using a variety of features including structural features.

The discourse acts of other targets were also explored in earlier works. Wang et al. (2007) evaluated a sequential learning technique with structural features of newsgroup argument codes. Kim et al. (2010a)

[1] https://www.reddit.com/

explored the classification of dialogue acts in chats with various features including structural ones with a sequential learner. Zhang et al. (2011) annotated tweets with speech acts and trained a classification model with word features and character features. These speech acts were examined further in Zhang et al. (2012) with semi-supervised learners including a graph-based label propagation. Ferschke et al. (2012) created a Wikipedia Talks corpus with dialog acts and trained a binary classifier with textual features and talk turn features. Zhang et al. (2017) annotated Reddit threads with discourse acts and designed sequential models with various features including structural features.

2.2 Reddit Analysis

Reddit, which has become a popular target for researchers for analyses of public opinion, consists of massive textual contents and visual contents with metadata that are readily accessible via an API. Within studies using Reddit as their data, automated analysis of the *karma score*, a post-level popularity of discussions in Reddit based on positive votes and negative votes of users, is widely studied. Jaech et al. (2015) proposed a karma score ranking task and used a pairwise classifier to rank them. Wei et al. (2016) designed a system that ranks comments based on karma scores as a reference to their persuasiveness. He et al. (2016) proposed a deep reinforcement learning architecture for the effective modeling of an online popularity prediction task. He et al. (2017) extended the model using a two-stage approach. Hessel et al. (2017) experimented on a relative popularity task using karma scores and demonstrated the effectiveness of multimodal features. Cheng et al. (2017) introduced a factored neural model that demonstrated improvements on predicting karma scores over standard document embedding methods. Zayats and Ostendorf (2018) presented a graph-structured neural architecture that outperformed a node-independent architecture in predicting karma scores.

Automatic analysis of targets other than karma scores were also investigated in earlier works analyzing Reddit discussions. The analysis of discourse acts (Zhang et al., 2017) is one such work, but some other attempts were also made. Buntain and Golbeck (2014) demonstrated the presence of answer-person roles and their identification with a supervised classifier. Tan et al. (2016) examined *ChangeMyView* subreddit to clarify the mechanism of persuasion and built a classifier for opinion malleability prediction. Lim et al. (2017) explored the estimation of relative user expertise through various content-agnostic approaches. Habernal et al. (2018) investigated ad hominem arguments and experimented with neural models for prediction.

2.3 Comparisons with Our Model

Our model, which we describe in Section 3 learns a discussion structure using tree learning approaches and a graph learning approach. The model processes textual information and structural information jointly within its architecture without structural features. For the classification of discourse acts, most earlier works used sequential learning approaches (Wang et al., 2007; Kim et al., 2010a; Kim et al., 2010b; Wang et al., 2011; Zhang et al., 2017; Liu et al., 2017). Carvalho and Cohen (2005) used a probabilistic graphical model, but it relied on relational features and pre-trained classifier trained with textual features. When we compare our model with Reddit analysis models, a close approach is taken in the model of Zayats and Ostendorf (2018). They extended Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) to accommodate graph structures to capture hierarchical and temporal conversations. However, their model includes designs for efficiently learning the popularity of Reddit. The popularity of a comment in Reddit is known to be associated strongly with the post timing and the post author (Jaech et al., 2015). We present in an experiment explained in Section 4 that a simple application of a popularity prediction model will not perform well on classifying discourse acts.

3 Model

Figure 2 illustrates the overview of our proposed model: Tree-LSTM GCN Hybrid. The model first encodes comments in a thread with an LSTM and max-pooling (Comment Encoder) to comment representations. The comment representations are then updated with tree-level LSTM processes (Parent-Branch Tree-LSTM and Child-Sum Tree-LSTM) and a graph-level convolutional networks process (GCN) to

