A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification

Ruizhe Li[♠], Chenghua Lin[♡], Matthew Collinson[♠], Xiao Li[♠] and Guanyi Chen[♣]
[♠]Department of Computing Science, University of Aberdeen, UK
{r02rl17, matthew.collinson, x.li}@abdn.ac.uk
[♡]Department of Computer Science, University of Sheffield, UK
c.lin@sheffield.ac.uk
[♣]Department of Information and Computing Sciences, Utrecht University, The Netherlands
g.chen@uu.nl

Abstract

Recognising dialogue acts (DA) is important for many natural language processing tasks such as dialogue generation and intention recognition. In this paper, we propose a dual-attention hierarchical recurrent neural network for DA classification. Our model is partially inspired by the observation that conversational utterances are normally associated with both a DA and a topic, where the former captures the social act and the latter describes the subject matter. However, such a dependency between DAs and topics has not been utilised by most existing systems for DA classification. With a novel dual task-specific attention mechanism, our model is able, for utterances, to capture information about both DAs and topics, as well as information about the interactions between them. Experimental results show that by modelling topic as an auxiliary task, our model can significantly improve DA classification, yielding better or comparable performance to the state-of-the-art method on three public datasets.

1 Introduction

Dialogue Acts (DA) are semantic labels of utterances, which are crucial to understanding communication: much of a speaker's intent is expressed, explicitly or implicitly, via social actions (e.g., questions or requests) associated with utterances (Searle, 1969). Recognising DA labels is important for many natural language processing tasks. For instance, in dialogue systems, knowing the DA label of an utterance supports its interpretation as well as the generation of an appropriate response (Searle, 1969; Chen et al., 2018). In the security domain, being able to detect intention in conversational texts can effectively support the recognition of sensitive information exchanged in emails or other communication channels, which is critical to timely security intervention (Verma et al., 2012).

A wide range of techniques have been investigated for DA classification. Early works on DA classification are mostly based on general machine learning techniques, framing the problem either as multi-class classification (e.g., using SVMs (Liu, 2006) and dynamic Bayesian networks (Dielmann and Renals, 2008)) or a structured prediction task (e.g., using Conditional Random Fields (Kim et al., 2010; Chen et al., 2018; Raheja and Tetreault, 2019, CRF)). Recent studies to the problem of DA classification have seen an increasing uptake of deep learning techniques, where promising results have been obtained. Deep learning approaches typically model the dependency between adjacent utterances (Ji et al., 2016; Lee and Dernoncourt, 2016). Some researchers further account for dependencies among both consecutive utterances and consecutive DAs, i.e., both are considered factors that influence natural dialogue (Kumar et al., 2018; Chen et al., 2018). There is also work exploring different deep learning architectures (e.g., hierarchical CNN or RNN/LSTM) for incorporating context information for DA classification (Liu et al., 2017).

It has been observed that conversational utterances are normally associated with both a DA and a topic, where the former captures the social act (e.g., promising) and the latter describes the subject matter (Wallace et al., 2013). It is also recognised that the types of DA associated with a conversation are likely to be influenced by the topic of the conversation (Searle, 1969; Wallace et al., 2013). For instance, conversations relating to topics about *customer service* might be more frequently associated with DAs of type Wh-question (e.g., *Why my mobile is not working?*) and a complaining statement (Bhuiyan et al., 2018); whereas meetings covering administrative

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 383–392
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

topics about resource allocation are likely to exhibit significantly more defending statements and floor grabbers (e.g., *Well I mean - is the handheld really any better?*) (Wrede and Shriberg, 2003). However, such a reasonable source of information, surprisingly, has not been explored in the deep learning literature for DA classification. We assume that modelling the topics of utterances as additional contextual information may effectively support DA classification.

In this paper, we propose a dual-attention hierarchical recurrent neural network with a CRF (DAH-CRF) for DA classification. Our model is able to account for rich context information with the developed dual-attention mechanism, which, in addition to accounting for the dependencies between utterances, can further capture, for utterances, information about both topics and DAs. Topic is a useful source of context information which has not previously been explored in existing deep learning models for DA classification. Second, compared to the flat structure employed by existing models (Khanpour et al., 2016; Ji et al., 2016), our hierarchical recurrent neural network can represent the input at the character, word, utterance, and conversation levels, preserving the natural hierarchical structure of a conversation. To capture the topic information of conversations, we propose a simple automatic utterance-level topic labelling mechanism based on LDA (Blei et al., 2003), which avoids expensive human annotation and improves the generalisability of our model.

We evaluate our model against several strong baselines (Wallace et al., 2013; Ji et al., 2016; Kumar et al., 2018; Chen et al., 2018; Raheja and Tetreault, 2019) on the task of DA classification. Extensive experiments conducted on three public datasets (i.e., Switchboard Dialog Act Corpus (SWDA), DailyDialog (DyDA), and the Meeting Recorder Dialogue Act corpus (MRDA)) show that by modelling the topic information of utterances as an auxiliary task, our model can significantly improve DA classification for all datasets compared to a base model without modelling topic information. Our model also yields better or comparable performance to state-of-the-art deep learning method (Raheja and Tetreault, 2019) in classification accuracy.

