# Benchmarking Bert-based models performances for multi-lingual Dialog Acts (DA) classification

**TORRES Clément***
ENSAE
clement.torres@ensae.fr

**DAHER MANSOUR Michel***
ENSAE
michel.dahermansour@ensae.fr

## Abstract

Dialog Acts (DA) classification plays an important role in chatbots and spoken dialogue systems. Such tasks are specific in two ways: First, they need to be robust to language switching within or between conversations. Second, each utterance must be understood within the context of the current dialog. In this work, we build on the rapid emergence of Deep Learning techniques applied to Natural language Processing (NLP) and the availability of pre-trained models, and propose to benchmark a series of Transformer based models on both French and English spoken (transcribed) data, with different settings to take into account the context of the dialog. Our experiments show that a baseline XLM-RobERTA can provide fairly robust performances across these two languages even if slighlty lower than mono-lingual model like CamemBERT or BERT on their own language. Also, a simple concatenation up to the last three utterances does not perform well to capture the conversation context. Our code is available on Github [1]

## 1 Introduction

Intent classification is a critical component of chatbot interaction as it helps the chatbot understand the user's intent and respond appropriately (Chen, 2021; Jaiwai et al., 2021). Intent classification involves analyzing the user's input and identifying the underlying purpose or objective of the user's message. This is typically done using natural language processing (NLP) techniques such as machine learning algorithms and deep learning models.

There are several reasons why intent classification is crucial for the success of chatbot interaction. Firstly, by accurately identifying the user's intent, the chatbot can provide more relevant and personalized responses (Colombo* et al., 2019; Jalalzai* et al., 2020; Colombo et al., 2021b). This, in turn, can improve the user experience and increase user engagement with the chatbot.

Secondly, intent classification allows the chatbot to handle a wider range of user requests and queries. By understanding the user's intent, the chatbot can route the request to the appropriate response, whether that be a specific pre-defined response or a more complex algorithmic response.

Thirdly, intent classification can help the chatbot improve its performance over time. By analyzing user interactions and identifying patterns in user requests and queries, the chatbot can learn to better understand user intent and respond more effectively in the future.

In the last decade, Deep Learning models have achieved an impressive progress tackling Natural Language Processing (NLP) downstream tasks such as text classification analysis, text generation, question answering, etc. Multiple studies have been conducted which resulted in several models such as Convolution Neural Network (CNN) in 2014 and later on a Recurrent NN (RNN) model for sentence classification (Kim, 2014; Witon* et al., 2018) and DA acts classification (Lee and Dernoncourt, 2016) with state of art results at the time. A breakthrough in this domain was achieved in 2017 with the release of the Transformers based-architecture (Vaswani et al., 2017) to produce word embedding to solve all kind of NLP, audio or computer vision tasks.

Despite all these advances, Dialog Act (DA) classification remains a challenging task because it requires to model the information at the utterance level as well as dependencies between utterances within a dialog (Colombo et al., 2020). Additionally, these systems could be often confronted to multilingual speakers (within a

[1] https://github.com/Michel93DM/Intent_Classification_NLP2023

same conversation), or to several languages across dialogues. In such situations, (Colombo et al., 2021a) show that pretraining a transformer model on speech domain data with new code switched inspired losses can significantly improve DA classification accuracy.

One of the most widely used pre-trained model is the The Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) which, thanks to its bidirectional behavior and the huge corpus on which it was trained, has provided text representations that have been fine tuned to achieve state of the art results for several NLP tasks related to classification (Sun et al., 2019). Other variants were then published such as RoBERTA (Liu et al., 2019) which improved BERT performances by better designing the pretraining (changing the number of batches, length of sequences, prediction objectives, making patterns, etc.).

Nevertheless, these BERT-based models were trained on English language (ENG wikipedia and BookCorpus) and thus its high performances is limited to this language. Therefore, multi-lingual versions were implemented such as mBERT and trained on a bigger multilingual dataset. Despite the high performance achieved with mBERT but these models are often larger, and their results can lag behind their monolingual counterparts for high-resource languages like the tasty French CamemBERT (Louis Martin, 2019) when it was released in 2020.

Therefore, in this work we present: (i) a comparison of BERT like model performance when we modify the pre-processing of the data to better take the convsersation context into account. (ii) a comparison of Monolingual and Multilingual performances for DA classification on French and English datasets.

