Weekend at BERTie's*: A Hierarchical Encoding For Dialog Act Classification

Rayane Mouhli †
ENSAE Paris
rayane.mouhli@ensae.fr

Thibaut Valour †
ENSAE Paris
thibaut.valour@ensae.fr

Abstract

Conversational agents like Siri, Alexa and ChatGPT have gained immense popularity due to their ability to comprehend the type of information being conveyed by the user and generate appropriate responses based on context. This task is known as Dialog Act Classification. In this paper, we propose a hierarchical encoding strategy to tackle this problem. We use a BERT model to encode each utterance of a dialog, then, a BiLSTM encodes the encoded utterances from a same dialog. Since we encode at the utterance-level then at the dialog level, our model has a well-understanding of the context of a sentence to classify it. We conducted our experiments on the dyda_da dataset from the SILICONE benchmark developed by HuggingFace, which contains everyday communication styles and diverse topics related to daily life. Our model outperforms the baseline BERT model, achieving an accuracy of 85% on the validation set. We also analyse the influence of context and imbalanced data on the performance of the model.

1 Context of the problem

Conversational agents have become powerful tools that significantly help people in their daily lives, as demonstrated by the widespread adoption of Siri, Alexa, and ChatGPT in recent years. A key factor contributing to their success is their ability to comprehend the type of information being conveyed by the user. This distinction is crucial, as the response generated will differ depending on whether the user requests a joke or inquires about the weather for the afternoon. This particular task is referred to as Dialog Act Classification. Numerous studies have explored language understanding at the sentence-level (Kim, 2014), focusing on determining whether a given sentence is a declarative statement, a question, and so on. In this project, we aim to expand this framework to the dialog-level (Chapuis et al., 2020; Colombo et al., 2021; Colombo* et al., 2020), which presents a more complex challenge as the context can alter the meaning of a word. The ultimate objective is to classify each sentence within a dialog.

Using the notation of (Colombo, 2021), let us consider $\mathcal{D}_i = (u_1, ..., u_{|\mathcal{D}_i|})$ a dialog with u_k the k-th utterance and considering that each utterance is composed of words $u_k = (w_1, ..., w_{|u_k|})$, the goal is to predict the labels associated to each utterance $\mathcal{Y}_i = (Y_1, ..., Y_{|\mathcal{D}_i|})$.

2 Experiments Protocol[1]

2.1 Dataset

We conducted our experiments using the SILICONE (Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE) benchmark from Hugging Face (Wolf et al., 2020), which comprises 10 datasets of various sizes (Godfrey et al., 1992; 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; Dinkar* et al., 2020; Mckeown et al., 2013), focusing on dialog act or sentiment analysis.

Specifically, we worked with the dyda_da (DailyDialog Act Corpus) dataset. The dialogs in this dataset mirror everyday communication styles and encompass a wide range of topics related to daily life. The dataset is annotated with both communication intention and emotion information, though our experiments centred on the communication intention task. Each sentence is categorized as either ”commissive”, ”directive,” ”inform,” or ”question,” corresponding to labels 0, 1, 2, and 3 in the table below.

[*] Weekend at Bernie's, 1989

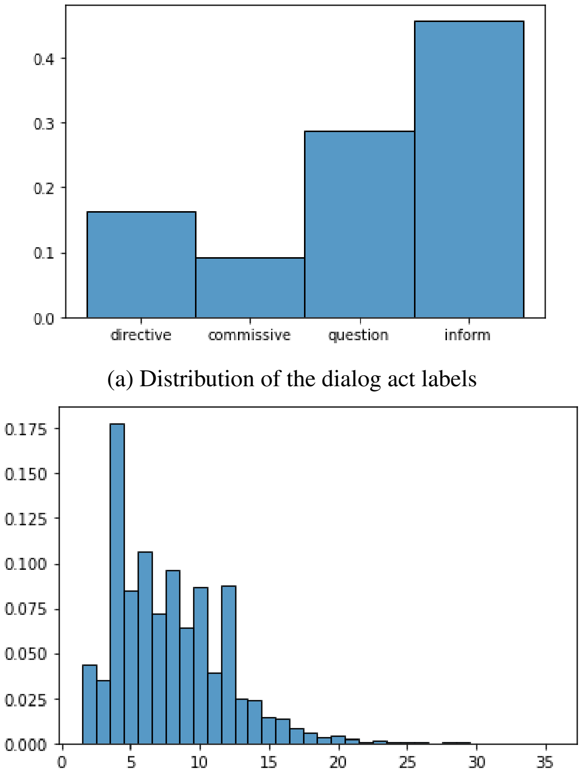
[†] stands for equal contribution

[1] https://github.com/thibautvalour/NLP_Intent_classification.git

Utterance	DA	ID
say, jim, how about going for a few beers after dinner ?	1	1
you know that is tempting but is really not good for our fitness	0	1
what do you mean ? it will help us to relax !	3	1
can you do push-ups ?	3	2
of course i can. I can do 30 push-ups a minute.	2	2
really ? i think that's impossible !	3	2