To summarise, the contributions of our paper are three-fold: (1) we propose to leverage topic information of utterances, a useful source of contextual information which has not previously been explored in existing deep learning models for DA classification; (2) we propose a dual-attention hierarchical recurrent neural network with a CRF which respects the natural hierarchical structure of a conversation, and is able to incorporate rich context information for DA classification, achieving better or comparable performance to the state-of-the-art; (3) we develop a simple topic labelling mechanism, showing that using the automatically acquired topic information for utterances can effectively improve DA classification.

2 Related Work

Broadly speaking, methods for DA classification can be divided into two categories: multi-class classification (e.g., SVMs (Liu, 2006) and dynamic Bayesian networks (Dielmann and Renals, 2008)) and structured prediction tasks including HMM (Stolcke et al., 2000) and CRF (Kim et al., 2010). Recently, deep learning has been widely applied in many NLP tasks, including DA classification. Kalchbrenner and Blunsom (2013) proposed to model a DA sequence with a RNN where sentence representations were constructed by means of a convolutional neural network (CNN). Lee and Dernoncourt (2016) tackled DA classification with a model built upon RNNs and CNNs. Specifically, their model can leverage the information of preceding texts, which can effectively help improve the DA classification accuracy. A latent variable recurrent neural network was developed for jointly modelling sequences of words and discourse relations between adjacent sentences (Ji et al., 2016). In their work, the shallow discourse structure is represented as a latent variable and the contextual information from preceding utterances are modelled with a RNN.

Kumar et al. (2018) proposed a hierarchical Bi-LSTM model with a CRF for DA classification, where the inter-utterance and intra-utterance information are encoded by a hierarchical Bi-LSTM and the dependency between DA labels is captured by a CRF. Chen et al. (2018) developed a CRF-Attentive Structured Network (CRF-ASN) for DA classification. They applied structured attention network to the CRF layer in order to model contextual utterances and corresponding DAs together. Raheja and Tetreault (2019) achieved the state-of-the-art performance on the SWDA dataset by employing a self-attention mechanism, a CRF layer and character-level embeddings.

In addition to modelling dependency between utterances, various contexts have also been explored for improving DA classification or joint modelling DA under multi-task learning. For instance, Wallace et al. (2013) proposed a generative joint sequential model to classify both DA and topics of patient-doctor conversations. Their model is similar to the factorial LDA model (Paul and Dredze, 2012), which generalises LDA to assign each token a K-dimensional vector of latent variables. We would like to emphasise that the model of Wallace et al. (2013), only assumed that each utterance is generated conditioned on the previous and current topic/DA pairs. In contrast, our model is able to model the dependencies of all preceding utterances of a conversation, and hence can better capture the effect between DAs and topics.

3 Methodology

Given a training corpus $\mathcal{D} = \langle (C_n, Y_n, Z_n) \rangle_{n=1}^{N}$, where $C_n = \langle u_t^n \rangle_{t=1}^{T}$ is a conversation containing a sequence of T utterances, $Y_n = \langle y_t^n \rangle_{t=1}^{T}$ and $Z_n = \langle z_t^n \rangle_{t=1}^{T}$ are the corresponding labels of DA and topics for C_n, respectively. Each utterance $u_t = \langle w_t^i \rangle_{i=1}^{K}$ of C_n is a sequence of K words. Our goal is to learn a model from $\mathcal{D}$, such that, given an unseen conversation C_u, the model can predict the DA labels of the utterances of C_u.

Figure 1 gives an overview of the proposed Dual-Attention Hierarchical recurrent neural network with a CRF (DAH-CRF). A shared utterance encoder encodes each word w_t^i of an utterance u_t into a vector $\mathbf{h}_t^i$. The DA attention and topic attention mechanisms capture DA and topic information as well as the interactions between them. The outputs of the dual-attention are then encoded in the conversation-level sequence taggers (i.e., $\mathbf{g}_t$ and $\mathbf{s}_t$), based on the corresponding utterance representations (i.e., $\mathbf{l}_t$ and $\mathbf{v}_t$). Finally, the target labels (i.e., y_t and z_t) are predicted in the CRF layer.

3.1 Shared Utterance Encoder

In our model, we adopt a shared utterance encoder to encode the input utterances. Such a design is based on the rationale that the shared encoder can transfer parameters between two tasks and reduce the risk of overfitting (Ruder, 2017). Specifically, the shared utterance encoder is implemented using the bidirectional gated recurrent unit (Cho et al., 2014, BiGRU), which encodes each utterance $u_t = \langle w_t^i \rangle_{i=1}^{K}$ of a conversation C_n as a series of hidden states $\langle \mathbf{h}_t^i \rangle_{i=1}^{K}$. Here, i indicates the timestamp of a sequence, and we define $\mathbf{h}_t^i$ as follows

$$\mathbf{h}_t^i = \overrightarrow{\mathbf{h}}_t^i \oplus \overleftarrow{\mathbf{h}}_t^i \tag{1}$$