## 2 Experiments and protocol

### 2.1 Data

Several dataset are available for DA classification (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; Mckeown et al., 2013). However, we choose to work with the multilingual benchmark MIAM. The benchmark MIAM is used in our work and can be found on the HuggingFace hub. It is composed of 5 datasets, each one representing DA in a different language such as French, English, German, Spanish and Italian. In the context of this project, we will focus on the English and French languages. In Fig 1 and Fig 8 we illustrate the dataset 'Maptask' for the English language and 'Loria' for French.

| Utterance | Label | Speaker | Dialogue_Act |
|---|---|---|---|
| start at the extinct volcano | 5 | g | instruct |
| go down around the tribal settlement and then | 5 | g | instruct |
| whereabouts is the tribal settlement | 6 | f | query_w |
| it's at the bottom it's to the left of the e--... | 10 | g | reply_w |
| right | 0 | f | acknowledge |

Figure 1: illustration of the Maptask dataset

| Utterance | Label | Dialogue_Act | Speaker |
|---|---|---|---|
| Bravo! Vous avez été rapides! | 5 | greet | Samir |
| Qu'est-ce que je peux faire pour vous? | 1 | ask | Samir |
| merci | 17 | next_step | Julie |
| Eh bien, il va falloir la fabriquer cette mane... | 7 | inform | Samir |
| Mais sinon, vous avez encore des questions sur... | 1 | ask | Samir |

Figure 2: illustration of the Loria dataset

We can see that both datasets have the same structure, so that we can easily apply our preprocessing and modelling steps to each of them. However, they have some specificites besides their the language difference : Loria has 31 unique labels while Maptask has only 12 unique labels. Consequently, Loria has on average 273 utterances per label while Maptask has on average 2115 utterances per label. In Figure 3 we show in details the count of labels for both datasets which reveal a strongly unbalanced distribution of labels for loria while more smoothed distribution in Maptask. This is likely to make our training on Loria challenging to learn and to predict rare labels (high variance expected).

We follow the Train / Validation / Test split from the MIAM benchmark available on HuggingFace (Table 1).

| Dataset | Train | Validation | Test |
|---|---|---|---|
| Maptask | 25 382 | 5 221 | 5 335 |
| Loria | 8 465 | 942 | 1 047 |

Table 1: Size of datasets

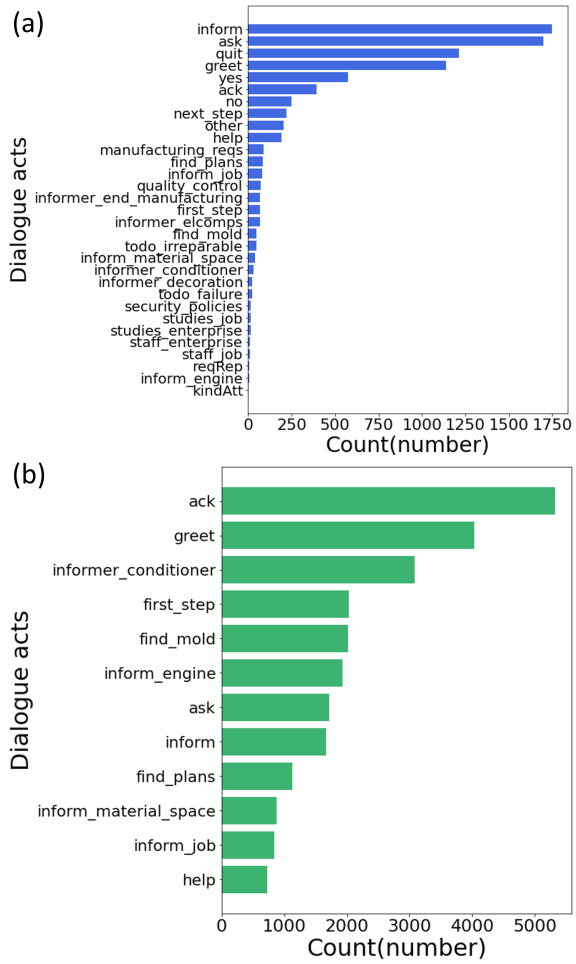


Figure 3: the count of labels in the datasets (a) loria and (b) Maptask

## 2.2 Our 3 experiments for DA classification

For all our experiments, we follow the same general approach: we use pretrained Transformer models and fine tune them on the data at hand to learn a downstream classification task.

Often, the pretraining is done with a mask language modelling objective (MLM), which enable to train models on large amounts of unlabeled data. However, this is a high resource consuming operation, necessitating several days of training on multiple GPUs (Chapuis et al., 2020). Instead, we can leverage on the rise of transfer learning methods and open source libraries developed by Huggingface (Wolf et al., 2020) to directly import "on the shelf" pretrained models that can perform well for zero-shot classification or with further fine tuning.