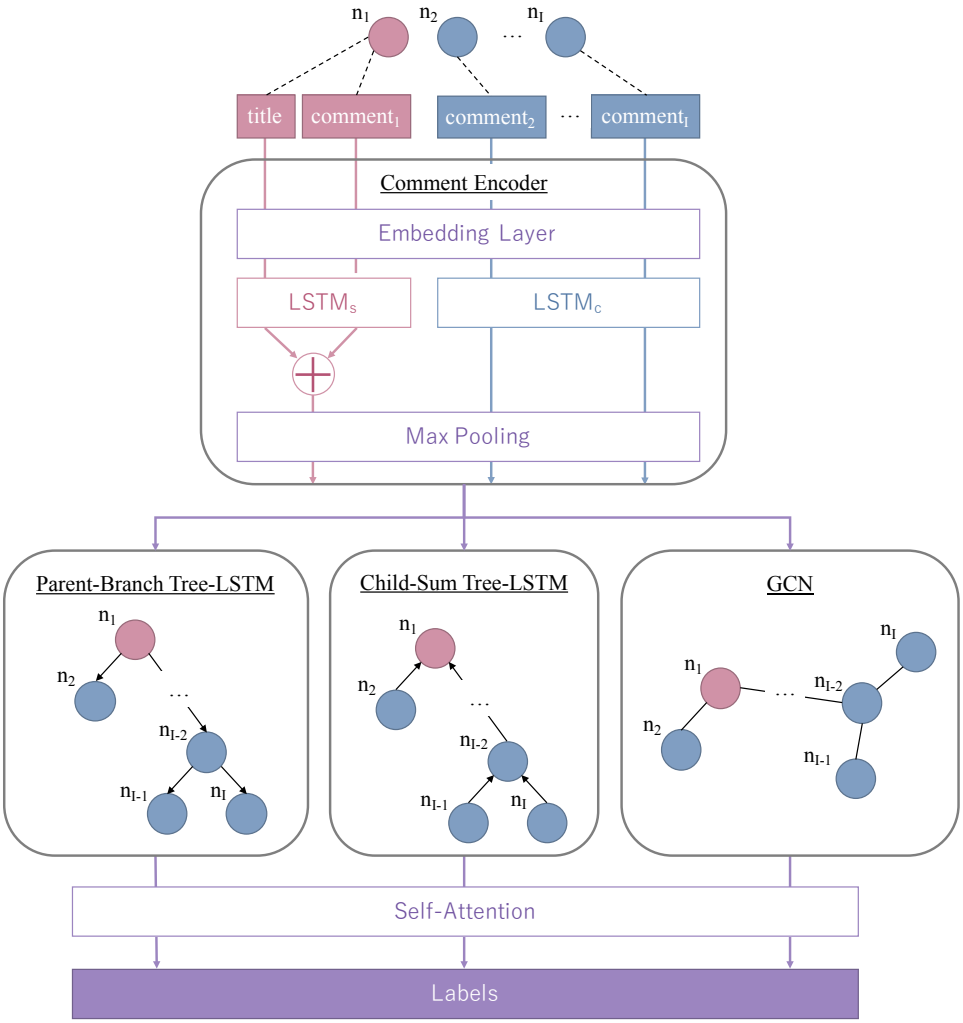

Figure 2: Overview of our proposed model: Tree-LSTM GCN Hybrid. In this model, comments in an input thread is encoded by Comment Encoder and then processed by tree processing components (Parent-Branch Tree-LSTM and Chid-Sum Tree-LSTM) and a graph processing component (GCN). Best viewed with colors.

introduce comment correlations. Finally, the updated comment representations are merged with an attention layer (Self-Attention) and are connected to labels with a fully-connected layer.

Comment Encoder

An input sequence consists of a title text $\boldsymbol{a}_{title}$ and comment texts $\boldsymbol{a}_{1...I}$. The words in these texts are embedded into $\boldsymbol{x}_{title}$ and $\boldsymbol{x}_{1...I}$ with an embedding matrix $\boldsymbol{E}$ in Embedding layer. Embedded inputs are then passed to bi-directional LSTM layers[2] to be processed using the following transition functions as

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_i\boldsymbol{x}_t + \boldsymbol{U}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \quad (1)$$
$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{U}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \quad (2)$$
$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{U}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \quad (3)$$
$$\tilde{\boldsymbol{c}}_t = \tanh\left(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}_c\boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right) \quad (4)$$
$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} \quad (5)$$
$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right) \quad (6)$$

where $\boldsymbol{i}_t$ is an input gate, $\boldsymbol{o}_t$ is an output gate, $\boldsymbol{f}_t$ is a forget gate, $\tilde{\boldsymbol{c}}_t$ is a cell gate, $\boldsymbol{c}_t$ is a cell state, $\boldsymbol{h}_t$ is a hidden state, $\boldsymbol{W}_*$ and $\boldsymbol{U}_*$ are weight matrices, $\boldsymbol{b}_*$ are bias vectors, σ is a logistic sigmoid function,