where $\oplus$ is an operation for concatenating two vectors, and $\overrightarrow{\mathbf{h}}_t^i$ and $\overleftarrow{\mathbf{h}}_t^i$ are the i-th hidden state of the forward gated recurrent unit (Cho et al., 2014, GRU) and backward GRU for w_t^i, respectively. Formally, the forward GRU $\overrightarrow{\mathbf{h}}_t^i$ is computed as follows

$$\overrightarrow{\mathbf{h}}_t^i = \mathrm{GRU}(\overrightarrow{\mathbf{h}}_t^{i-1}, \mathbf{e}_t^i) \tag{2}$$

where $\mathbf{e}_t^i$ is the concatenation of the word embedding and the character embedding of word w_t^i. Finally, the backward GRU encodes u_t from the reverse direction (i.e. $w_t^K \rightarrow w_t^1$) and generates $\langle \overleftarrow{\mathbf{h}}_t^i \rangle_{i=1}^{K}$ following the same formulation as the forward GRU.

3.2 Task-specific Attention

Recall that one of the key challenges of our model is to capture for each utterance, information about both DAs and topics, as well as information about the interactions between them. We address this challenge by incorporating into our model a novel task-specific dual-attention mechanism, which accounts for both DA and topic information extracted from utterances. In addition, DAs and topics are semantically relevant to different words in an utterance. With the proposed attention mechanism, our model can also assign different weights to the words of an utterance by learning the degree of importance of the words to the DA or topic labelling task, i.e., promoting the words which are important to the task and reducing the noise introduced by less important words.

For each utterance u_t, the DA attention calculates a weight vector $\langle \alpha_t^i \rangle_{i=1}^{K}$ for $\langle \mathbf{h}_t^i \rangle_{i=1}^{K}$, the hidden states of u_t. u_t can then be represented as an attention vector $\mathbf{l}_t$ computed as follows

$$\mathbf{l}_t = \sum_{i=1}^{K} \alpha_t^i \mathbf{h}_t^i \tag{3}$$

In contrast to the traditional attention mechanism (Bahdanau et al., 2015), which only depends on one set of hidden vectors from the Seq2Seq decoder, the DA attention of our model relies on two sets of hidden vectors, i.e., $\mathbf{g}_{t-1}$ of

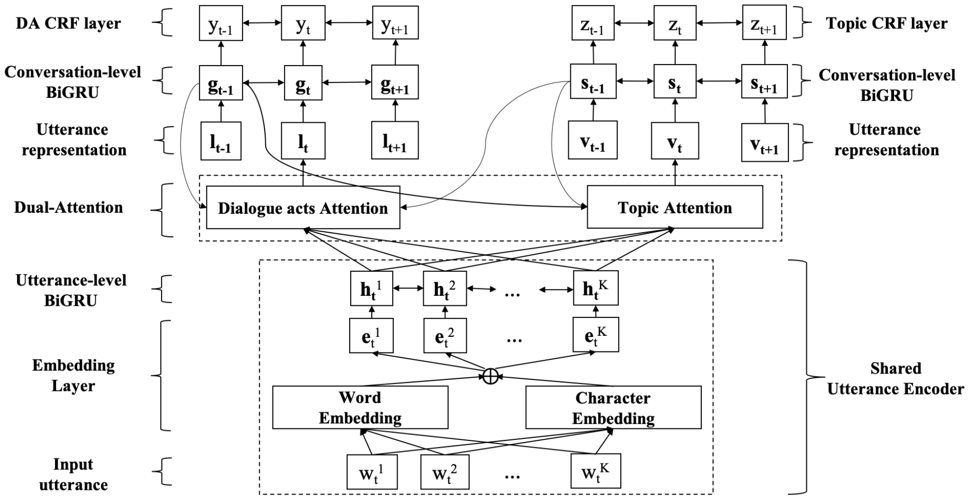
DA CRF layer
Conversation-level BiGRU
Utterance representation
Dual-Attention
Utterance-level BiGRU
Embedding Layer
Input utterance
Topic CRF layer
Conversation-level BiGRU
Utterance representation
Shared Utterance Encoder
y_{t-1} y_t y_{t+1}
z_{t-1} z_t z_{t+1}
$\mathbf{g}_{t-1}$ $\mathbf{g}_t$ $\mathbf{g}_{t+1}$
$\mathbf{s}_{t-1}$ $\mathbf{s}_t$ $\mathbf{s}_{t+1}$
$\mathbf{l}_{t-1}$ $\mathbf{l}_t$ $\mathbf{l}_{t+1}$
$\mathbf{v}_{t-1}$ $\mathbf{v}_t$ $\mathbf{v}_{t+1}$
Dialogue acts Attention
Topic Attention
$\mathbf{h}_t^1$ $\mathbf{h}_t^2$ ... $\mathbf{h}_t^K$
$\mathbf{e}_t^1$ $\mathbf{e}_t^2$ ... $\mathbf{e}_t^K$
$\oplus$
Word Embedding
Character Embedding
w_t^1 w_t^2 ... w_t^K

Figure 1: Overview of the dual-attention hierarchical recurrent neural network with a CRF.