We simply add a "classification head" which will be in the form of a multi-layer perceptron which basically applies some dropout for robustness and then reduces the size of the embedding representation of the last layer (i.e. generally 768 for BERT based models) to our number of labels. To transform word embeddings in sentence embeddings, we use the default specification of *XLMRobertaClassificationHead* and *BertForSequenceClassification* from the HuggingFace transformer library which consists in taking the [CLS] token corresponding to the beginning of sentence.

The fine tuning is then done by training the model (the encoder layers + the classification head) in a supervised way with the DA labels of the training set for the Loria and Maptask dataset (separately).

Our training hyperparamters are described in the appendix.

### 2.2.1 Using the dialog context with data preprocessing

Each Dialog Act label $Y$ is associated to an utterance $U$. But each of these utterances is spoken in the context of a conversation $C_i = (u_1, u_2, \ldots, u_{|C_i|})$ which can enrich the word embeddings representation.

Several papers have highlighted the need to take the context into account for Dialog Act classification. For example, (Bothe et al., 2018) use a RNN with the context of the few previous utterances, and (Duran et al., 2021) investigate several aspects of text pre-processing like the number of words to keep in the vocabulary or input sequences length. In the same vein, we test two simple settings: one where each utterance is taken independently, and another where we attempt to get some contextual information by concatenating up to 3 consecutive utterances within each dialog.

We checked that the longest length (#of tokens) of concatenated utterances was always under the maximum length that BERT based model typically process (i.e. 512), see the appendix. We also add some dynamic padding so that the padding is done up to longest utterance for each batch taken independently.

For this section, we use XLM-Roberta (Conneau et al., 2020), the multi-lingual version of RoBERTA ("XLM" stands for "Cross-lingual Language Model"), which can be used both for Loria (French) and Maptask (English).

### 2.2.2 Comparing Monolingual and Multilingual performances

According to its designers (Conneau et al., 2020), XLM-RoBERTA is very competitive with strong monolingual models on GLUE and XNLI benchmarks. We want to test that too on DA classification tasks on the MIAM benchmark. We will therefore compare its performance with BERT on English DA classification, and with CamemBERT for French DA classification.

### 2.2.3 Comparing Encoder and Seq2seq model

We compare our previous decoder models with BART (Bidirectional and Auto-Regressive Transformers). It is a sequence-to-sequence (Seq2seq) model, which means that it takes a variable-length sequence as input and produce another variable-length sequence as output. We want to use its Auto-Regressive property to sequentially include the previous output label as an input for the next label classification.

Formally, we can say that we try to estimate the probability $P(Y_1, ...Y_{|C_i|})|u_1, ..., u_{|C_i|})$ where $|C_i|$ is the number of utterances in conversation $i$, which we can express as $\prod_{t=1}^{|C_i|} P(Y_t|Y_{<t}, u)$.

## 3 Results and discussion

In this section, we will present the results we obtained and discuss them, for more details on the models check the appendix.

### 3.1 Dialog context with concatenation

We do not find that simply concatenating up to 3 utterances within each conversation helps predict the dialog acts, on the contrary. The accuracy of XLM-RoBERTA models in this setting is 21% for Maptask and 17% for Loria (see Table 2 and Table 3). The model basically adopt a naive strategy of tend to predict mostly for the most frequent label(s) seen in the training set (see Fig 4 and Fig 5). It could be that such concatenation is too crude to capture well the context and express the specificity of the current utterance at the same time.

### 3.2 Monolingual and Multilingual performances

For baseline models without any concatenation, we find that XLM-RoBERTa models slightly under-perform compared to models specifically trained on one language (BERT for English and

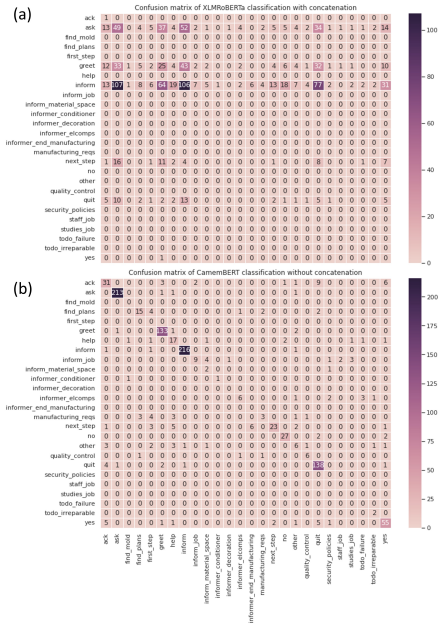


Figure 4: Confusion matrices for Loria (a) with concatenation and (b) without concatenation

| Model | Accuracy | F1 micro | F1 macro |
|---|---|---|---|
| BERT | 0.72 | 0.72 | 0.68 |
| XLM-Roberta baseline | 0.63 | 0.63 | 0.54 |
| XLM-Roberta concat3S | 0.21 | 0.21 | 0.029 |

Table 2: Accuracy, F1 micro and F1 macro scores for DA predictions on Maptask.