[2]We prepared two separate LSTMs for an initial comment ($\mathrm{LSTM_s}$) and for later comments ($\mathrm{LSTM_c}$).

and $\odot$ is an element-wise multiplication operator. Bi-directional LSTM outputs are concatenated and processed with a max-over time process to obtain a comment representation $\boldsymbol{m} = \max\left(\overrightarrow{\boldsymbol{h}} \| \overleftarrow{\boldsymbol{h}}\right)$. For an initial comment, the title output ($\boldsymbol{h}_{title}$) and the comment output ($\boldsymbol{h}_1$) are added before the max-over time process as $\boldsymbol{m} = \max\left(\overrightarrow{\boldsymbol{h}}_{title} \| \overleftarrow{\boldsymbol{h}}_{title} \oplus \overrightarrow{\boldsymbol{h}}_1 \| \overleftarrow{\boldsymbol{h}}_1\right)$, where $\oplus$ is an element-wise addition operator.

Parent-Branch Tree-LSTM

This component allows a discourse act of a comment to be predicted with information from its parent. The comments representations are processed from the root to the leaves with LSTM. We conducted this process with a simple extension to LSTM by replacing a previous time state $\boldsymbol{h}_{t-1}$ in Eq. 1–4 with a parent state $\boldsymbol{h}_{parent}$. A comment can have multiple children. Therefore, $\boldsymbol{h}_{parent}$ are likely to be shared by multiple comments. For latter processes, $\boldsymbol{h}$ is used as an updated comment representation $\boldsymbol{r}_P$.

Child-Sum Tree-LSTM

This component allows a discourse act of a comment to be predicted with information from its children. The comments representations are processed from the leaves to the root with LSTM. We used Child-Sum Tree-LSTM (Tai et al., 2015) for this process. Child-Sum Tree-LSTM extends LSTM by expressing a state transition of children to parent with a summation. More specifically, a previous time state $\boldsymbol{h}_{t-1}$ in Eq. 1, 2, and 4 are replaced by previous child states $\tilde{\boldsymbol{h}}_t$, with updates to forget gate $\boldsymbol{f}_t$ (Eq. 3) and cell gate $\boldsymbol{c}_t$ (Eq. 5) by

$$\tilde{\boldsymbol{h}}_t = \sum_k \boldsymbol{h}_k \tag{7}$$

$$\boldsymbol{f}_{tk} = \sigma\left(\boldsymbol{W}_f \boldsymbol{m}_t + \boldsymbol{U}_f \boldsymbol{h}_k + \boldsymbol{b}_f\right) \tag{8}$$

$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t + \sum_k \boldsymbol{f}_{tk} \odot \boldsymbol{c}_k \tag{9}$$

where $\boldsymbol{h}_k$ is a hidden state of a child and $\boldsymbol{c}_k$ is a cell state of a child. For latter processes, $\boldsymbol{h}$ is used as an updated comment representation $\boldsymbol{r}_C$.

GCN

This component allows a discourse act of a comment to be predicted with information from surrounding comments. The comments representations are processed with a convolution over a graph. For each node, the representations of the surrounding nodes are considered with a convolution filter. We used the graph convolution approach of Defferrard et al. (2016) which approximates a convolutional filter over a graph with Chebyshev expansion (Hammond et al., 2011). In this approach, input representations $\boldsymbol{h}_l$ are processed with the following transition functions as

$$\boldsymbol{h}_{l+1} = U g_\theta\left(\Lambda\right) U^T \boldsymbol{h}_l \tag{10}$$

$$g_{\boldsymbol{\theta}}\left(\Lambda\right) = \sum_k^{K-1} \boldsymbol{\theta}_k T_k(\tilde{\Lambda}) \tag{11}$$

where U is the Fourier basis of a normalized Laplacian in an input graph, $\boldsymbol{\theta}_k$ is a vector of Chebyshev coefficients, and $T_k(\tilde{\Lambda})$ is a Chebyshev polynomial of order k. This graph convolution process can be stacked easily to perform convolutions over a graph iteratively. We prepared two graph convolution processes with rectified linear unit (ReLU) as an intermediate activation as

$$\boldsymbol{r}_G = U g_\theta\left(\Lambda\right) U^T \mathrm{ReLU}\left(U g_\theta\left(\Lambda\right) U^T \boldsymbol{m}\right) \tag{12}$$

to obtain an updated comment representation $\boldsymbol{r}_G$.

