Dialogue Act Classification Methods in Natural Language Processing

Omar ELGHAFFOULI*[1] and Salwa HAJJI*[1]
https://nlp-ensae.github.io/[1]

[1]ENSAE Paris
[1]firstname.lastname@ensae.fr

Abstract

Speech Act Classification, which consists on determining the communicative intent of an utterance, has been investigated widely over the past few years as a standalone task due to the tremendous growth of NLP-based systems and AI chat assistants such as ChatGPT. It aims to classify an utterance with respect to the function it serves in a dialogue, i.e. the act the speaker is performing. In this paper, we focus on building a dialogue act classifier using the Meeting Recorder Dialogue Act Corpus (MRDA). We approach the problem as a sequence labeling task by using a recurrent neural networks model bidirectional LSTM (BiLSTM) with different embedding models. we add also a context-aware self-attention property and compare results with baseline performance in literature. We notice an increase in model accuracy using the BERT encoding model with self-attention and context-aware mechanism.

1 Introduction

Dialogue act (DA) classification is an important task in natural language processing that involves identifying the intended purpose or function of an utterance in a conversation. The task involves categorizing the different types of communicative acts that people perform during a conversation, such as making a statement, asking a question, expressing an opinion, or making a request[1].

Dialog act (DA) classification is a fundamental task in natural language processing that enables a conversational agent to understand the user's intent and generate an appropriate response. By identifying the type of speech act being performed in a given utterance (e.g., question, statement, request), the agent can tailor its response to better meet the user's needs. This is particularly important in spoken dialog systems (Dinkar* et al., 2020), where the goal is to create an engaging and effective conversation between the user and the agent.

Without the ability to classify DAs, the agent would not be able to effectively condition (Modi et al., 2020; Jalalzai* et al., 2020) its response based on the user's intent. For example, if a user asks a question but the agent responds with a statement, the conversation may quickly become confusing or frustrating for the user. However, by correctly identifying the user's DA and generating an appropriate response, the agent can create a more natural and effective dialogue.

Overall, DA classification plays a critical role in the development of conversational agents. By accurately identifying and understanding the user's intent, these systems can generate more appropriate and effective responses, leading to better user experiences and increased engagement.

DA classification is approached either as a sequence labeling task or a text classification problem. Deep learning models were widely used for DA (Lee and Dernoncourt, 2016),(Li and Wu, 2016). However, machine learning models such as logistic regression or naive bayes have also been used and showed an acceptable performance. (Lendvai and 2, 2007).

2 Related Work

There are two main classes of approaches which have been used in DA classification. the first considers the problem as a classification task where each utterance is classified in isolation. (Lee and Dernoncourt, 2016) have used CNN with 84.6% accuracy and (Lendvai and 2, 2007) have used Naive Bayes with 82% accuracy. The second

[*]stands for equal contribution.

Speaker	Utterance	DA label
A	Okay.	Other
A	Um, what did you do this weekend?	Question
B	Well, uh, pretty much spent most of my time in the yard.	Statement
B	[Throat Clearing]	Non Verbal
A	Uh-Huh.	Backchannel
A	What do you have planned for your yard?	Question
B	Well, we're in the process of, revitalizing it.	Statement

Figure 1: A snippet of a conversation sample from the MRDA Corpus. Each utterance has a corresponding dialogue act label.

approaches the problem as a sequence labeling task learning dependencies from previous utterances (Li and Wu, 2016). The most used models for this task are recurrent neural networks and more precisely BiLSTMs along with an embedding model. A research team in IBM India has built a hierarchical recurrent neural network using BiLSTM as a base unit and the conditional random field (CRF) as the top layer to classify each utterance into its corresponding dialogue act and achieved 90.9% of model accuracy (IBM Research, 2017) . Others have also tried to increase BiLSTM performance by integrating the context using Attention or self-attention mechanisms. (Tetreault, 2017). Scientists had also been interested in understanding different speech features such as fillers which carry valuable information about the speaker's level of confidence, hesitation, or uncertainty (Tanvi Dinkar, 2020). Others had worked on the switching languages in the same conversation issue as there is a growing interest in understanding dialogue in a multilingual fashion (Emile Chapuis, 2021).