the conversation-level DA tagger and $\mathbf{s}_{t-1}$ of the conversation-level topic tagger, where dual attention mechanism can capture, for utterances, information about both DAs and topics as well as the interaction between them. Specifically, the weights $\langle\alpha_t^i\rangle_{i=1}^K$ for the DA attention are calculated as follows:

$$\alpha_t^i = \text{softmax}(o_t^i) \tag{4}$$

$$o_t^i = \mathbf{w}_a^\top \tanh\left(\mathbf{W}^{(\text{act})}(\mathbf{s}_{t-1} \oplus \mathbf{g}_{t-1} \oplus \mathbf{h}_t^i) + \mathbf{b}^{(\text{act})}\right) \tag{5}$$

The topic attention layer has a similar architecture to the DA attention layer, which takes as input both $\mathbf{s}_{t-1}$ and $\mathbf{g}_{t-1}$. The weight vector $\langle\beta_t^i\rangle_{i=1}^K$ for the topic attention output $\mathbf{v}_t$ can be calculated similar to Eq. 3 and Eq. 4. Note that $\mathbf{w}_a$, $\mathbf{W}^{(\text{act})}$, and $\mathbf{b}^{(\text{act})}$ are vectors of parameters that need to be learned during training.

3.3 Conversational Sequence Tagger

CRF sequence tagger for DA. The conversational CRF sequence tagger for DA predicts the next DA y_t conditioned on the conversational hidden state $\mathbf{g}_t$ and adjacent DAs (c.f. Figure 1). Formally, this conditional probability of the whole conversation can be formulated as

$$p\left(y_{1:T}|C;\theta\right) = \frac{\prod_{t=1}^T \Psi\left(y_{t-1}, y_t, \mathbf{g}_t; \theta\right)}{\sum_Y \prod_{t=1}^T \Psi\left(y_{t-1}, y_t, \mathbf{g}_t; \theta\right)} \tag{6}$$

$$\begin{aligned}\Psi\left(y_{t-1}, y_t, \mathbf{g}_t; \theta\right) &= \Psi_{emi}\left(y_t, \mathbf{g}_t\right)\Psi_{tran}\left(y_{t-1}, y_t\right) \\ &= \mathbf{g}_t\left[y_t\right]\mathbf{P}_{y_t, y_{t-1}}\end{aligned} \tag{7}$$

Here the feature function $\Psi(\cdot)$ includes two score potentials: emission and transition. The emission potential Ψ_{emi} regards utterance representation $\mathbf{g}_t$ as the unary feature. The transition potential Ψ_{tran} is a pairwise feature constructed from a $T \times T$ state transition matrix $\mathbf{P}$, where T is the number of DA classes, and $\mathbf{P}_{y_t,y_{t-1}}$ is the probability of transiting from state y_{t-1} to y_t. $C = \langle u_t\rangle_{t=1}^T$ is the sequence of all utterances seen so far, θ is the parameters of the CRF layer. $\mathbf{g}_t$ is calculated in a BiGRU similar to Eq. 1 and Eq. 2:

$$\mathbf{g}_t = \overrightarrow{\mathbf{g}}_t \oplus \overleftarrow{\mathbf{g}}_t \tag{8}$$

$$\overrightarrow{\mathbf{g}}_t = \text{GRU}(\overrightarrow{\mathbf{g}}_{t-1}, \mathbf{l}_t) \tag{9}$$

CRF sequence tagger for topic. The conversational CRF sequence tagger for topic is designed to predict topic z_t conditioned on $\mathbf{v}_t$ and adjacent topics, which can be calculated similar to the formulation of the CRF tagger for DA.

Training the model. Let Θ be all the model parameters that need to be estimated for DAH-CRF. Θ then is estimated based on $\mathcal{D} = \langle(C_n, Y_n, Z_n)\rangle_{n=1}^N$ (i.e., a corpus with N conversations) by maximising the following objective function

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^N &\left[\log\left(p\left(y_{1:T}^n|C_n;\Theta\right)\right)\right. \\ &\left.+\alpha\log\left(p\left(z_{1:T}^n|C_n;\Theta\right)\right)\right]\end{aligned} \tag{10}$$

The hyper-parameter α controls the contribution of the conversational topic tagger towards the objective function. In our experiments, $\alpha = 0.5$ is determined using the validation datasets. During

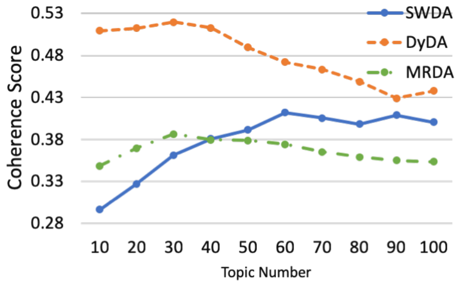

Figure 2: Coherence score of LDA on three datasets.

the test, the optimal DA or topic sequence is calculated using the Viterbi algorithm (Viterbi, 1967).

$$Y' = \arg \max_{y_{1:T} \in Y} p(y_{1:T}|C, \Theta) \tag{11}$$

3.4 Automatically Acquiring Topic Labels

To avoid expensive human annotation and to improve the generalisability of our model, we propose to label the topic of each utterance of the datasets using LDA (Blei et al., 2003). While perplexity has been widely used for model selection for LDA (Lin, 2011; He et al., 2012), we employ a topic coherence measure proposed by (Röder et al., 2015) to determine the optimal topic number for each dataset, which combines the indirect cosine measure with the normalised pointwise mutual information (Bouma, 2009, NPMI) and the Boolean sliding window. Empirically, we found the latter yields much better topic clusters than perplexity for supporting DA classification.