Self-Attention

Outputs of three components are merged with an attention layer (Self-Attention) with a self-attentive style (Yang et al., 2016; Lin et al., 2017). A comment representation $\boldsymbol{s}_i$ is updated as a weighted sum of

$\boldsymbol{r}_{ji}$ with weight α_{ji} as

$$\boldsymbol{s}_i = \sum_{j \in \{P,C,G\}} \alpha_{ji} \boldsymbol{r}_{ji} \tag{13}$$

$$\alpha_{ji} = \frac{\exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_{ji}\right)}{\sum_{j' \in \{P,C,G\}} \exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_{j'i}\right)} \tag{14}$$

$$\boldsymbol{u}_{ji} = \tanh\left(\boldsymbol{W}_\alpha \boldsymbol{r}_{ji} + \boldsymbol{b}_\alpha\right) \tag{15}$$

where $\boldsymbol{v}_\alpha$ is a weight vector, $\boldsymbol{W}_\alpha$ is a weight matrix, and $\boldsymbol{b}_\alpha$ is a bias vector.

4 Experiment

4.1 Baselines

Rule 5-ACTS

We prepared a simple rule-based classifier as a non-machine learning baseline. Given a thread, an initial comment is classified as *Question* if it includes a question mark and as *Announcement* for another case. In later comments, a comment is classified as *Question* if it includes a question mark and as *Appreciation* if it includes "thank", "thanks", "thx", "thxs", or "tks", and as *Answer* or *Elaboration* for another case. In the *Answer* or *Elaboration* case, a comment is classified as *Answer* if it is preceded by a comment with *Question* and as *Elaboration* for another case. Note that comments are never classified with the remaining 4 acts (*Agreement*, *Disagreement*, *Humor*, and *Negative Reaction*) in this classifier.

CRF Vote

We implemented the best model in Zhang et al. (2017) as CRF Vote. This model decomposes a thread into root-to-leaf sequences. Features of content, punctuation, structure, author, thread, and community are extracted and used to train a Conditional Random Field (CRF) with Orthant-Wise Limited-memory Quasi-Newton training algorithm and L1 regularization. The strong effect of structural features were confirmed on an ablation test. In this model, a comment often receives multiple prediction labels since a thread is decomposed into sequences. In that case, the label of a comment is decided by a majority vote of the given labels.

LSTM-CRF Vote

We prepared a straightforward extension of CRF Vote combining Comment Encoder (Section 3), an LSTM layer, and a CRF layer. An architecture combining an LSTM and a CRF is known to be effective for sequential labeling tasks (Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016). As in CRF Vote, a thread is first decomposed into root-to-leaf sequences. The decomposed sequences are then processed with Comment Encoder to obtain comment representations. The comment representations are then updated with an LSTM layer to introduce comment correlations. Finally, the updated comment representations are passed to a CRF layer to include label correlations. Like in CRF Vote, the label of a comment with multiple prediction labels is decided by a majority vote of the given labels.

Graph-LSTM

We implemented the graph-structured LSTM in Zayats and Ostendorf (2018) as Graph-LSTM. In this model, LSTM is extended to process parent–child relationships and sibling relationships in a Reddit thread. Relationships of these two kinds are introduced to LSTM by preparing a forget gate for hierarchical (reply) connections and a forget gate for timing (sibling) connections. An input to the extended LSTM is comments represented with concatenated word representation averages and submission context features. Submission context features include structural features such as graph locations and graph responses. This model is not designed for discourse acts, but it has shown superior performances for predicting the popularity of Reddit comments.

Discourse Act	Count	Initial Comment Rate
Answer	40,723	0.00%
Elaboration	18,513	0.00%
Question	17,105	41.32%
Appreciation	8,380	0.00%
Agreement	4,815	0.00%
Disagreement	3,263	0.00%
Humor	2,291	2.71%
Announcement	2,002	100.00%
Negative Reaction	1,773	0.00%

Table 1: Numbers of discourse acts and their initial comment rates in the dataset that we used in the experiment.

