Do biliguanality and orality matter ? Comparative fine-tuning roBERTa and XLM-RoBERTa for dialog act classification

Arthur Katossky
ENSAE
arthur.katossky@ensae.fr

Loris Bulliard
ENSAE
loris.bulliard@ensae.fr

Abstract

The task of classifying dialog act intents is crucial in the development of intelligent dialog systems as it helps to guide the generation of responses and avoid generic, unnatural replies.

In this study, we study how the inclusion of (a) oral and (b) multilingual examples in a pre-training phase may improve downstream intent classification. We benchmarks model variants based on roBERTa and XLM-roBERTa and pre-trained on OpenSubtitles by masked language modeling. We then perform dialog act classification on Maptask (English) and Loria (French) datasets from the MIAM collection, and demonstrate the accuracy improvements achieved through this method.

1 Introduction

Creating an artificially intelligent personal assistant in the form of a chatbot requires a thorough understanding of the human's intent in conversation (Colombo, 2021). Both the emotion and dialog act of the human's utterance play critical roles in conditioning the generation of the chatbot's response. The dialog act refers to the type of speech act conveyed by the utterance, such as a question, statement, or request for action. Therefore, accurately identifying the dialog act is essential for generating an appropriate response. Additionally, the emotion (Garcia* et al., 2019; Jalalzai* et al., 2020) conveyed in the utterance plays a significant role in determining the tone and style of the response generated by the chatbot (Colombo* et al., 2019). Therefore, considering both emotion and dialog act is essential for constructing an effective and natural-sounding chatbot. Although both are important we will focus on dialog act in this paper.

1.1 Problem Framing

However, the task is far from trivial and has occupied researchers at least since the turn of the 21st century. Indeed, intent is highly dependent of context and language in general can be vague and ambiguous. Multiple intent-labeling schemes exist for dialog acts due to absence of consensus. Other important issues include the fact that dialogue context is crucial to understand individual conversation intent ; that oral language lacks records for model training ; and that many humans naturally shift language in the course of a discussion. This report focus on the latter two issues.

1.2 Spoken language

Large language models such as BERT, trained on written corpora, are not suitable for spoken dialogues due to distributional drift in terms of vocabulary, grammar and the presence of filler words (Dinkar et al., 2020). Indeed, models such as BERT are trained on a written texts while the input of our task is oral.

1.3 Code-switching and multilinguality

Another line of research is handling multilinguality inside an uterance. Indeed, most people are bilingual and switch between languages. An improvement for dialog agent would be to generalize to multilinguality as proposed by Chapuis et al. (2021).

1.4 Scope of the report

Our goal is to study whether knowledge of (a) orality and (b) multilinguality improves the intent classification. As we are limited by time and computation power, we are not able to test all possible variants to bullet-proof our preliminary results but we hope to give directions for future research.

2 Experiments Protocol

We fine-tune roBERTa (Liu et al. 2019 ; English-only) and XLM-roBERTa (Conneau et al. 2019 ; multilingual) on the task of intent classification

on two dataset from the Silicone collection, with and without pre-fine-tuning the models on OpenSubtitles for specialisation on oral speech. Complete code for reproducing our results is available at https://github.com/katossky/nlp-intent.

Contrary to Chapuis et al. 2020 and Dai et al. 2020, we do not model the contextual nature of dialog acts, and we model each utterance as independent texts separately. We do not either model so-called code-switching, as in Chapuis et al. 2021. This may be done in future research.

2.1 Datasets

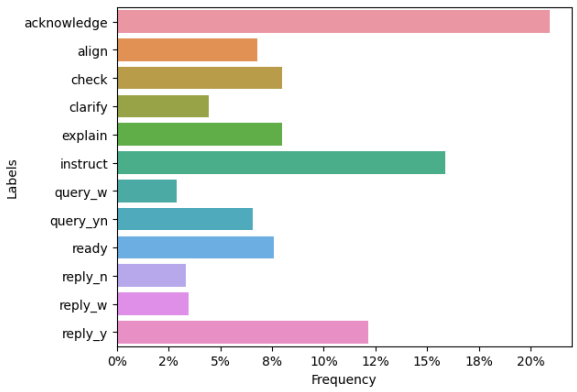
We use two textual sources. OpenSubtitles is set of movie dialogues aligned in different languages, that is adequate for teaching models orality and multilinguality. Miam is a collection of annotated monolingual dialogues that we use for studying how pre-training on OpenSubtitles affects downstream intent classification.