We treat each conversation as a document and train topic models using Gensim with topic number settings ranging from 10 to 100 (using an increment step of 10). Gibbs sampling is used to estimate the model posterior and for each model we run 1,000 iterations. For each trained model, we calculate the averaged coherence score of the extracted topics using Gensim[1], an implementation following (Röder et al., 2015). Figure 2 shows the topic coherence score for each topic number setting for all datasets, from which we determine that the optimal topic number setting for SWDA, DyDA, and MRDA are 60, 30, and 30, respectively.

Based on the optimal models (i.e., a trained LDA model using the optimal topic number setting), we assign topic labels to the datasets with two different strategies, i.e., conversation-level labelling (*conv*) and utterance-level labelling (*utt*).

Dataset	$\|C\|$	$\|T\|$	$\|V\|$	Training	Validation	Testing
SWDA	42	66	20K	1003/193K	112/23K	19/5K
DyDA	4	10	22K	11K/92.7K	1K/8.5K	1K/8.2K
MRDA	5	-	15K	51/77.9K	11/15.8K	11/15.5K

Table 1: $|C|$ is the number of DA classes, $|T|$ is the number of manually labelled conversation-level topic classes, $|V|$ is the vocabulary size. Training, Validation and Testing indicate the number of conversations/utterances in the respective splits.

For conversation-level labelling, we assign the topic label with the highest marginal probability to the conversation based on the corresponding per-document topic proportion estimated by LDA. Every utterance of the conversation then shares the same topic label of the conversation. For utterance-level labelling, there is an additional step to perform inference on every utterance based on corresponding optimal model (e.g., for every utterance of SWDA, we do inference using the LDA trained on SWDA with 60 topics), and assign the topic label with the highest marginal probability to the utterance. Therefore, the topic labels of the utterances of the same conversation could be different for utterance-level labelling.

4 Experimental Settings

4.1 Datasets

We evaluate the performance of our model on three public DA datasets with different characteristics, namely, Switchboard Dialog Act Corpus (Jurafsky, 1997, SWDA), Dailydialog (Li et al., 2017, DyDA), and the Meeting Recorder Dialogue Act corpus (Shriberg et al., 2004, MRDA).

SWDA[2] consists of 1,155 two-sided telephone conversations manually labelled with 66 conversation-level topics (e.g., *taxes*, *music*, etc.) and 42 utterance-level DAs (e.g., *statement-opinion*, *statement-non-opinion*, *wh-question*).

DyDA[3] contains 13,118 human-written daily conversations, manually labelled with 10 conversation-level topics (e.g., *tourism*, *politics*, *finance*) as well as four utterance-level DA classes, i.e., *inform*, *question*, *directive* and *commissive*. The former two classes are information transfer acts, while the latter two are action discussion acts.

MRDA[4] contains 75 meeting conversations anno-

[1] https://radimrehurek.com/gensim/models/coherencemodel.html

[2] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

[3] http://yanran.li/dailydialog

[4] http://www1.icsi.berkeley.edu/~ees/dadb/

tated with 5 DAs, i.e., Statement (S), Question (Q), Floorgrabber (F), Backchannel (B), and Disruption (D). The average number of utterances per conversation is 1,496. There are no manually annotated topic labels available for this dataset.

4.2 Implementation Details

For all experimental datasets, the top 85% highest frequency words were indexed. For SWDA and MRDA, we split training/validation/testing datasets following (Stolcke et al., 2000; Lee and Dernoncourt, 2016). For DyDA, we used the standard split from the original dataset (Li et al., 2017). The statistics of the experimental datasets are summarised in Table 1. We represented input data with 300-dimensional Glove word embeddings (Pennington et al., 2014) and 50-dimensional character embeddings (Ma and Hovy, 2016). We set the dimension of the hidden layers (i.e., $\mathbf{h}_t^i$, $\mathbf{g}_t$ and $\mathbf{s}_t$) to 256 and applied a dropout layer to both the shared encoder and the sequence tagger at a rate of 0.2. The Adam optimiser (Kingma and Ba, 2015) was used for training with an initial learning rate of 0.001 and a weight decay of 0.0001. Each utterance in a mini-batch was padded to the maximum length for that batch, and the maximum batch-size allowed was 50.

