Intents Classification for Neural Text Generation

Sirine Louati *
ENSAE IP Paris
sirine.louati@ensae.fr

Jérémie Stym-Popper *
ENSAE IP Paris
jeremie.stym-popper@ensae.fr

Abstract

Sequence labelling tasks like Dialog Acts (DA) and Emotion/Sentiment (E/S) are very important in spoken dialog systems. In fact, they allow distinguishing different dialog acts and emotions from different conversations.
In this work, we propose an intent classifier that allows to identify different labels such as communicative intent or dialog acts from a given conversation. We evaluate this classifier on the SILICONE benchmark introduced by Chapuis et al. [2020]. Our experiments show that using an encoder with the BERT model enables to build a strong intent classifier achieving good performances. All our numerical experiments and codes are located here : https://github.com/Jeremstym/NLP_intent_class

1 Introduction

Sequence labeling tasks are a type of natural language processing (NLP) task that involves assigning a label or category to each element in a sequence of input data. The input data can be any kind of sequential data, such as words in a sentence, letters in a word, or sounds in speech.
In NLP, sequence labeling tasks are commonly used for tasks such as part-of-speech (POS) tagging, named entity recognition (NER), and sentiment analysis. In POS tagging, the task is to assign a grammatical tag (such as noun, verb, adjective, etc.) to each word in a sentence. In NER, the task is about identifying ang then classifying entities (such as people, organizations, and locations) in a text. In sentiment analysis, the task is to classify the sentiment or emotion expressed in a sentence or document.
We focus in this document on the third utility namely identifying communicative intent, dialog acts and emotions/sentiments. This identification is very useful in a way that it helps improving model performance on a dialog task. It especially avoids having unspecified responses that can fit to many utterances ([Colombo et al., 2019, 2021a]).
In this paper, we will first introduce some works that have been done before explaining our method and how it can obtain good performances. We will then set the experiments settings such as the dataset and the model used and we will finally expose the results.
The goal of this work is to train a supervised model on the SILICONE database [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] in order to perform an intent classification on Dialog Acts and identify questions, commissive, directive or informative acts.

2 Related works

A first work was done by Chapuis et al. [2020] where a hierarchical encoder is built to predict both the label of a text C_i (a conversation, a monologue, a letter, etc.) and the label of an utterance (*e.g.* in a dialogue, there are questions, answers, opinions etc.). and then, at a lower level, predict the label of an utterance (*e.g.* in a dia-

*equal contribution

logue, there are questions, answers, opinions etc.). First we will use the labels of this paper, *i.e.* Dialogue Acts (DA) and Emotion/Sentiment (E/S) for a conversation set $D = (C_1, \ldots, C_{|D|})$ and then the labels: questions (qw), statement non-opinion (sd) backchannel (b) and response acknowledgment (bk), where (b) concerns interjections and transition words. Besides, even though using large corpora can allow learning complex models like it was developed in Colombo et al. [2020], it also requires using more annotated data. As so, the work of Chapuis et al. [2020] showed that it can be not always optimal to use such large corpora and that using hierarchical encoders can even achieve high performances with less parameters while working on a reduced number of GPUs. Moreover, using hierarchical encoder has been proved in the work of Garcia et al. [2019] as being an efficient procedure to capture dependencies at different levels which is very important in dialog embedding because it allows reaching grained levels and it reduces the number of model parameters allowing in consequence faster learning.

3 Method

3.1 Limits of previous works

In the first article cited, we could discuss the labels used for the classification. Opposing DA and E/S allows to make a binary classification, which has many advantages such as simplicity of the model and readability of the metrics (confusion matrix, ROC curve, etc.). Moreover, the proposed labels are interesting because they encapsulate several modalities of a conversation, and then can enable a model to reproduce them. However, in our paper, we intend to clarify how an encoder can be very powerful to enable a classifier to identify all the above-mentioned labels. We will also check other metrics and analyze the performance of the model in terms of computation (number of data used, computation time).