Table 1: Extract from dyda_da (SILICONE) where DA represents the dialog act label and ID the id of the conversation.

The dataset is split into a train set (87 170 utterances) and a validation set (7 740 utterances). As illustrated in 1a, the distribution of labels in the dataset is imbalanced, reflecting the natural variation in the frequency of different sentence types in everyday conversations. We will discuss this aspect in section 3. On average, a dialog consists of 8 utterances, indicating that we are working with relatively brief conversations.



(a) Distribution of the dialog act labels

(b) Distribution of the number of utterances per dialog

Figure 1: Statistics on dyda_da (SILICONE)

2.2 Hierarchical encoding strategy

Our approach to tackle dialog act classification is grounded in a hierarchical encoding methodology, as demonstrated in (Chapuis et al., 2020). The architecture of our model is depicted in Figure 2.

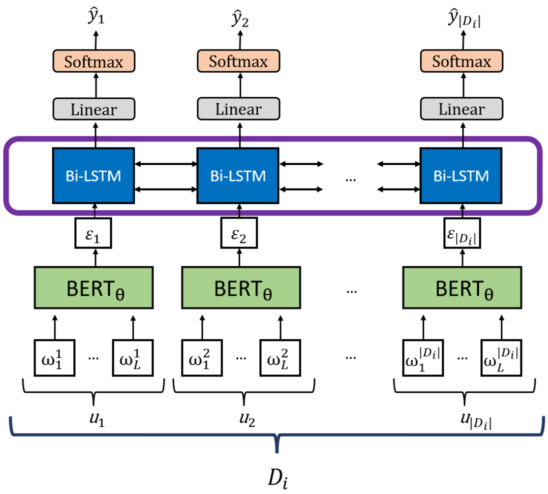


Figure 2: Architecture of our model

This architecture comprises two primary layers: BERT and BiLSTM (Staudemeyer and Morris, 2019). Initially, BERT encodes each utterance independently of its context, followed by concatenating the encoded utterances from the same dialog. The resulting vector, representing a single conversation, serves as input for the BiLSTM layer. This hierarchical architecture encodes at two levels: BERT at the utterance level and BiLSTM at the conversation level. Lastly, we incorporate a linear and softmax layer to generate the classification output.

2.2.1 BERT model

In 2018, Google AI researchers introduced a pretrained model for Natural Language Processing known as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). The primary concept is that bidirectionally trained language models can gain a more profound comprehension of the context of a discussion than unidirectional models. To achieve this, the researchers presented the notion of Masked Language Modelling (MLM) to facilitate bidirectional training.

BERT is built upon the attention mechanism of Transformers (Vaswani et al., 2017), enabling the understanding of the relationships between words within a sentence. More specifically, our focus lies on the encoder component of the Transformers. To accomplish this, a portion of the words intended for encoding is replaced with a [MASK] token, and the model attempts to predict the original value of the masked words using the remaining words.

2.2.2 BiLSTM

Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are a subclass of Recurrent Neural Networks (RNN). They are designed to overcome the long-term dependency issues encountered by traditional RNNs. LSTMs leverage the information provided by the first k-1 terms to predict the k-th term of a sequence, making them particularly useful for language-related problems. LSTMs employ three gates (forget, input, and output) to manage the retention or discarding of information. The latter are stored in a cell state, representing the current long-term memory of the network, and in the hidden state, which constitutes the output at the previous time step.

In dialog act classification, the objective is to classify an utterance based on the context of the dialog to which it belongs. Following the same idea as BERT, we utilize a bidirectional LSTM to comprehend the context of the utterance. BiLSTM involves combining two independent LSTMs to obtain both backward and forward information about the sequence at every time step. In our model, we stacked two layers of BiLSTM.

In summary, each utterance from a dialog is fed into a BERT model for encoding at the utterance level. The encoded utterances are then inputted into a BiLSTM that encodes and decodes at the conversation level. Finally, the linear and softmax layers provide the prediction for the classification task.