4.3 Baselines

We compare the proposed DAH-CRF model incorporating utterance-level topic labels extracted by LDA (denoted as DAH-CRF+LDA$_{utt}$) against five strong baselines and two variants of our own models:

JAS[5]: A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication (Wallace et al., 2013);

DRLM-Cond[6]: A latent variable recurrent neural network for DA classification (Ji et al., 2016);

Bi-LSTM-CRF[7]: A hierarchical Bi-LSTM with a CRF to classify DAs (Kumar et al., 2018);

CRF-ASN: An attentive structured network with a CRF for DA classification (Chen et al., 2018);

SelfAtt-CRF: A hierarchical Bi-GRU with self-attention and CRF (Raheja and Tetreault, 2019);

DAH-CRF+MANUAL$_{conv}$: Use the manually annotated conversation-level topic labels (i.e., each utterance of the conversation shares the same topic) for DAH-CRF model training rather than the topic labels automatically acquired from LDA;

DAH-CRF+LDA$_{conv}$: Use conversation-level topic labels automatically acquired from LDA for DAH-CRF model training.

	Model	SWDA	MRDA	DyDA
Baselines	JAS	71.2	81.3	75.9
	DRLM-Cond	77.0†	88.4	81.1
	Bi-LSTM-CRF	79.2†	90.9†	83.6
	CRF-ASN	80.8†	91.4†	-
	SelfAtt-CRF	**82.9**†	91.1†	-
Ours	DAH-CRF + MANUAL$_{conv}$	80.9	-	86.5
	DAH-CRF + LDA$_{conv}$	80.7	91.2	86.4
	DAH-CRF + LDA$_{utt}$	82.3	**92.2**	**88.1**
	Human Agreement	84.0	-	-

Table 2: DA classification accuracy. † indicates the results which are reported from the prior publications.

Note that only JAS (a non-deep-learning model) has attempted to model both DAs and topics, whereas all the deep learning baselines do not model topic information as a source of context for DA classification. All the baselines mentioned above use the same test dataset as our models for all experimental datasets.

5 Experimental Results

5.1 Dialogue Acts Classification

Table 2 shows the DA classification accuracy of our models and the baselines on three experimental datasets. We fine-tuned the model parameters for JAS, DRLM-Cond and Bi-LSTM-CRF in order to make the comparison as fair as possible. The implementation of CRF-ASN and SelfAtt-CRF are not available so we can only report their results for SWDA and MRDA based on the original papers (Chen et al., 2018; Raheja and Tetreault, 2019).

It can be observed that by jointly modelling DA and topics, DAH-CRF+LDA$_{utt}$ outperforms the two best baseline models SelfAtt-CRF and CRF-ASN around 1% on the MRDA dataset. Our model also gives similar performance to SelfAtt-CRF, the baseline which achieved the state-of-the-art performance on the SWDA dataset (i.e., 82.3% vs. 82.9%). While both manually annotated and automatically acquired topic labels are effective, we see that DAH-CRF+LDA$_{utt}$ outperforms both DAH-CRF+MANUAL$_{conv}$ and DAH-CRF+LDA$_{conv}$, i.e., with over 1.6% gain on DyDA and over 1.4% on SWDA (significant; paired t-test $p < .01$). It is also ob-

[5] https://github.com/bwallace/JAS

[6] https://github.com/jiyfeng/drlm

[7] https://github.com/YanWenqiang/HBLSTM-CRF

Model	SWDA	MRDA	DyDA
SAH	76.2	88.5	82.5
SAH-CRF	78.4	89.6	84.1
DAH + LDA$_{utt}$	79.5	91.1	86.0
DAH-CRF + LDA$_{utt}$ (without Dual-Att)	81.0	91.3	86.3
DAH-CRF + LDA$_{utt}$	82.3	92.2	88.1

Table 3: Ablation studies of DA classification.

served that DAH-CRF+MANUAL$_{conv}$ and DAH-CRF+LDA$_{conv}$ perform very similar to each other.

5.2 Ablation Study Results

We conducted ablation studies (see Table 3) in order to evaluate the contribution of the components of our DAH-CRF+LDA$_{utt}$ model, and more importantly, the effectiveness of leveraging topic information for supporting DA classification.

DAH-CRF+LDA$_{utt}$ (without Dual-Att) removes the dual-attention component from DAH-CRF+LDA$_{utt}$, and DAH+LDA$_{utt}$ removes the CRF from DAH-CRF+LDA$_{utt}$ but retaining the dual-attention component. SAH is a Single-Attention Hierarchical RNN model without a CRF, i.e., a simplified version of DAH+LDA$_{utt}$ that only models DAs with topical information omitted. As can be seen in Table 3, DAH+LDA$_{utt}$ achieves over 3% averaged gain on all datasets when compared to SAH, which clearly shows that leveraging topic information can effectively support DA classification. It is also observed that both the dual-attention mechanism and the CRF component are beneficial, but are more effective on the SWDA and DyDA datasets than MRDA.

In summary, while all the analysed model components are beneficial, the biggest gain is obtained by jointly modelling DAs and topics.

5.3 Analysing the Effectiveness of Joint Modelling Dialogue Act and Topic

In this section, we provide detailed analysis on why DAH-CRF+LDA$_{utt}$ can yield better performance than SAH-CRF by jointly modelling DAs and topics. Due to the page limit, our discussion focuses on SWDA and DyDA datasets.