3.2 Method

Let us recall here an important difference. Fine-tuning of a model (*e.g.* an encoder) means that we start with a pre-trained model and, during the training phase, every parameters are updated with respect to the labels. In this case we retain the whole model. On the other hand, the feature extraction techniques entail a pre-trained model as well, but only the final layer is updated, the layer from which the prediction are made. In our case, since it was less expensive, we adopted the last technique by adding an extra layer to the BERT model, as shown in the figure 1, even if we use the word "fine-tuning" for simplicity.

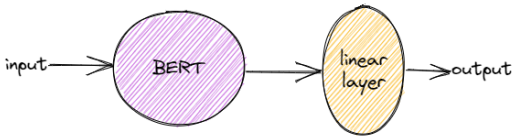

Figure 1: Architecture of the baseline model, using BERT

We used the "bert-base-uncased" model from HuggingFace Devlin et al. [2019] in order to represent every utterance in a vector space. The sentences with the same intent are more likely to be close to each other. Then, we added a linear layer to learn in order to predict the classes. We use the same method as Hinton et al.[2012] that randomly zeroes some of the elements of the input tensor because it has proven to be an effective technique for regularization and preventing the co-adaptation of neurons. It can explain that sometimes the train error is bigger than the test error. We use both the tokenizer and the encoder (self-attention) of BERT before "fine-tuning" the classifier. We also trained a binary model (the labels are "question" and "non-question") in order to see the performance of the encoder with a binary classifier.

4 Experimental settings

4.1 Dataset

Here, we use the SILICONE dataset (Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE), coming from Chapuis et al. [2020] which constitutes a good benchmark in order to measure the performance of the various methods that we are going to experiment afterwards. The dataset utterances are essentially in English and cover a variety of domains including daily life, scripted scenarios, joint task completion, phone call conversations and television dialogue. The labels classify the dialogue act utterances in 4 categories: commissive/responsive ("all right"), directive ("I suggest..."), inform ("It is important...") and question ("do you believe this ?"), each of them corresponds to the above-mentioned labels in the related works. Here, the purpose is to obtain a classifier able to recognize these DA categories in each utterance with the methods presented in the section.

4.2 Baseline model: feature extracting with BERT

4.2.1 Architecture : BERT

We use BERT (Bidirectional Encoder Representations from Transformers) which is a pre-trained language representation model developed by Devlin et al. [2019] from Google AI Language in 2018 for natural language processing (NLP) tasks. It is based on the transformer architecture, which is a type of neural network that is particularly effective for handling sequential data. This language model is trained on a large corpus of text data using an unsupervised learning approach. The model learns to predict missing words in a sentence based on the surrounding context, using a technique called masked language modeling (MLM). BERT also learns to identify the relationships between words in a sentence by processing the input text in both forward and backward directions, hence the term "bidirectional encoder". The pre-training process of BERT results in the model learning contextual representations of words, meaning that it understands the meaning of a word in a sentence based on the other words around it. This contextual understanding makes BERT highly effective for a variety of NLP tasks such as sentiment analysis, named entity recognition, question answering, and text classification (see figure A.1 in appendix)

4.2.2 Model : Attention model

We use an attention model which is a type of neural network architecture commonly used in natural language processing (NLP) tasks. It is a mechanism that allows the network to selectively focus on the most important parts of the input data when making predictions or generating output. In a traditional neural network, each input feature is given equal weight when making predictions. However, in NLP tasks, some parts of the input sequence may be more important than others for making accurate predictions. An attention model addresses this issue by allowing the network to learn to weight the importance of each input feature or sequence element dynamically, based on its relevance to the current prediction task. This is achieved by calculating attention weights for each input element, which reflect the degree of importance assigned to that element for the current task.

4.2.3 Cross entropy loss

We use for this model the cross-entropy loss which is a commonly used loss function in machine learning, particularly in classification tasks. It measures the difference between the predicted probability distribution and the true probability distribution of the target class. The loss is calculated by summing the logarithm of the predicted probability for the true class label, across all classes. The cross-entropy loss encourages the model to output high probabilities for the true class labels and low probabilities for the false class labels, and is therefore commonly used as a training objective for classification models. The model

learns to minimize the cross-entropy loss by adjusting its weights and biases during training using optimization algorithms such as gradient descent. The cross entropy loss is defined as

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

where y is a one-hot encoded vector representing the true class label and $\hat{y}$ is a vector of predicted probabilities for each class. The loss is calculated by summing the logarithm of the predicted probability for the true class label, across all classes.