2.1.1 Miam

Miam is a collection of 5 datasets introduced by Chapuis et al. (2021) and available at https://huggingface.co/datasets/miam. All datasets contain dialogues, cut into fragments (utterances), respectively in German, Italian, French, Spanish and English, of which we use only English (so-called

Figure 2: Frequency of intent categories on Miam:Maptask (English)

Note the slight class imbalance. Predicting 'acknowledge' deterministically would achieve ca. 20% accuracy.



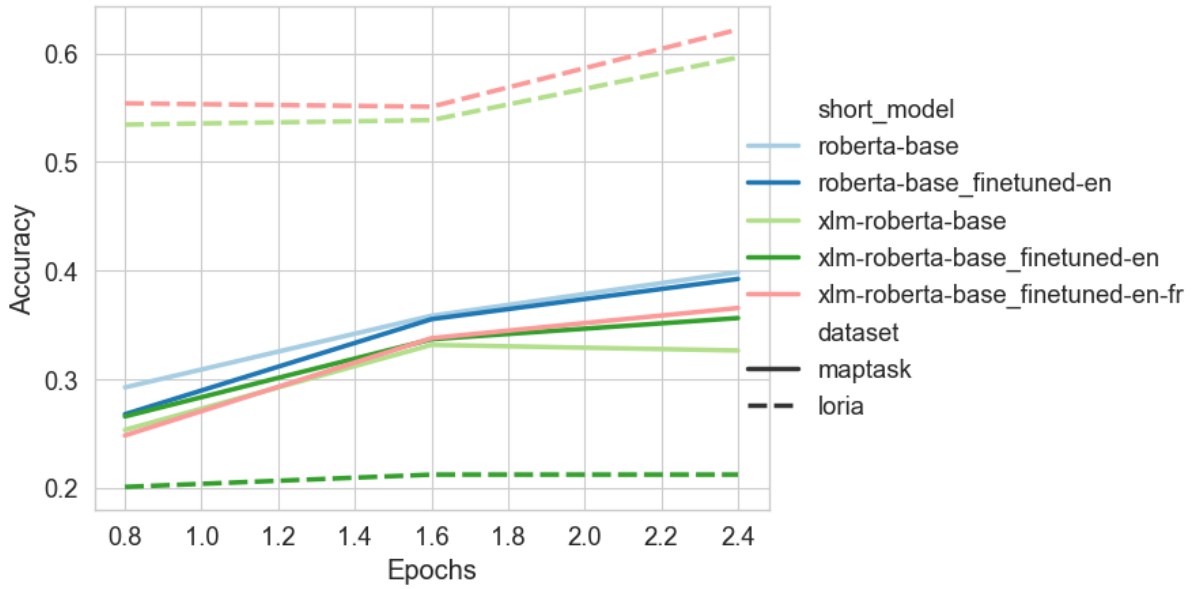
"Maptask" dataset) and French ("Loria"). Each utterance is labelled according to a dialogue-act scheme, which differ between the datasets. Because of class imbalance, it is possible to achieve 20% accuracy on both Maptask and Loria with a deterministic prediction (fig. 2 and 3).

2.1.2 OpenSubtitles

OpenSubtitles is a dataset extracted by Jörg Tiedemann in 2009 from http://www.opensubtitles.org/

Figure 1: Accuracy depends on models' knowledge of orality and multilinguality

We train 8 models and report their accuracy for intent classification on two intent-classification test sets, either in French (Loria) or English (Maptask). The base models are either roBERTa or XLM-roBERTa and they are either pre-fine-tuned or not on the OpenSubtitles dataset. This fine-tuning happens either in English only or in both English and French. All models are then post-fine-tuned on the intent-classification datasets of their corresponding language.



and regularly updated since then. The latest version (2016) is available at https://opus.nlpl.eu/OpenSubtitles-v2016.php. We use the version available through HuggingFace repository at https://huggingface.co/datasets/open_subtitles.

The dataset was initially intended for oral language translation, and thus Jörg Tiedemann and his collaborators have performed considerable work to detect dialog units and, inside dialogues, paired utterances across languages. That's what enables Chapuis et al. 2021 for instance to implement code-switching by alternating languages at each utterance inside the same dialog[1]. In total, OpenSubtitles contains 2.60G fragments (utterances), 17.09G tokens (words) in a total of 65 languages.