Figure 4 shows the normalized confusion matrix derived from 10 DA classes of SWDA for both SAH-CRF and DAH-CRF+LDA$_{utt}$ models. It can be observed that DAH-CRF+LDA$_{utt}$ yields improvement on recall for many DA classes compared to SAH-CRF, e.g., 23.8% improvement on *bk* and 11.7% on *sv*. For *bk* (`Response Acknowledge`) which has the highest improvement level, we see that the improvement largely comes from the reduction of misclassifing *bk* to *b* (`Acknowledge Backchannel`). The key difference between *bk* and *b* is that an utterance labelled with *bk* has to be produced within a question-answer context, whereas *b* is a "continuer" simply representing a response to the speaker (Jurafsky, 1997). It is not surprising that SAH-CRF makes poor prediction on the utterances of these two DAs: they share many syntactic cues, e.g., indicator words such 'okay', 'oh', and 'uh-huh', which can easily confuse the model. When comparing the topic distribution of the utterances under the *bk* and *b* categories (cf. Figure 3), we found topics relating to personal leisure (e.g., buying cars, music, and exercise) are much more prominent in *bk* than *b*. By leveraging the topic information, DAH-CRF+LDA$_{utt}$ can better handle the confusion cases and hence improve the prediction for *bk* significantly.

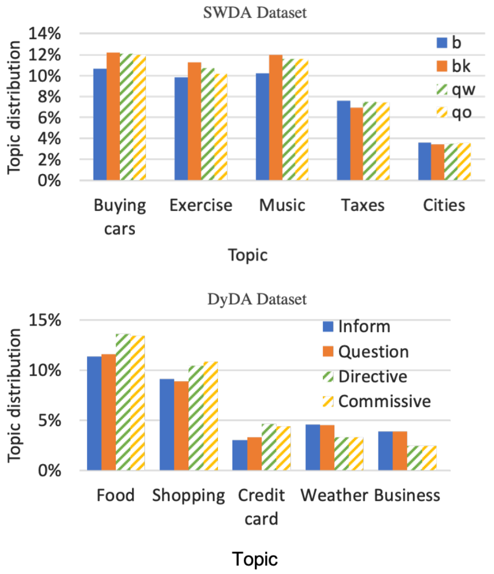

Figure 3: We highlight the prominent topics for some example DAs. The topic distribution of a topic k under a DA label d is calculated by averaging the marginal probability of topic k for all utterances with the DA label d.

There are also cases where DAH-CRF+LDA$_{utt}$ performs worse than SAH-CRF. Take the DA pair of *qo* (`Open Question`) and *qw* (`wh-questions`) as an example. *qo* refers to questions like '*How about you?*' and its variations (e.g., '*What do you think?*'), whereas *qw* represents wh-questions which are much

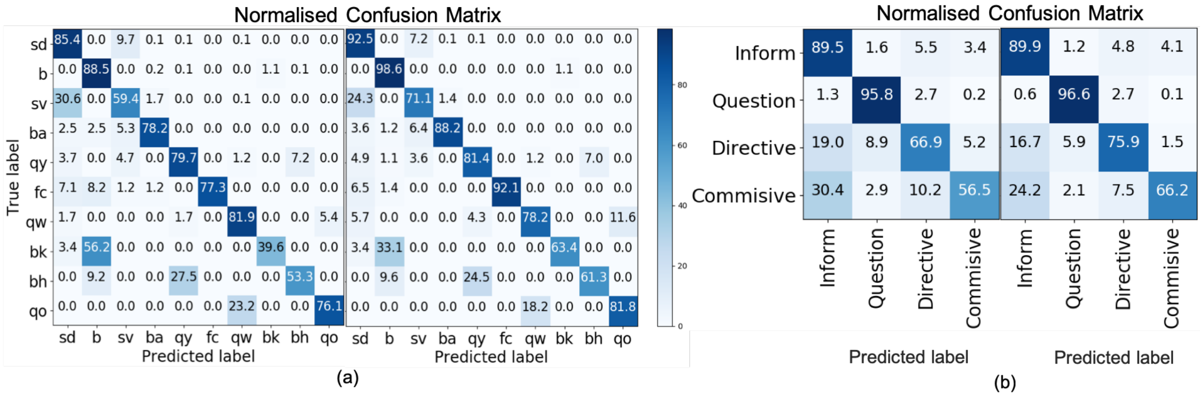

Figure 4: The normalized confusion matrix of DAs using SAH-CRF (left) and DAH-CRF+LDA$_{utt}$ (right) on SWDA (a) and DyDA (b).

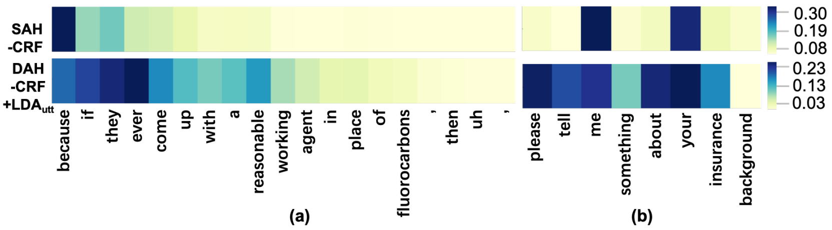

Figure 5: DA Attention visualisation using SAH-CRF and DAH-CRF+LDA$_{utt}$ on (a) SWDA and (b) DyDA datasets. The true labels of the utterances above are *sd* (*statement-non-opinion*) and *Directive*, respectively. SAH-CRF misclassified the DA as *sv* (*statement-opinion*) and *Inform* whereas DAH-CRF+LDA$_{utt}$ gives correct prediction for both cases.