4.2 Dataset and Evaluation

To evaluate the proposed model and the baseline models, we used the dataset of Zhang et al. (2017)[3], which consists of 9,483 threads with 115,827 comments from 2,837 communities (subreddits). These comments are annotated with the following 10 discourse acts: *Answer*, *Elaboration*, *Question*, *Appreciation*, *Agreement*, *Disagreement*, *Humor*, *Announcement*, *Negative Reaction*, and *Other*. Following the setting of Zhang et al. (2017), we discarded the comments that have no majority annotation or which have *Other* as a majority annotation. We used the resulting 9,131 threads with 98,865 comments to evaluate our models. The numbers of respective discourse acts and their initial comment rates in the resulting dataset are presented in Table 1.

All models are evaluated with ten-fold cross validation following the setting of Zhang et al. (2017). Accuracy, precision, recall, and F_1 score are used for evaluation metrics. For neural models, the dataset is split into train:validation:test in the ratio of 8:1:1 for each fold to provide validation data. A best performing model in a validation data in terms of F_1 score is used to evaluate the corresponding test data.

4.3 Model Configurations

Pre-training of Word Representations

We pre-trained the word-embeddings using randomly sampled comments from Reddit dumps during 2006–2016[4,5]. We restricted communities to those which appeared in the dataset of Section 4.2, and extracted approximately 230M comments. For pre-training, we used word2vec (Mikolov et al., 2013) with the skip-gram algorithm of parameters dimension=100, learning rate=0.025, window size=5, negative sample size=5, and epoch=5. The pre-trained word representations are used in LSTM-CRF Vote, Graph-LSTM, and Tree-LSTM GCN Hybrid.

Hidden Unit Sizes and Maximum Number of Words

Several layers in our models have hidden units as their parameters. For LSTM layers, we have set hidden unit sizes to $\mathrm{LSTM}_s = 300$, $\mathrm{LSTM}_c = 300$, and the LSTM layer in LSTM-CRF Vote to 600. Parent-Branch Tree-LSTM, Child-Sum Tree-LSTM, and GCN also have hidden units. Their size were set to 600. Reddit comments are sometimes long consisting of thousands of words. We restricted the maximum number of words for each comment to speedup training. The maximum numbers were set to 400 for initial comments and 100 for later comments.

Optimization

We trained LSTM-CRF Vote, Graph-LSTM, and Tree-LSTM GCN Hybrid by stochastic gradient descent. As an optimization objective, the score of CRF is used for LSTM-CRF Vote and cross-entropy loss is used for Graph-LSTM and Tree-LSTM GCN Hybrid. The learning rate is selected from $\{0.01, 0.1\}$ using a validation data along with momentum=0.9 and gradient clipping=3.0 for parameters of stochastic

[3]https://github.com/google-research-datasets/coarse-discourse

[4]https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts

[5]https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

Model	Accuracy	Precision	Recall	F_1 score
Rule 5-ACTS	0.590	0.513	0.590	0.536
CRF Vote (Zhang et al., 2017)	0.764	0.752	0.764	0.748
LSTM-CRF Vote	0.769	0.757	0.769	0.759
Graph-LSTM (Zayats and Ostendorf, 2018)	0.738	0.720	0.738	0.724
Proposed (Parent-Branch Tree-LSTM)	0.772	0.760	0.772	0.763
Proposed (Child-Sum Tree-LSTM)	0.631	0.609	0.631	0.611
Proposed (GCN)	0.767	0.755	0.767	0.756
Proposed (Tree-LSTM GCN Hybrid)	0.779	0.768	0.779	0.770

Table 2: Accuracy, precision, recall, and F_1 score of the baseline models and the proposed model. For precision, recall, and F_1 score, the weighted average with the number of positive instances over 9 discourse acts.

gradient descent. To avoid overfitting, we introduced dropout (Srivastava et al., 2014) with a rate of 0.5 to the LSTM layers of all models and the intermediate layer of GCN for regularization.