We rather orality of movie language as the main benefit of using this dataset. Among all available languages, we select English and French as they are both subsets of the languages available in XLM-roBERTa and the Miam collection. The dataset of paired French-English fragments consists in 41G utterances[2]. Due to time constraints, we only use a small share of 100 000 fragment instances.

[1]We do not take advantage of this alignment, as we do not implement code-switching, as we focus on orality and multilinguality in this report.

Table 1: The Miam dataset collection

Dataset	Language	# Classes	# Dialogs	# Fragments
Dihana	Spanish	11	723	19 063
Ilisten	Italian	15	36	1 986
Loria	French	31	1 019	8 465
Maptask	English	12	121	25 382
Ilisten	German	31	36	25 060

A class is a dialog act category. All counts are for the training split.

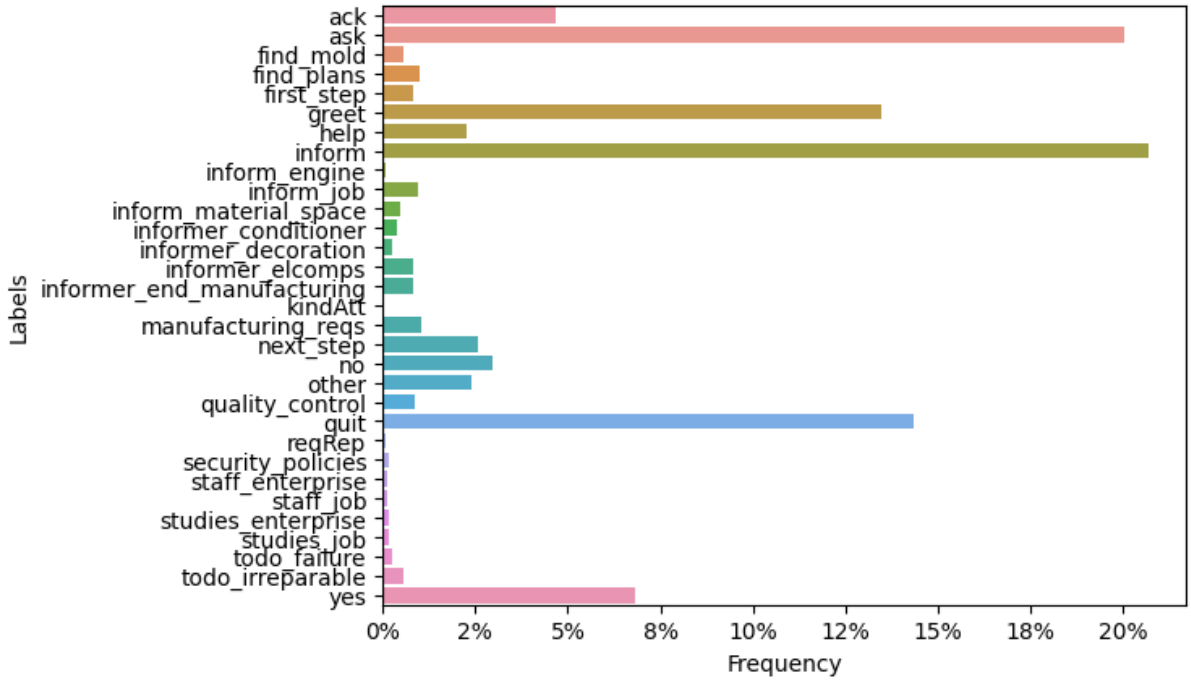
2.2 Baseline models

Due to the lack of a large enough dataset of spoken, multilingual language for training a model from scratch (and to a lesser extent due to a lack of computing resources), we use transfer learning from the BERT model family in order to adapt to spoken language. We thus succinctly present BERT and the two of its successors, roBERTa (monolingual) and XLM-roBERTa (multilingual), that we use in our experiments.

[2]These utterances are clearly grouped into dialogues, but his information is not available through the HuggingFace API. We could not find canonical train-test division either.

Figure 3: Frequency of intent categories on Miam:Loria (French)

Note the high class imbalance and the few outliers with unexpected class labels: staff_enterprise, inform_engine, etc. Predicting 'inform' deterministically would achieve ca. 20% accuracy.



2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. 2018) model is a pre-trained language model that uses transformer architecture. The model is trained using masked language, where the task is to predict a word that has been purposefully removed from a sequence. This task can be readily transposed to OpenSubtitles. However roBERTa improves over BERT using more input and better training strategies, wile XLM-roBERTa introduced multilinguality.