CamemBERT for French). It might still be advantageous to have multilingual model in production if we expect that interactions may happen with people from different countries and languages, with sometimes code switching within the same conversation.

| Model | Accuracy | F1 micro | F1 macro |
|---|---|---|---|
| CamemBERT | 0.86 | 0.86 | 0.47 |
| XLM-Roberta baseline | 0.74 | 0.74 | 0.18 |
| XLM-Roberta concat3S | 0.17 | 0.18 | 0.03 |

Table 3: Accuracy, F1 micro and F1 macro scores for DA predictions on Loria.

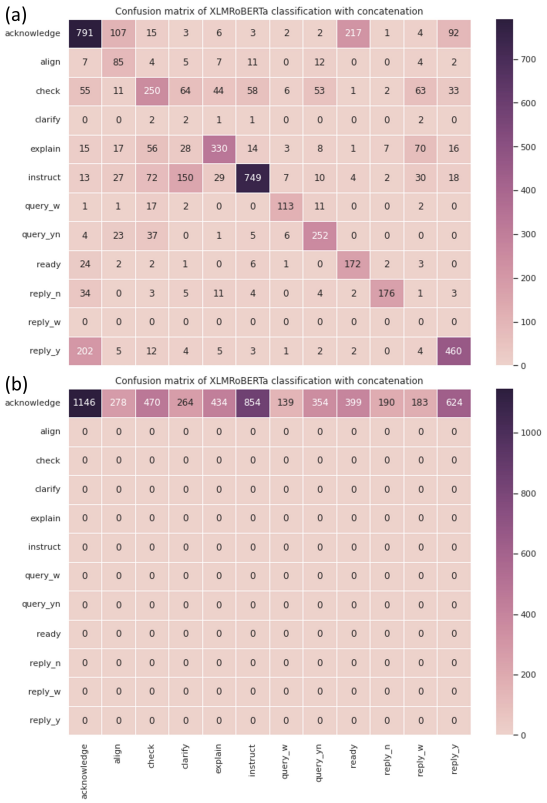


Figure 5: Confusion matrices for maptask (a) without concatenation and (b) with concatenation

## 4 Conclusion

In this work, we compared the performances of Bert-based models for dialogue act classification as a first step. In a second step, we concatenated three sequences for each dialogue and assign to it the label of the last one, the performance of this experiment was compared with the ones of the first step. For both languages, French and English, the concatenated-sequence experiment showed lower performances(lower f1 score). These results suggests that simple concatenation does not work well to capture the dependencies of the utterances within the same dialogue, and that other approaches like Seq2Seq models such as BART might obtain better results.

Generally, to increase the performances further studies could be pursued, such as domain adaptation (with OpenSubtitles dialogues, and MLM loss), taking into account a change of speaker, using previously predicted labels, adding weights for the loss function to take into account label imbalances, and fine tuning some specific layers as suggested by (Sun et al., 2019).

Another important research avenue is related to fairness. It is a critical aspect of intent classification systems (Pichler et al., 2022; Colombo et al., 2022), especially as these systems become more prevalent in our daily lives. Intent classification systems are used in various applications, such as chatbots, virtual assistants, and recommendation systems. These systems are designed to analyze user input and provide responses or recommendations that are relevant to the user's intent. However, if these systems are not designed with fairness in mind, they can perpetuate biases and reinforce discriminatory practices.

## 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.

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. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 30.

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

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez Yoann Dupont Laurent Romary Éric Villemonte de la Clergerie Djamé Seddah Benoît Sagot Louis Martin, Benjamin Muller. 2019. Camembert: a tasty french language model.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding Attention in Sequence-to-Sequence Models for Dialogue Act Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7594–7601.