4.4 Result

Table 2 presents the evaluation results. For Tree-LSTM GCN Hybrid, we prepared models that used only a single component as Proposed (*component name*). Results show that our proposed model Tree-LSTM GCN Hybrid outperforms the previous best approach of Zhang et al. (2017) (CRF Vote) by 1.5% in accuracy and 2.2 in F_1 score. LSTM-CRF Vote also outperformed CRF Vote but in a smaller magnitude compared to Tree-LSTM GCN Hybrid. This result suggests the effectiveness of both the simple neural extension and the neural structural learning approaches[6]. Graph-LSTM performed moderately but lower than CRF Vote. This result implies that a popularity predictions and a discourse act classification differs even within Reddit. Rule 5-ACTS performed substantially lower than other machine learning models. This result indicates that simple rules are not enough to solve this task. For computational times, Tree-LSTM GCN Hybrid took approximately 21 hours to perform the ten-fold cross validation with an NVIDIA Titan X gpu. This is about five times slower to CRF Vote, which took approximately 4 hours on the evaluation with an Intel Core i7 cpu core.

Performances of the single component proposed models vary among active components. Results show that Parent-Branch Tree-LSTM performs best, with GCN slightly lower than Parent-Branch Tree-LSTM, and with Child-Sum Tree-LSTM scoring quite low. These results are intuitive because a discussion flow in a thread occurs from the root to the leaves with the chain of replies. GCN considers surrounding nodes in distance of two. Therefore, its performance suggests that the effect of distant comments is not strong to classify the discourse act of a comment.

5 Discussions

5.1 Strategy for Combining Three Components

Our proposed model combines three components to classify discourse acts. We analyzed attention probabilities in Self-Attention layer to see how Tree-LSTM GCN Hybrid merges the three components. Figure 3a shows the estimated probability density functions of all comments. As in the single component models in Section 4.4, Parent-Branch Tree-LSTM is most preferred in the model. However, the next preferred component is Child-Sum Tree-LSTM, which performed quite poorly as a single component model. This observation implies that the performance of an individual component might not relate directly to importance in a hybrid model. Figure 3b shows the estimated probability density functions of initial comments. Initial comments have no parents. Therefore, Child-Sum Tree-LSTM and GCN have stronger preference compared to the all comments case.

We further analyzed attention probabilities in terms of the number of replies. In the case of comments with none or a small number of replies, the probability density function shows a similar trend to the all

[6]We confirmed statistically significant improvements between Tree-LSTM GCN Hybrid and LSTM-CRF Vote for all four metrics with the confidence level of 0.05 by Fisher-Pitman permutation test. The improvements between LSTM-CRF Vote and CRF Vote were also statistically significant with the confidence level of 0.10 for accuracy and the confidence level of 0.05 for precision, recall, and F_1 score.

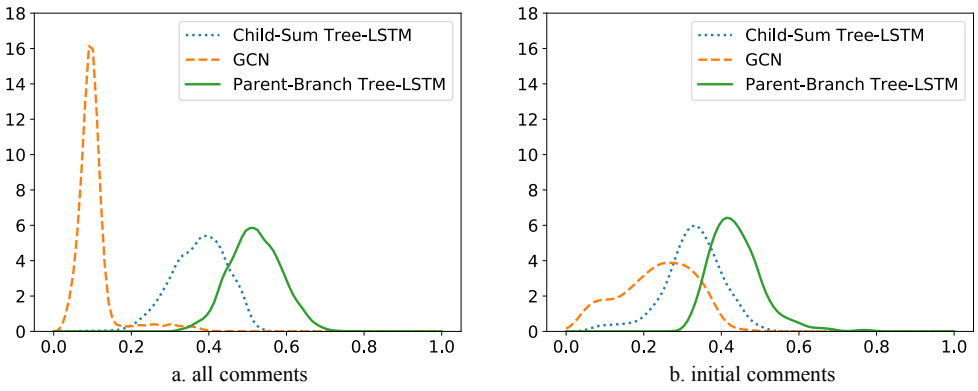

Figure 3: Estimated probability density functions of attention probabilities in the Self-Attention layer.