2.2.2 roBERTa and XLM-roBERTa

roBERTa (Liu et al., 2019) is a later iteration over BERT optimized with a better choice of hyperparameters and leads to better downstream task performance according to their authors. We use this as a reference for English language and continue the language-masking on OpenSubtitles. The model is available at https://huggingface.co/roberta-base.

XLM-roBERTa (Conneau et al., 2019) is yet an other iteration, still trained with language-masking, which has a cross-lingual representation. We pre-fine-tune the version available at https://huggingface.co/xlm-roberta-base by training on OpenSubtitles in both English and French.

2.3 Pretraining on OpenSubtitles for masked language modeling

We fine-tune both models on OpenSubtitles, in order to obtain representations closer to the spoken language. We train roBERTa on English, and create two versions of XLM-roBERTa, one monolingual English-trained, and one bilingual French-English trained.

Fine-tuning here consists in randomly masking 15 % of tokens in each sequence. Each utterance is tokenized and then padded to reach the size of 758 tokens needed for BERT-like model inputs. The task being to predict the masked words, it boils down to classification and thus cross-entropy loss is used during optimization. Because our computation power is limited, we explicitly freeze the 9 first modules of roBERTa but allow learning of the weights of 3 last modules. Among the 280 millions trainable parameters, we retrain solely 85 millions parameters. 100k examples of OpenSubtitles were given to each model.

At this end of this first stage, we obtain 5 models to be used in second stage for intent classification:

1. the two baseline models roBERTa and XLM-roBERTa
2. the same two, but finetuned on oral English
3. XLM-roBERTa, but finetuned on oral English *and* French

2.4 Fine-tuning on intent classification

For each of the five previous models, we further fine-tune the model on the downstream task of intent classification. We replace the token-masking head by a classification head consisting in a two-layer perceptron ($768 \to 768 \to \mathrm{k}$) with a 10% dropout, where 768 is the width of BERT-like models' last layer's output and k is the number of intent classes (see table 1).

More specificallly, we fine-tune the two English-only roBERTa-based models on Miam:Maptask and the three English-French XLM-roBERTa-based models on both Miam:Maptask and Miam:Loria, separately. We thus end up with a total of 8 models on which to base our discussion, as summarized in table 2.

Pre-fine-tuning XLM-roBERTa *in English* for an intent classification *on French fragments* is intentional, as we want to test how much pre-fine-tuning in one language degrades the performances in other languages. Given our time constraints, we do not pre-fine-tune XLM-roBERTa on OpenSubtitles in French.

Table 2: Model summary

Short name	Base model	Open Subt. language	Miam language
roberta–maptask	roBERTa	—	en
roberta–maptask–en	roBERTa	en	en
xlm-roberta–maptask	xlm-roBERTa	—	en
xlm-roberta–loria	xlm-roBERTa	—	fr
xlm-roberta–maptask–en	xlm-roBERTa	en	en
xlm-roberta–loria–en	xlm-roBERTa	en	fr
xlm-roberta–maptask–en-fr	xlm-roBERTa	en & fr	en
xlm-roberta–loria–en-fr	xlm-roBERTa	en & fr	fr

Since roBERTa and XLM-roBERTa share most of their architecture, there are for each of the 9 models a total of $(768 + 1) \times 768 + (768 + 1) \times k$ parameters to learn, for which we use the Adam W optimizer with respect to the standard cross-entropy loss. In order to judge overfitting, we also compute the cross-entropy loss on an evaluation sample (fig. 4). For comparing the models together, however, we focus directly on accuracy, where the predicted class is the one with highest probability (fig. 1).

In order to improve comparison between models trained on different datasets, we train all of

them on the same number of instances, no matter how many languages we use. Otherwise, a model could just be improved because it has seen more examples, not necessarily because it has been exposed to different sources[3]. Thus we align on the smallest dataset in the Miam collection, namely Ilisten for Italian, with 1 986 training and 971 test instances[4].

All 8 models actually learn from the fine-tuning. Fig. 4 shows none of the 8 models have reached a minimal loss on a test set after 3 epochs of training, suggesting that further computation effort may lead to better intent classification. Said differently, we find no sign of overfitting.

2.5 Hardware

For the first stage with OpenSubtitles, we use a cluster with ca. 10 Go (private) disk space, 36 (shared) 16-Go CPUs and one (shared) 16 Go GPU. Training takes an average of 0.13 s per instance.