2.3 Training details

2.3.1 Padding

To utilize the BERT model, it is necessary to ensure that all input sentences have the same length. This is achieved through padding. Additionally, the [CLS] (classification) and [SEP] (separator) tokens are incorporated, as they were employed in the original BERT training to indicate the beginning and end of sentences, respectively. To further prepare our data for compatibility with the BERT-BiLSTM model, we also applied padding and truncation to the dialogs to create a uniform size, effectively formatting the data for seamless integration into our model. The padding size for dialog was chosen based on a trade-off between minimizing training time and maximizing information retention. By examining Figure 1b, we opted for a value of 16.

2.3.2 Dropout

Our model, particularly the BERT component, has a large number of trainable parameters (110M). Given that our dataset is relatively small, it is crucial to prevent overfitting. Dropout (Srivastava et al., 2014) is an effective method for addressing this issue. Dropout is a regularization technique where a proportion of neurons is randomly selected, and their outputs are temporarily set to zero during training. This prevents the model from relying too heavily on individual neurons, thus promoting generalization and reducing the risk of overfitting. We chose a dropout rate of 50%, striking a balance between maintaining the model's capacity to learn and preventing overfitting on the limited dataset. Dropout layers were incorporated at two points: between the BERT output and the LSTM, as well as between the LSTM output and the linear layer. These dropout layers are not depicted in Figure 2.

2.3.3 Freezing BERT Layers

During the training process of our BERT-BiLSTM model, we found that training all the parameters on a single GPU was computationally feasible, but the results were not satisfactory. To address this issue, we experimented with training only the n last layers of BERT, while still training all the parameters of the LSTM layers. This approach yielded improved performance. We optimized the hyperparameter n and found that the best results were achieved when $n = 3$.

3 Results

3.1 Baseline

First, we experiment a baseline model consisting in just a BERT at the utterance-level. It means that we ignore the context and we fed independently our model with sentences. The BERT model consists of 110M trainable parameters. It is then still possible to train all of them with a single GPU. After 2 epochs, the loss is constant, and we get a 67% accuracy on the validation set. Our results on the baseline are disappointing compared to the 82% accuracy achieves with BERT in (Chapuis et al., 2020).

3.2 BERT-BiLSTM Model

The hierarchical model, presented in the section 2.2, has been run with the hyperparameters presented in the table 2 in the Appendix. It has largely better results, since after 2 epochs of training we reach 85 % of accuracy on the validation set.

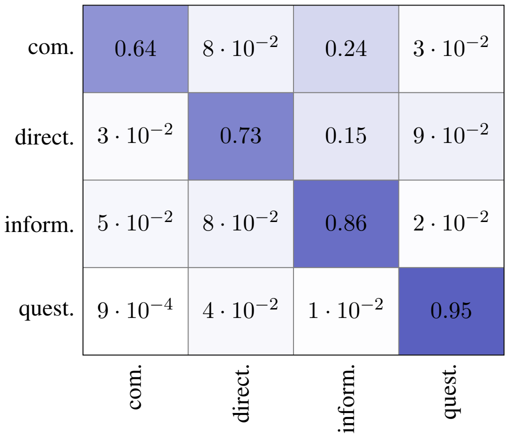
First, we observe that we succeed to improve the baseline thanks to the BiLSTM layer. It confirms the idea that the context adds some relevant information for the understanding of a sentence. In (Chapuis et al., 2020), they reach an 80% accuracy, which is slightly less than our model.

4 Discussion

4.1 Imbalanced data

In the section 2.1, we observed that the dataset was imbalanced. Indeed, 46% of the utterances are labelled 'inform', 29% are 'question', 16% are 'directive' and 9% are 'commissive'. Imbalanced data in the classification case can be problematic, since the model can predict poorly an underrepresented class without decreasing significantly the accuracy.

The confusion matrix reveals the model's performance across different classes in the dataset. In summary, while the model shows decent performance in identifying more prevalent classes, there is room for improvement when it comes to the least represented class.



	com.	direct.	inform.	quest.
com.	0.64	$8 \cdot 10^{-2}$	0.24	$3 \cdot 10^{-2}$
direct.	$3 \cdot 10^{-2}$	0.73	0.15	$9 \cdot 10^{-2}$
inform.	$5 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	0.86	$2 \cdot 10^{-2}$
quest.	$9 \cdot 10^{-4}$	$4 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0.95

Figure 3: Confusion matrix

To enhance our model's performance across various classes, we employed a weighted loss strategy, which utilizes the inverse proportion of each class in the training set as the weight. This method encourages the model to focus on learning from underrepresented classes. The resulting confusion matrix can be found in the appendix. While this approach does lead to improved accuracy for underrepresented classes, the overall loss and accuracy have not seen substantial improvements.