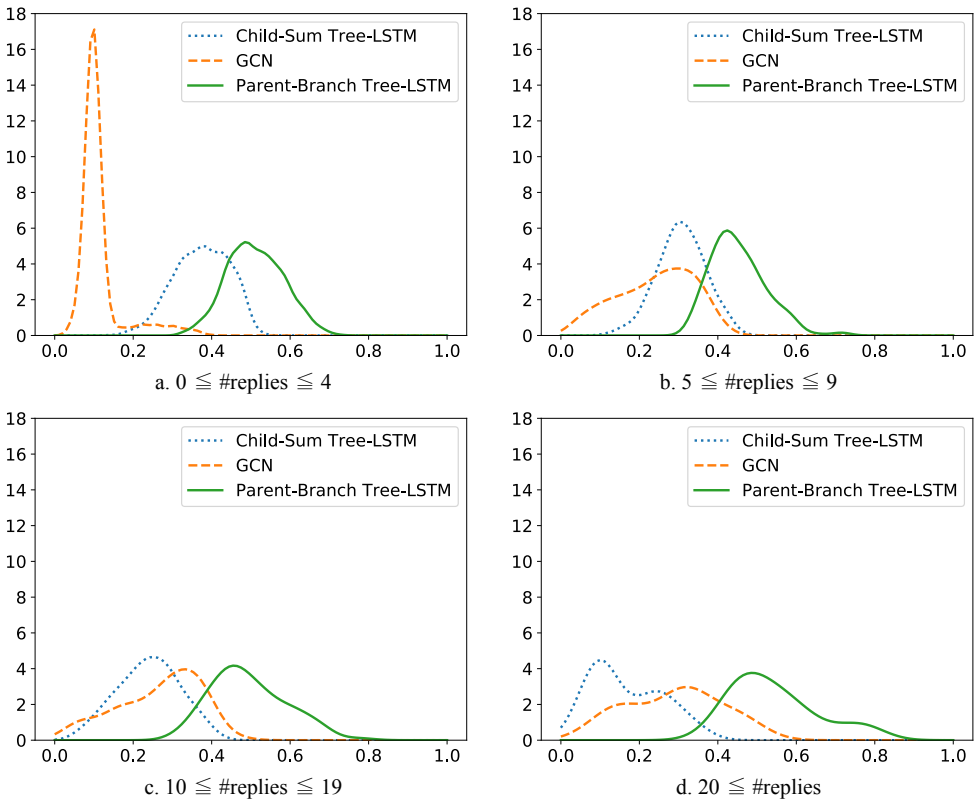

Figure 4: Estimated probability density functions in terms of the number of replies.

comments case (Figure 4a). The preference of Child-Sum Tree-LSTM and the preference of GCN become close in the case with a moderate number of replies (Figure 4b). The preference of GCN surpasses the preference of Child-Sum Tree-LSTM with larger numbers of replies (Figure 4c, 4d). These results imply that when there is a sufficient number of replies, considering nearby comments like in GCN is an effective approach for predicting discourse acts.

5.2 Strengths of Neural Models

In the experiment, we targeted 9 discourse acts for prediction. To clarify the effects of the proposed model, we analyzed the performances of each discourse act. Table 3 shows the detailed F_1 scores of Rule 5-ACTS, CRF Vote, LSTM-CRF Vote, Parent-Branch Tree-LSTM, and Tree-LSTM GCN Hybrid. The analysis shows that the neural models perform on par or better for most acts. Strong improvements in the neural models against CRF Vote are observed in *Disagreement*, *Humor*, and *Negative Reaction*.

Discourse Act	Rule 5-ACTS	CRF Vote	LSTM-CRF Vote	Parent-Branch Tree-LSTM	Tree-LSTM GCN Hybrid
Answer	0.685	0.884	0.892	0.897	0.900
Elaboration	0.325	0.647	0.657	0.660	0.667
Question	0.754	0.853	0.869	0.870	0.877
Appreciation	0.575	0.730	0.704	0.713	0.720
Agreement	0.000	0.460	0.458	0.462	0.475
Disagreement	0.000	0.219	0.280	0.259	0.299
Humor	0.000	0.180	0.285	0.269	0.290
Announcement	0.679	0.787	0.779	0.787	0.819
Negative Reaction	0.000	0.189	0.280	0.302	0.313

Table 3: F_1 scores for each discourse acts in selected baseline models and proposed models.

Model	Precision	Recall	F_1 score
Rule 5-ACTS	0.836	0.439	0.575
CRF Vote (Zhang et al., 2017)	0.794	0.676	0.730
LSTM-CRF Vote	0.721	0.689	0.704
Proposed (Parent-Branch Tree-LSTM)	0.722	0.703	0.713
Proposed (Tree-LSTM GCN Hybrid)	0.750	0.693	0.720

Table 4: Precisions, recalls, and F_1 scores of *Appreciation* in selected baseline models and proposed models.