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 Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mathawan Jaiwai, Kanokwatt Shiangjen, Surangkana Rawangyot, Sakpan Dangmanee, Thanakrit Kunsuree, and Autchadaporn Sa-nguanthong. 2021. Automatized educational chatbot using deep neural network. In *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, pages 5–8.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021a. 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. Association for Computational Linguistics.

jinha Park Hahyeon Kim Dimitrios Chrysostomou Chen, Li. 2021. How can i help you? an intelligent virtual assistant for industrial robots. *Association for Computing Machinery doi.org/10.1145/3434074.3447163*.

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

Nathan Duran, Steve Battle, and Jim Smith. 2021. Sentence encoding for dialogue act classification. *Natural Language Engineering*, pages 1–30.

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*.

.

# 5 Appendix

In this section, we will provide more details about the datasets and the model's preprocessing and implementation.

## 5.1 Datasets

In Figure. 6 for Loria and Figure. 7 for Maptask, we present the length of our concatenated utterances after tokenization. In these figures we show that even when we concatenate three sentences we are still below the capacity, 512, of the XLM-Roberta model.

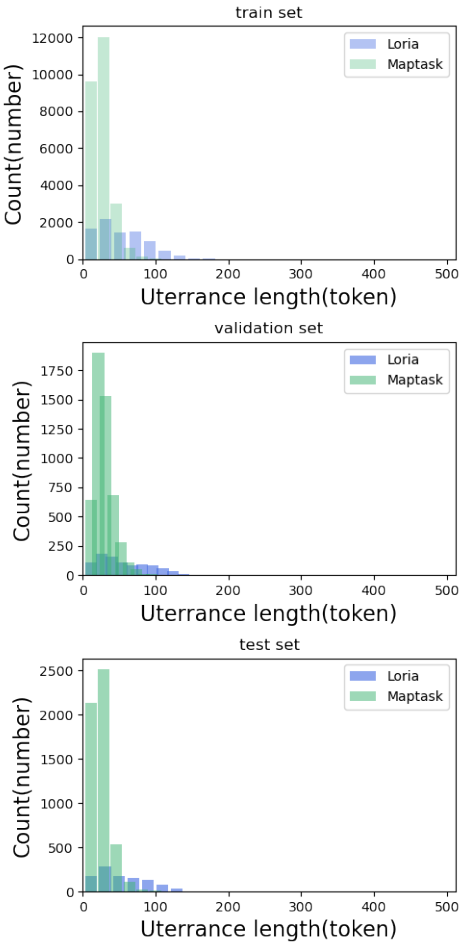


Figure 6: the Utterance length for the three concatenated sequence, blue for Loria and green for Maptask

## 5.2 Models

The parameters we used in our Model are presented in this Table 4.

### 5.2.1 Bert

We present here, the results of the classification with the Bert model.

| Hyperparamter name | value |
|---|---|
| learnign rate | 5e-5 |
| Epochs | 3 |
| Batch size | 8 or 16 |
| Weight decay | 0.01 |
| Loss | CrossEntropy |
| Optimizer | AdamW |

Table 4: Hyperparameters of the models we implemented.

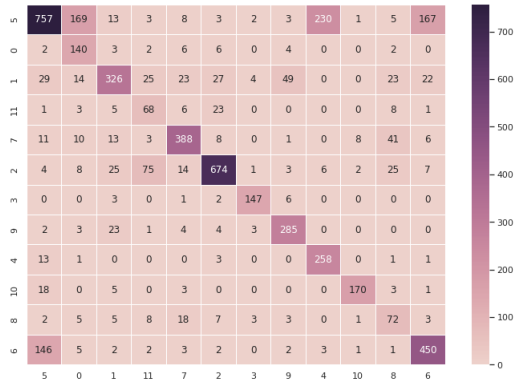

Figure 7: Confusion matrix for the Bert model on the Maptask test dataset

### 5.2.2 CamemBert

We present here, the results of the classification with the CamemBert model.

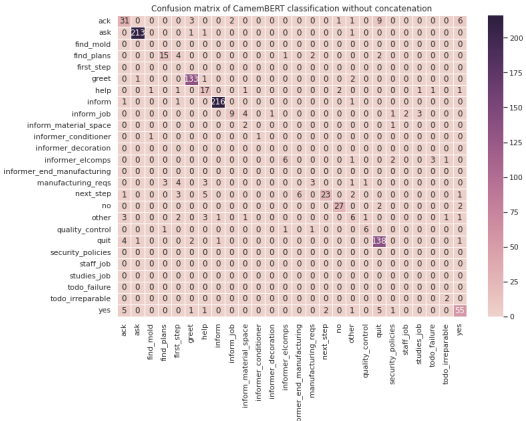


Figure 8: Confusion matrix for the CamemBert model on the Loria test dataset

For more information and details, please click here.