A potential avenue for further improvement could involve employing an iterative loss strategy. Initially, the loss would be weighted based on the inverse class proportion, but as training progresses, it would gradually transition toward a uniform loss after a specified number of epochs, as suggested by (ValizadehAslani et al., 2022).

4.2 Impact of context

To evaluate the influence of context in our model, we shuffle the sentences within each dialog in the validation set. The findings were unexpected, as the overall accuracy remained unchanged. This suggests that the order of the text is not crucial; instead, the overall context of the dialog is what matters. To investigate the significance of the global context further, additional researches would be required, such as incorporating sentences from different dialogs.

5 Conclusion

In conclusion, we have successfully developed a hierarchical BERT-BiLSTM model for the task of dialog act classification. Our model outperforms the baseline model, demonstrating the importance of incorporating context in understanding and classifying dialog acts.

Our approach showed that the order of sentences within a dialog may not be as significant as initially expected. However, additional researches are required to further explore the influence of global context on the model's performance. Additionally, our analysis on imbalanced data suggests that class imbalance does not significantly impact our model's performance.

This work contributes to the ongoing development of more effective conversational agents capable of understanding and responding to user input in a contextually relevant manner. In addition to the type of request made by the user, emotions (Witon* et al., 2018) can also play a critical role in generating an appropriate response. For example, if a user is feeling sad and requests a joke, the response generated should be empathetic and uplifting. On the other hand, if the user is feeling anxious about the weather forecast, the response generated should be informative and reassuring. Thus, taking into account the emotional state of the user is an important factor in creating a personalized and effective response (Colombo* et al., 2019; Jalalzai* et al., 2020).

References

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

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*.

Ralf C Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

Thomas Wolf, Quentin Lhoest, Patrick von Platen, Yacine Jernite, Mariama Drame, Julien Plu, Julien Chaumond, Clement Delangue, Clara Ma, Abhishek Thakur, Suraj Patil, Joe Davison, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angie McMillan-Major, Simon Brandeis, Sylvain Gugger, François Lagunas, Lysandre Debut, Morgan Funtowicz, Anthony Moi, Sasha Rush, Philipp Schmidd, Pierric Cistac, Victor Muštar, Jeff Boudier, and Anna Tordjmann. 2020. Datasets. *GitHub. Note: https://github.com/huggingface/datasets*, 1.

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*.

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*.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog.

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

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

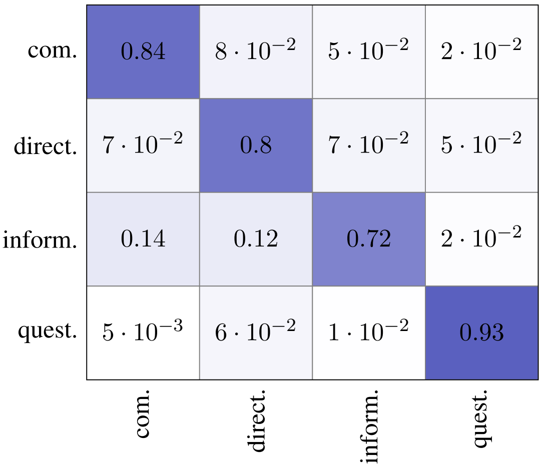
Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Taha ValizadehAslani, Yiwen Shi, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2022. Two-stage fine-tuning: A novel strategy for learning class-imbalanced data.

6 Appendix

Parameter	**Value**
Optimizer	Adam
Learning rate	10^{-4}
Epochs	3
Batch size	4
Sentence size after truncation/padding	32
Dialog size after truncation/padding	16
Trainable Bert layers	3
LSTM hidden dimension	768
LSTM layers	2
Dropout rate	0.5

Table 2: Model and training hyperparameters



	com.	direct.	inform.	quest.
com.	0.84	$8 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$2 \cdot 10^{-2}$
direct.	$7 \cdot 10^{-2}$	0.8	$7 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
inform.	0.14	0.12	0.72	$2 \cdot 10^{-2}$
quest.	$5 \cdot 10^{-3}$	$6 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0.93

Figure 4: Confusion matrix with the weighted loss