3 Problem Framing

In this paper, we build a DA classifier using different techniques for sequence labeling found in literature. We then compare them based on each model accuracy with baseline model performance. Before describing the used models in detail, the sequence labeling problem can be modeled as set D of N conversations where $D = (C^1, C^2, ..., C^N)$ with $(Y^1, Y^2, ..., Y^N)$ the corresponding target. Each conversation C^i itself is a sequence of R_i utterances $C^i = (u_1, u_2, ..., u_{R_i})$ with $Y^i = (y_1, y_2, ..., y_{R_i})$ the corresponding target.

This means that for each utterance u_j in each conversation , we have a corresponding target label y_j. Each utterance u_j in return is itself a sequence of S_j words stringed together, ie., $u_j = (w_1, w_2, ..., w_{S_j})$.

The overall architecture of our model involves four main components : (1) encoding information within the utterances using GloVe (Jeffrey Pennington and Manning, 2014), Keras or BERT (Devlin, 2019) embedding; (2) a word-level Hierarchical Recurrent Encoder using BiLSTM instead of a linear transformation (IBM Research, 2017); (3) a context-based and attention feature concatenating word representation into utterances representations; (4) a conversation-level including a Softmax function in final layer for classification. We provide a schema of those blocs below for a detailed description.

4 Data

We train and assess our models on the Meeting Recorder Dialogue Act Corpus (MRDA) (Ang and Shriberg, 2005). It contains 72 hours of naturally occurring multi-party meetings that were first converted into 75 word level conversations, and then hand annotated with 11 general Tags/labels and 39 specific tags. In this work and for the sake of simplicity, we use the basic annotation which contains 5 classes : Statements, Questions, Floorgrabber, Backchannel, Disruption.

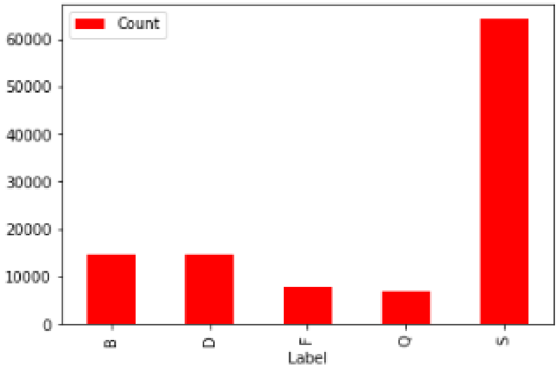


Figure 2: Basic tags distribution in MRDA corpus. Statements(S); Questions(Q); Floorgrabber(F); Backchannel(B); Disruption(D).

The Data seems to be balanced. In a natural conversation, statements are more frequently occurring than questions or backchannels. The use of accuracy as an evaluation metric is legit.

Model	Acc(%)
Hierachical BiLSTM-CRF (IBM Research, 2017)	90.9
LSTM-softmax (Khanpour and Nielsen, 2016)	86.8
CNN (Lee and Dernoncourt, 2016)	84.6
Naiive Bayes (Lendvai and 2, 2007)	82.0
BiLSTM + Attention+ Context (Tetreault, 2017)	87.7

Table 1: Literature review of model performance in related work papers.

5 Experiments Protocol

In this section, we describe the modeling protocol deployed in the notebook code. We use five different ways to build the DA classifier. (A) GloVe Embedding + BiLSTM; (B) Keras Embedding + BiLSTM; (C) Keras Embedding + BiLSTM + self-attention; (D) Keras Embedding + BiLSTM + self-attention, context-awareness; (E) BERT Embedding + BiLSTM + Attention + Context-awareness.

In each case, we either use a different embedding or add context and attention mechanisms to the BiLSTM in order to achieve the highest accuracy.

5.1 Utterance level encoding

For each word in an utterance, we use different word embeddings such as GloVe (Jeffrey Pennington and Manning, 2014) or BERT (Devlin, 2019). The word embedding is then followed by a BiLSTM layer that serves as input to the utterance-level context-aware self-attention mechanism which learns the final utterance representation.

5.2 Context-awareness self-attention encoding

We use the previous hidden state from the last layer, which we will explain in the next section, to provide the context of the conversations. This is combined with the hidden states of all the words in an utterance using a self-attentive encoder (Lin and al, 2021), which computes a $2D$ representation of each utterance. The output is then fed to a linear layer and concatenated with the previous utterance output[3].