These acts are less frequent than others (Table 1), and the improvements imply the strength of the neural models to capture infrequent characteristics.

The strength of the hybrid approach (Tree-LSTM GCN Hybrid) is observed in *Announcement* with an improvement of 3.2 in F_1 score. A unique characteristic of *Announcement* is that it only appears in initial comments (Table 1). Therefore, the strength of the hybrid approach supports the observation in Section 5.1 that Child-Sum Tree-LSTM and GCN are more preferred in initial comments. One exception where the proposed model did not work well is *Appreciation*. Table 4 shows the detailed evaluation values of *Appreciation*. In terms of precision, the non-neural models (Rule 5-ACTS and CRF Vote) performed better than the neural models. This analysis suggests that non-neural approaches have certain strength for classifying discourse acts that have strong linguistic clues (e.g. thank).

6 Conclusion

As described in this paper, we proposed a model that integrates discussion structures with neural networks to analyze discourse acts. Parent-Branch Tree-LSTM, Child-Sum Tree-LSTM, and GCN are used as components of the proposed model to capture discussion structures. Results show that, by combining the three components with a self-attentive process, improvements of 1.5% in accuracy and 2.2 in F_1 score are achieved compared to the previous best model. Interactions among the three components are also explored via analysis of attention probabilities in the self-attentive layer of the model. As future works of this study, we first plan to apply our model to discussion in different domains. Reddit comments have their strength in data size, but they are known to have strong association with non-textual attributes such as timing and authors. We are planning to expand our models to explore a more flexible architecture for analyzing discussions.

Acknowledgements

We would like to thank the members of Okumura–Takamura Group at Tokyo Institute of Technology for having fruitful discussions about tree learning approaches and graph learning approaches using neural networks. We would also like to thank the anonymous reviewer for their comments to improve this paper.

References

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2012. Classifying user messages for managing web forum data. In *Proceedings of the 15th International workshop on the Web and Databases*.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*.

Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in Reddit using network structure. In *Proceedings of the Workshop on Modeling Social Media: Mining Big Data in Social Media and the Web*, pages 615–620.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech act". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 345–352.

Vitor Carvalho and William Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 35–41.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2017. A factored neural network model for characterizing online discussions in vector space. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2296–2306.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29*, pages 3844–3852. Curran Associates, Inc.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 386–396.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150.

Ji He, Mari Ostendorf, Xiaodong He, Jianshu Chen, Jianfeng Gao, Lihong Li, and Li Deng. 2016. Deep reinforcement learning with a combinatorial action space for predicting popular Reddit threads. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1838–1848.

Ji He, Mari Ostendorf, and Xiaodong He. 2017. Reinforcement learning with external knowledge and two-stage Q-functions for predicting popular Reddit threads. *arXiv preprint arXiv:1704.06217*.

Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 927–936.

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

Jun Hu, Rebecca Passonneau, and Owen Rambow. 2009. Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. In *Proceedings of the SIGDIAL 2009 Conference*, pages 357–366.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010a. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010b. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Wern Han Lim, Mark James Carman, and Sze-Meng Jojo Wong. 2017. Estimating relative user expertise for content quality prediction on Reddit. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 55–64.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*.

Fei Liu, Timothy Baldwin, and Trevor Cohn. 2017. Capturing long-range contextual dependencies with memory-enhanced conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 555–565.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Adinoyi Omuya, Vinodkumar Prabhakaran, and Owen Rambow. 2013. Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 802–807.

Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 129–136.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624.

Yi-Chia Wang, Mahesh Joshi, and Carolyn Rose. 2007. A feature based approach to leveraging context for classifying newsgroup style discussion segments. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 73–76.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 13–25.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 195–200.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Victoria Zayats and Mari Ostendorf. 2018. Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics*, 6:121–132.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in Twitter. In *Proceedings of the AAAI-11 Workshop on Analyzing Microtext*, pages 86–91.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2012. Towards scalable speech act recognition in Twitter: Tackling insufficient training data. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 18–27.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th AAAI International Conference on Web and Social Media*, pages 357–366.