Comparing BERT and a BCBAH model for Dialogue act classification

Ayoub Ammy-Driss
ENSAE
ayoub.ammy-driss@ensae.fr

[1]

Abstract

Dialogue act classification is a key task in natural language processing that involves identifying the intended purpose or function of a particular utterance in a conversation. In recent years, deep learning models like BERT have achieved state-of-the-art performance on this task. However, the performance of BERT can still be improved by incorporating other deep learning models. In this report, we present a comparison between the performance of BERT and a BERT-CNN-BiGRU-Attention Hybrid (BCBAH) model on the *"dyda_da"* dataset from the SILICONE dataset for dialogue act classification (Emile Chapuis, 2021). The hybrid model combines the strengths of different deep learning models to improve the accuracy and efficiency of the task. We conducted experiments on the dataset to evaluate the performance of both models.

1 Introduction

Dialog Act classification is a crucial component of chatbot technology that enables machines to understand and respond to natural language inputs from users (Colombo* et al., 2020). A Dialog Act is an action performed by a speaker in a conversation, such as making a statement, asking a question, giving an opinion, expressing agreement or disagreement, or making a request (Li et al., 2017; Leech and Weisser, 2003; Busso et al., 2008; Passonneau and Sachar., 2014; Thompson et al., 1993; Poria et al., 2018; Shriberg et al., 2004; Mckeown et al., 2013).

Dialog Act classification involves identifying the intent behind a user's message and categorizing it into one of several predefined categories. These categories typically correspond to common conversational actions such as requesting information, giving instructions, or expressing gratitude (Colombo* et al., 2019).

Chatbots like Siri rely heavily on Dialog Act classification to accurately understand and respond to user inputs. By analyzing the language used by the user and categorizing it according to the appropriate Dialog Act, chatbots can generate responses that are relevant, informative, and engaging (Colombo, 2021).

Overall, Dialog Act classification is an important tool for enhancing the natural language processing capabilities of chatbots and improving the quality of the user experience. However, due to the complexity and variability of human language, dialogue act classification is challenging (Colombo et al., 2021a).

2 Related Work

To tackle DA classification, various approaches have been proposed in the literature, including rule-based methods, machine learning-based methods, and deep learning-based methods. Deep learning models, in particular, have shown remarkable performance on this task in recent years. The introduction of BERT(Devlin, 2018) marked a significant milestone. BERT utilizes a transformer-based architecture to capture contextual relations in language modeling tasks and can be fine-tuned for various downstream tasks, including dialogue act classification. BERT was evaluated on various natural language understanding tasks, and the results were impressive. This model outperformed other existing models, including those based on convolutional and recurrent neural networks. But BERT is not the only revolutionary model in this field. Another game-changer is "Attention is All You Need"(Vaswani, 2017), which introduced the transformer architecture. This architecture uses

[1] https://github.com/AyoubAmmyDriss/NLP_ENSAE

self-attention mechanisms to capture long-range dependencies in sequences without the need for recurrence or convolution operations. It has become a popular architecture for dialogue act classification. To further improve the accuracy and efficiency of short text classification tasks,(Tong, 2021) proposed a hybrid model that combined BERT with convolutional neural networks (CNNs) and attention-based bidirectional gated recurrent units (BiGRUs). The hybrid model outperformed existing models, including BERT, in terms of accuracy and efficiency. The authors of this paper showed that combining different deep learning models could improve the accuracy and efficiency of the task. In this project, we aim to replicate the results achieved in the paper by implementing and comparing the performance of BERT and a BCBAH hybrid model for natural language understanding tasks.

3 Dataset presentation

In this report, we focus on comparing the performance of two deep learning models for dialogue act classification: BERT and a BCBAH hybridmodel. We evaluate the performance of these models on the *"dyda_da"* dataset from the SILICONE dataset (Li et al., 2017; Leech and Weisser, 2003; Busso et al., 2008; Passonneau and Sachar., 2014; Thompson et al., 1993; Poria et al., 2018; Shriberg et al., 2004; Mckeown et al., 2013), which contains labeled examples of dialogue acts in various domains.

The *"dyda_da"* dataset consists of 102,979 utterances and 4 dialogue act labels(*"question"*, *"commissive"*, *"directive"*, *"inform"*). The dataset is split into training, validation, and test sets with 80%, 10%, and 10% of the examples, respectively.