5 Results

Because of the cost in terms of computation time, we have chosen to train the baseline model only on 10 epochs, which already provides convincing results. We observe in the figure 2 the evolution of the test loss on the SILICONE dataset.

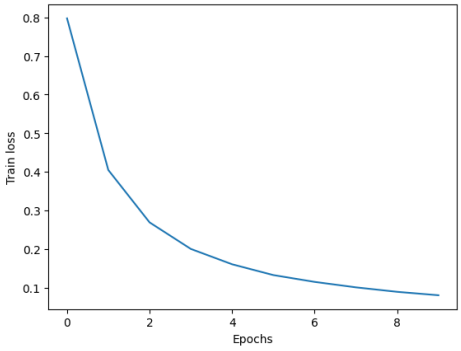

Figure 2: Test loss during the training of the baseline model

We have selected some random data in the test dataset to measure some metrics about the classifier (see table 1).

	precision	recall	f1-score	support
commissive	1.00	0.03	0.06	32
directive	0.56	0.38	0.45	50
inform	0.69	0.89	0.78	142
question	0.79	0.83	0.81	96
accuracy			0.71	320
macro avg	0.76	0.53	0.53	320
weighted avg	0.73	0.71	0.67	320

Table 1: Usual classification metrics on the labels of SILICONE

Some results may seem surprising here, but it is due to the fact that we had chosen only some data in the whole dataset (because of computation time). The global accuracy of the model, measure on the test dataset, is 74.8%, which is close to the value given above. Moreover, we can see that it is much easier for the classifier to detect question than the other categories, which seems plausible because of some literal indicators as "?".

We also have similar results for the binary model. We reach an accuracy of 89.8% (see the confusion matrix in appendixes, A.2).

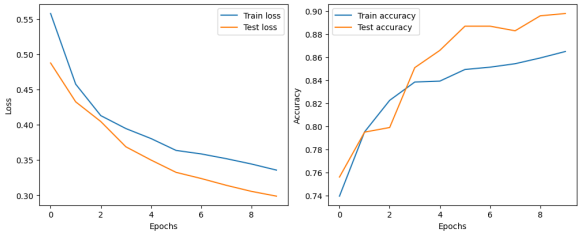

Figure 3: Evolution of the loss and accuracy for the binary classifier

	precision	recall	f1-score	support
non question	0.90	0.95	0.92	220
question	0.88	0.77	0.82	100
accuracy			0.89	320
macro avg	0.89	0.86	0.87	320
weighted avg	0.89	0.89	0.89	320

Table 2: Usual classification metrics on the labels of SILICONE (for the binary model)

6 Conclusion

The intent classification plays an important role in generation language or text identification in that it can guide the responses and improves the performance of language models (including generative language). Here, we have trained a supervised model with different labels drawn from the SILICONE database. In order to perform that classification, we have used an encoder with the BERT model from HuggingFace. Then we added an additional linear layer in order to adapt the transformers (fine-tuning) to our classifier. The results are robust and consistent in that, within 10 epochs,

the accuracy of the model reaches 74.8% on the SILICONE test dataset. This classification is even better with a binary model that distinguishes the question from other types of sentence. We have displayed several metrics that highlight these performances.

References

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.239. URL https://aclanthology.org/2020.findings-emnlp.239.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. 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 3734–3743, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1374. URL https://aclanthology.org/N19-1374.

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

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

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008. doi: 10.1007/s10579-008-9076-6.

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

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. The hcrc map task corpus: natural dialogue for speech recognition. 01 1993. doi: 10.3115/1075671.1075677.

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

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 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, April 30 - May 1 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-2319.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 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, 08 2013. doi: 10.1109/T-AFFC.2011.20.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. 2020. doi: 10.48550/ARXIV.2002.08801. URL https://arxiv.org/abs/2002.08801.

Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. From the token to the review: A hierarchical multimodal approach to opinion mining. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5539–5548, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1556. URL https://aclanthology.org/D19-1556.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. May 2019. doi: arXiv:1810.04805. URL https://doi.org/10.48550/arXiv.1810.04805.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie. Brew. Huggingface's transformers: State-of-the-art natural language processing. 2019. doi: arXiv:1910.03771. URL https://doi.org/10.48550/arXiv.1910.03771.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL http://arxiv.org/abs/1207.0580.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. Code-switched inspired losses

for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.656. URL https://aclanthology.org/2021.emnlp-main.656.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

Daniel Salber and Joëlle Coutaz. A wizard of oz platform for the study of multimodal systems. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 95–96, 1993.

Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. Quantum-inspired interactive networks for conversational sentiment analysis. 2019.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.641. URL https://aclanthology.org/2020.emnlp-main.641.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2cfa3753d6a524711acb5fce38eeca1a-Paper.pdf.

Appendices

A Extra figures

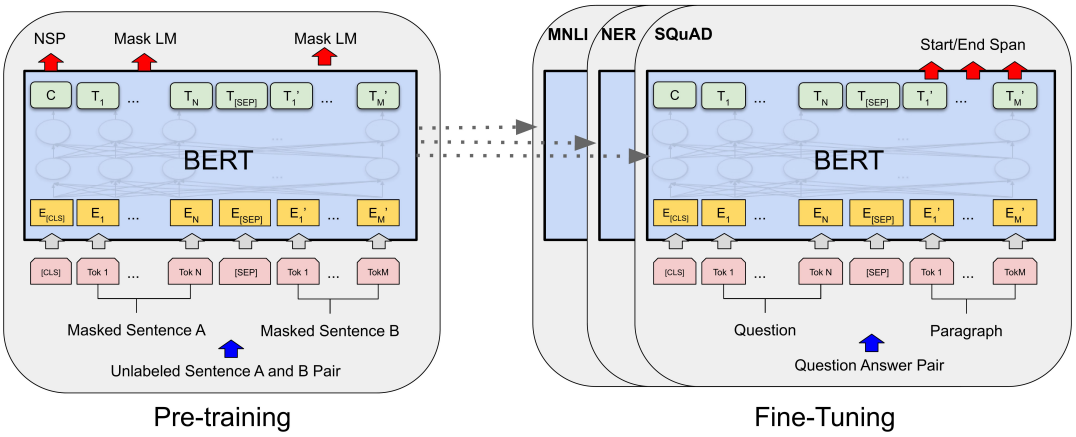

Figure A.1: Pre-training and fine-tuning procedures for BERT

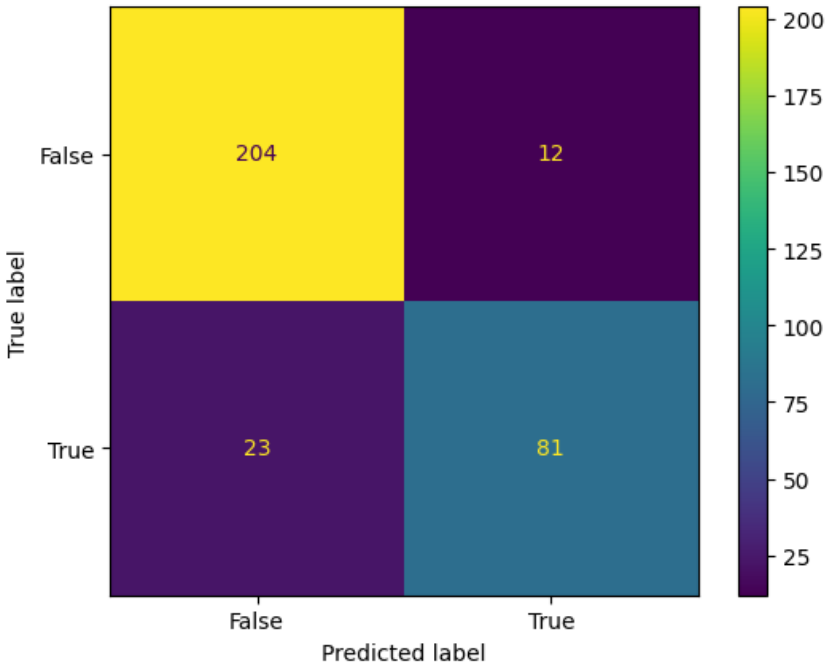

Figure A.2: Confusion matrix of the binary classifier