5.3 Last layer

To incorporate context dependence of the previous utterance, we feed outputs of the previous step into a self-attention encoder and then we use the Softmax activation function for DA classification. The use of the Softmax has several advantages and was mainly based on its ability to handle multiclass classification problems and provide probabilistic output, which can be useful for decision-making in conversational systems.

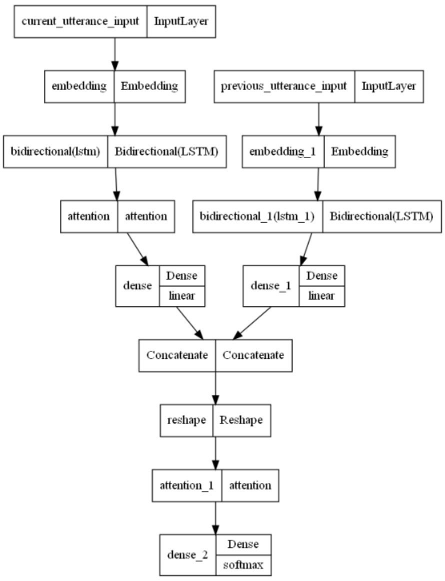


Figure 3: General modeling architecture

5.4 hyperparameter tuning

The training process is performed in mini-batches of a maximum batch-size of 32 example and each utterance was padded to a maximum length of 100. We used $L2$ regularization of $1e-3$. The word vectors were initialized with the 300-dimensional. Dropout was applied to the embedding obtained from the output of each encoder. The learning rate is set by default by the keras library. Early stopping is also used on the validation set with a patience of 3 epochs. We use the adam optimizer with the loss function set to categorical cross entropy. The validation metric is the accuracy. We did not use any specific method to optimize the number of layers or neurons in each layer. They were set arbitrarily.

6 Results

Based on the hyperparameters tuning discussed in the previous section, we compare the approach composed of BERT embedding, BiLSTM, self-attention and context-awarness with four other approaches. We observe a remarkable increase in accuracy as shown in table 2.

Note that in this project, we did not reinvent the wheel. All these models already exist. We have simply tried different settings and combinations to achieve the best model accuracy, as in ChatGPT. Nothing new or disruptive but only well arranged according to Yunn LeCan.

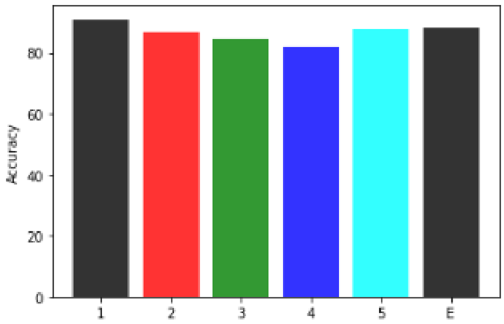
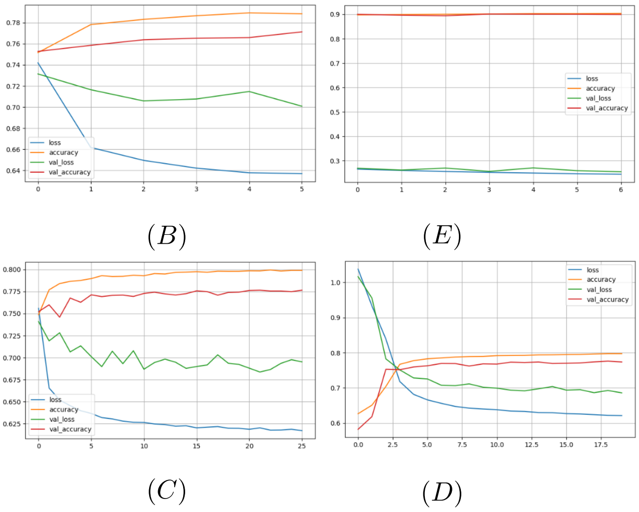


Figure 4: Accuracy comparison: 1 : (IBM Research, 2017); 2 : (Khanpour and Nielsen, 2016); 3:(Lee and Dernoncourt, 2016); 4 : (Lendvai and 2, 2007); 5 : (Tetreault, 2017); E : 2

.

We used then methods referenced in table 1 as baseline models. Notice that our model is performing as nearly as (Tetreault, 2017) because they are the same. However, based on the data processing techniques , the embedding models and the choice of hyperparameters, the accuracy obtained may differ.