Dialog Act	Count	Percentage
Inform	46532	45.2%
Question	29428	28.6%
Directive	17295	16.8%
Commissive	9724	9.4%

Table 1: Dialog Act Repartition in Dailydialog

4 Experiments Protocol

We conducted experiments to compare two models for dialogue act classification: BERT and a BCBAH hybrid model. The hybrid model incorporates different techniques to improve performance on short text classification tasks. We compared the accuracy of both models on dialogue act classification tasks.

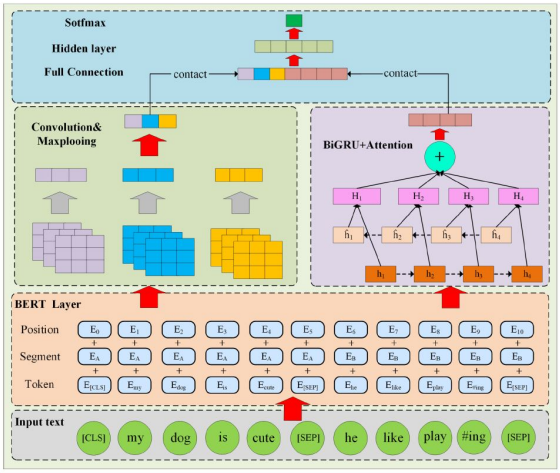

Figure 1: Model Architecture

To input preprocessed text sequences into the hybrid model, we first passed them through the BERT embedding layer to obtain fixed-length vectors for each word. These embeddings were then simultaneously fed into both the CNN and BiGRU layers. The CNN layer applied filters to extract local features from the embeddings, while the BiGRU layer processed the sequence in a bidirectional manner to capture global context. The outputs of BiGRU layers were then concatenated and fed into an attention layer, and then concatenated with the output of the CNN layers. Finally, the output of the attention layer was passed through a fully connected layer to obtain the predicted class probabilities.

Overall, we compared the performance of the BERT model and the BCBAH hybrid model on dialogue act classification tasks and evaluated their accuracy. To provide a more in-depth analysis of our models and the evaluation metrics, we will now delve into each component.

4.1 BERT Tokenizer

Before text can be processed by the BERT model, it must first be tokenized into individual pieces that the model can understand. BERT uses a specialized tokenizer that performs a technique called subword tokenization, which breaks down words into smaller units that can be represented by the model.

The BERT tokenizer also adds special tokens to the beginning and end of the input sequence. The [CLS] token is added to the beginning of the sequence and is used as a representation of the entire input sequence for certain downstream tasks. The [SEP] token is added between sentence pairs to indicate the end of one sentence and the beginning of the next.

4.2 BERT & Word embeddings

To tackle NLP tasks, the first crucial step is to translate text into a format that machines can understand. This step is called word embeddings. There are various ways to achieve this, and in our study, we opted to leverage the power of the BERT model. In BERT, each word is transformed into a fixed-length vector representation by passing it through a deep neural network. Specifically, BERT uses a transformer-based architecture, which consists of multiple layers of self-attention and feedforward neural networks. The model is trained on a large corpus of text, using an unsupervised learning objective known as the masked language modeling (MLM) task.

4.3 Convolutional Neural Networks (CNN)

The main idea behind CNNs is to apply filters to the input data to extract local features. In the case of text data, these filters can be viewed as sliding windows of fixed length that move over the text sequence and perform element-wise multiplications with the input data. The result of this operation is a new sequence of values, called a feature map, which captures the presence or absence of certain patterns in the input data.

After the convolution operation, a pooling operation is applied to the resulting feature map to reduce its dimensionality and capture the most salient features. In our code, we used Max Pooling.

In our model, we applied multiple filters with different window sizes to the input data to capture features at different scales. The resulting feature maps were then concatenated.

4.4 BiGRU Layer

In addition to the CNN layer, our model also incorporates a bidirectional gated recurrent unit (BiGRU) layer.

The basic idea behind BiGRUs is to use a hidden state that is updated at each time step, and which contains information about the previous inputs that have been processed. BiGRUs are a variant of GRUs that process the input sequence in both forward and backward directions, allowing them to capture both past and future dependencies.

The output of the BiGRU layer is a sequence of hidden states, which contain information about both the forward and backward context of each input token. These hidden states are then fed into the attention layer, which assigns different weights to the features based on their importance for the classification task.

4.5 Attention Mechanism

After the BiGRU layers, we add an attention layer to further improve the classification performance.

4.6 Fully Connected Layer

The output of the attention layer is concatenated with the output of the CNN layer and passed through a fully connected layers to obtain the predicted class probabilities.