more specific in general (e.g. '*What other long range goals do you have?'*). SAH-CRF gives quite decent performance in distinguishing *qw* and *qo* classes. This is somewhat reasonable, as linguistically the utterances of these two classes are quite different, i.e., the *qw* utterance expresses very specific question and is relatively lengthy, whereas *qo* utterances tends to be very brief. We see that DAH-CRF+LDA$_{utt}$ performs worse than SAH-CRF: a greater number of *qw* utterances are misclassified by DAH-CRF+LDA$_{utt}$ as *qo*. This might be attributed to the fact that topic distributions of *qw* and *qo* are similar to each other (see Figure 3), i.e., incorporating the topic information into DAH-CRF may cause these two DAs to be less distinguishable for the model.

We also conducted a similar analysis on the DyDA dataset. As can be seen from the confusion matrices shown in Figure 4, DAH-CRF+LDA$_{utt}$ gives improvement over SAH-CRF for all the four DA classes of DyDA. In particular, `Directives` and `Commissive` achieve higher improvement margin compared to the other two classes, where the improvement are largely attributed to less number of instances of the `Directives` and `Commissive` classes being mis-classified into `Inform` and `Questions`. Examining the topic distributions in Figure 3 reveals that `Directives` and `Commissive` classes are more relevant to the topics such as *food*, *shopping*, and *credit card*. In contrast, the topics of `Inform` and `Questions` classes are more about *business*, and *weather*.

Finally, Figure 5 shows the DA attention visualisation examples of SAH-CRF and DAH-CRF+LDA$_{utt}$ for an utterance from SWDA and DyDA. For SWDA, it can be seen that SAH-CRF gives very high weight to the word "because" and de-emphasizes other words. However, DAH-CRF+LDA$_{utt}$ can capture more important words (e.g., "if", "reasonable", etc.) and correctly predicts the DA label as *sd*. For DyDA, SAH-CRF only focuses on "me" and "your", but DAH-CRF+LDA$_{utt}$ captures more words relevant to `Directive`, such as "please", "tell", etc. To summarise, DAH-CRF+LDA$_{utt}$ can capture more significant words related to the corresponding DA, by modelling both DAs and topic information with the dual-attention mechanism.

6 Conclusion

In this paper, we developed a dual-attention hierarchical recurrent neural network with a CRF for DA classification. With the proposed task-specific dual-attention mechanism, our model is able to capture information about both DAs and topics, as well as information about the interactions between them. Moreover, our model is generalised by leveraging an unsupervised model to automatically acquire topic labels. Experimental results based on three public datasets show that modelling utterance-level topic information as an auxiliary task can effectively improve DA classification, and that our model is able to achieve better or comparable performance to the state-of-the-art deep learning methods for DA classification.

We envisage that our idea of modelling topic information for improving DA classification can be adapted to other DNN models, e.g., to encode topic labels into word embeddings and then concatenate with the utterance-level or conversation-level hidden vectors of our baselines, e.g. SelfAtt-CRF. It will also be interesting to explicitly take into account speaker's role in the future.

Acknowledgment

This work is supported by the award made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P011829/1).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mansurul Bhuiyan, Amita Misra, Saurabh Tripathy, Jalal Mahmud, and Rama Akkiraju. 2018. Don't get Lost in Negation: An Effective Negation Handled Dialogue Acts Prediction Algorithm for Twitter Customer Service Conversations. In *Proc. of ICWSM workshop on Chatbots*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234. ACM.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

Alfred Dielmann and Steve Renals. 2008. Recognition of dialogue acts in multiparty meetings using a switching DBN. *IEEE transactions on audio, speech, and language processing*, 16(7):1303–1314.

Yulan He, Chenghua Lin, and Amparo Elizabeth Cano. 2012. Online sentiment and topic dynamics tracking over the streaming data. In *International Confernece on Social Computing*, pages 258–266.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. 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. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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)*, pages 986–995.

Chenghua Lin. 2011. *Probabilistic topic models for sentiment analysis on the Web*. Ph.D. thesis, University of Exeter.

Yang Liu. 2006. Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus. In *Ninth International Conference on Spoken Language Processing*.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using Context Information for Dialog Act Classification in DNN Framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

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 (Volume 1: Long Papers)*, pages 1064–1074.

Michael Paul and Mark Dredze. 2012. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*, pages 2582–2590.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca 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.

Rakesh Verma, Narasimha Shashidhar, and Nabil Hossain. 2012. Detecting phishing emails the natural language way. In *European Symposium on Research in Computer Security*, pages 824–841. Springer.

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.

Byron C Wallace, Thomas A Trikalinos, M Barton Laws, Ira B Wilson, and Eugene Charniak. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1775.

Britta Wrede and Elizabeth Shriberg. 2003. Relationship between dialogue acts and hot spots in meetings. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 180–185. IEEE.