As we can see in figure 4, our approach's accuracy performance (E) is lower than the one of the Hierachical BiLSTM-CRF model (IBM Research, 2017) which incorporates also the self-attention context-awarness mechanism. This is due to the use of CRF layer which enables to model as well dependencies among labels, aside the dependencies among utterances which has already been captured by the bidirectional encoders.



(B) (E)

(C) (D)

Figure 9: Evolution of the cross entropy loss and model accuracy both on train and validation data over number of epochs.

To mitigate the problem of computational resources, we used the ENSAE SSP cloud service with GPU access. As shown in figure 9, our model's accuracy and loss converged after few epochs with a fixed batch-size of 32.

7 Discussion and Conclusion

The BERT embedding (Devlin, 2019) has shown good performance compared to other models. We used the same architecture across all our approaches to make the fairest possible comparison. Our goal was to find the best recipe and combination to achieve the best model accuracy.

The proposed architecture effectively captures long-term dependencies between words within an utterance as well as across different utterances, enabling the generation of vector representations for each utterance in a conversation. The efficacy of capturing dependencies, whether at the word-level or at the utterance-level, is highly dependent on the data encoding and the encoding model's ability to capture relevant dependency features. This is reflected in the varying levels of accuracy observed across the different implemented approaches.

Our work can be extended by including the CRF based classifier to model the dependencies between the Dialog Act labels and the utterance representations (IBM Research, 2017) and evaluate the models on other datasets as SwDA.

Modeling approaches	Acc
GloVe embedding + BiLSTM (A)	54%
Keras embedding + BiLSTM (B)	77%
Keras embedding + BiLSTM + Attention (C)	78%
Keras embedding + BiLSTM + Attention + Context (D)	78%
BERT embedding + BiLSTM + Attention + Context (E)	88%

Table 2: DA classifier accuracy of modelling approaches

In the context of dialog systems, DA classification is a crucial task that enables the system to understand the user's intention and provide an appropriate response. Many approaches have been proposed to solve the DA classification task, ranging from traditional machine learning algorithms to deep learning models such as neural networks. These models vary in their architecture, training methods, and input representation.

As the field of natural language processing continues to evolve, researchers are exploring new ways to improve the performance of DA classification models. One such approach is the incorporation of multimodal data (Garcia* et al., 2019; Colombo et al., 2021), such as images or video, to supplement the textual data used in classification. Adding multimodal data can provide additional context that can help disambiguate user intent and improve model accuracy.

As sentiment analysis (Witon* et al., 2018; Colombo* et al., 2019) can be used for a variety of applications, we can consider, in further work, incorporating emotion detection in the previous implemented models as a next step in trying to better understand the context of the conversations and the speakers themselves.

.

References

Liu Ang and Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. *ICASSP*.

Lendvai and Geertzen 2. 2007. Token-based chunking of turn-internal dialogue act sequences. *In SIGDIAL Workshop on Discourse and Dialogue*.

Richard Socher Jeffrey Pennington 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. Association for Computational Linguistics*.

Guntakandla Khanpour and Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. *n COLING*.

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1970–1979. The COLING 2016 Organizing Committee*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural network. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 515–520. Association for Computational Linguistic*.

Vipul Raheja Joel Tetreault. 2017. Dialogue act classification with context-aware self-attention. *Dialogue Act Classification with Context-Aware Self-Attention*.

India IBM Research. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. *Dialogue Act Sequence Labeling using Hierarchical encoder with CRF*.

Wojciech Witon*, Pierre Colombo*, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at iest 2018: Predicting emotions using an ensemble. In *Wassa @EMNP2018*.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

al Devlin. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Google AI Language*.

Alexandre Garcia*, Pierre Colombo*, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. *EMNLP 2019*.

Ashutosh Modi, Mubbasir Kapadia, Douglas A Fidaleo, James R Kennedy, Wojciech Witon, and Pierre Colombo. 2020. Affect-driven dialog generation. US Patent 10,818,312.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*.

Matthieu Labeau Chloe Clavel Tanvi Dinkar, Pierre Colombo. 2020. The importance of fillers for text representations of speech transcripts. *EMNLP*.

Matthieu Labeau Chloé Clavel Emile Chapuis, Pierre Colombo. 2021. Code-switched inspired losses for generic spoken dialogue representations. *EMNLP*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. *EMNLP 2021*.

Lin and al. 2021. Self-attention-based conditional random fields latent variables model for sequence labeling. *ELSEVIER*.