5 Results

5.1 BERT Model

Looking at the classification report for the validation set, we can see that the Bert model has performed well on all classes. The model has the highest f1-score for the inform class (0.93), which means it can accurately identify informative utterances. The directive and commissive classes also have high f1-scores, indicating that the model can classify these types of utterances with high accuracy. However, the f1-score for the question class is comparatively lower (0.54) for the Bert model.

Class	Precision	Recall	F1-score	Support
Question	0.59	0.49	0.54	925
Commissive	0.74	0.72	0.73	1775
Directive	0.80	0.84	0.82	3125
Inform	0.92	0.94	0.93	2244
Accuracy	0.80			
Macro Avg	0.76	0.75	0.75	8069
Weighted Avg	0.80	0.80	0.80	8069

Table 2: Classification Report on Validation

5.2 BCBAH

5.2.1 Parameters

The hyperparameters used in this study were based on a BERT-based hybrid short text classification

Hyperparameter	Value
Embedding Dimension	768
Hidden Dimension	128
Number of Layers	2
Number of Out Channels	16
Kernel Sizes	[3, 4, 5]
Dropout Rate	0.5

Table 3: Hyperparameters Used in the BERT-CNN-BiGRU Hybrid Model

Class	Precision	Recall	F1-score	Support
Question	0.59	0.50	0.54	925
Commissive	0.92	0.94	0.93	2244
Directive	0.79	0.85	0.82	3125
Inform	0.76	0.71	0.73	1775
Accuracy	0.80			
Macro Avg	0.77	0.75	0.76	8069
Weighted Avg	0.80	0.80	0.80	8069

Table 4: Classification Report on Validation

model that incorporated CNN and attention-based BiGRU, as described in (Tong, 2021). The embedding dimension was set to 768 to match the BERT embeddings used in the model. The hidden dimension was set to 128, with two layers, and the number of output classes was determined by the dataset being used. The CNN component of the model utilized 16 output channels and kernel sizes of 3, 4, and 5. Finally, a dropout rate was applied to the model to prevent overfitting. These hyperparameters were chosen based on the performance of the model in the original paper and were adapted to fit our own dataset.

5.2.2 Results

To evaluate the performance of the model on the validation set, we looked at the classification report. The results showed that the BCBAH has achieved high f1-scores for the commissive, directive, and inform classes, indicating that it can accurately classify these types of utterances. However, the f1-score for the question class is comparatively lower, which means the model struggles to identify question utterances.

The precision and recall values for the BCBAH are consistent across all classes, with higher precision and recall values for the inform class. This means that the model can accurately identify informative utterances.

Overall, the Bert CNN BiGRU Attention Hybrid model has shown comparable performance to the Bert model.

6 Discussion/Conclusion

In this study, we evaluated the performance of two models, BERT and BERT-CNN-BiGRU Hybrid, on a text classification task. Our results indicate that both models achieved similar accuracy and F1-scores on the validation dataset. Specifically, BERT achieved an accuracy of 80%, while the hybrid model achieved an accuracy of 79%. Furthermore, the F1-scores for each class were also comparable between the two models. These findings suggest that the added complexity of the hybrid model did not result in any significant improvements in performance.

Based on our results, we conclude that for this text classification task, using a simpler model like BERT is sufficient to achieve good performance. While it is important to consider more complex models in certain scenarios, such as when dealing with larger datasets or more complex classification problems, our findings suggest that in this case, the simpler model is adequate. Additionally, using a simpler model can also result in faster training times and reduced computational resources, which may be important considerations in practical applications.

Overall, this study highlights the importance of evaluating the performance of different models on specific tasks, and suggests that more complex models may not always lead to significant improvements in performance. For the futur, it is important to consider fairness (Colombo et al., 2021b; Pichler et al., 2022; Colombo et al., 2022) in the design and implementation of dialog act models. Dialog acts are the linguistic actions performed by speakers in a conversation, such as asking a question or making a statement. These models can have significant impacts on social interactions, particularly in areas such as customer service and healthcare.

References

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: natural dialogue for speech recognition.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues.

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, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

R. Passonneau and E. Sachar. 2014. Loqui human-human dialogue corpus (transcriptions and annotations).

Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

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

Pierre Colombo*, Emile Chapuis*, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *AAAI 2020*.

Tong. 2021. A bert-based hybrid short text classification model incorporating cnn and attention-based bigru. *Journal of Organizational and End User Computing Volume 33 Issue 6*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021a. Code-switched inspired losses for spoken dialog representations. In *EMNLP 2021*.

Matteo Manica Matthieu Labeau Chloe Clavel1 Emile Chapuis, Pierre Colombo. 2021. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *ICML 2022*.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*.