For the second stage with Miam, we use a personal computer running under MacOS with 1 To disk space, 32 Go CPU-memory also accessible from the (programmable) GPU. Making Torch run on GPU on this computer required a significant effort of configuration and even after configuration, a few computation operations were not supported on Apple GPUs and were deferred to the CPU instead. Training on this machine takes an average of 0.46 s per instance.

[3]Doing so, we assume that each fragment contains a similar number of tokens. One may however doubt this assumption given that, per language, the number of fragments per dialogue varies a lot. A similar variability in the number of tokens per fragment may be present accross the datasets in the Miam collection (see table 1).

[4]Even though we do not use Italian in this report, the initial plan was to exploit every languages available in the Miam collection, hence the choice.

3 Results

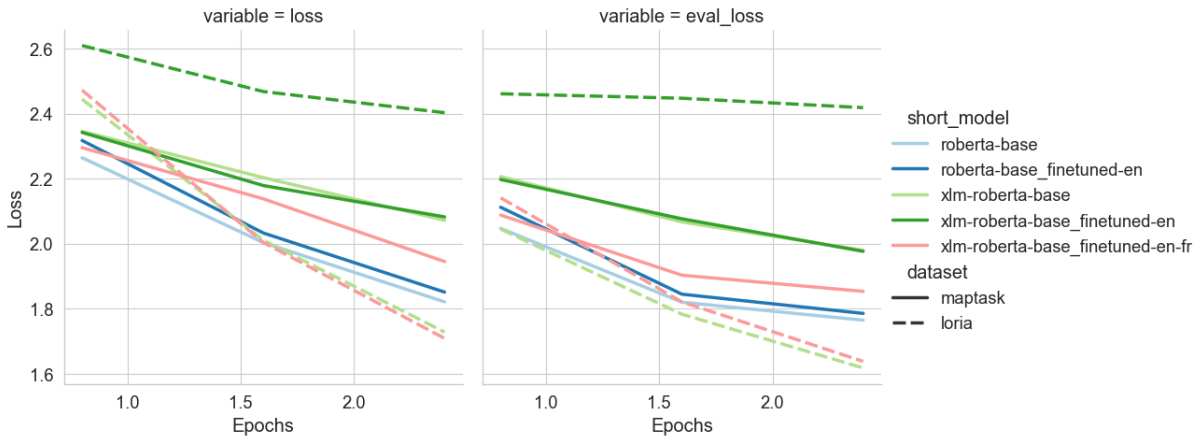
Figure 1 presents accuracy results for the 8 variants.

Orality. Finetuning roBERTa to OpenSubtitle virtually does not change predictions (blue lines), with an accuracy around 35% after 3 epochs. This is better than the deterministic approach consisting in predicting the most frequent class, which has accuracy of 20%. Finetuning XLM-roBERTa to OpenSubtitle in English appears to slightly increase accuracy by a couple of percentage points (solid green lines). However, this relative improvement is paid by a huge reduction in accuracy on the French dataset Loria (dashed lines), from 60% down to 20%, which is as bad as the deterministic prediction.

Multinguality. XLM-roBERTa has a slightly lower accuracy than roBERTa on Miam:Maptask, irrespective of whether the model is finetuned or not on oral language. In French, however, the accuracy is good, higher than 50% on the Loria dataset. Being exposed to French *and* English in the pre-fine-tuning actually improves the predictions on both the English and French downstream classifications by a couple of percentage points. This is surprising since, as mentioned in section 2.1.2, the total number of training instances is con-

Figure 4: Loss diminution during training

The loss of all models decreases during training. The absence of plateau on the evaluation dataset suggests that performance might have continue to increase with further training.



stant, which means that XLM-roBERTa fine-tuned on French and English has seen half the number of English instances, when compared with XLM-roBERTa fine-tuned on English only. We would thus normally expect prediction power in both languages to be slightly worse, not better.

Interplay. Fine-tuning XLM-roBERTa on oral English could have led to an increase in accuracy points when tested on French dialog acts. Indeed, orality is a feature of both languages, and thus learning orality in one could have led to improvement in the other. The fact that it is not the case points toward a specificity of oral speech in each language.

4 Discussion

Our research provides valuable insights into the effectiveness of training a language model on a small subset of OpenSubtitle content using various optimizers. However, there are several limitations to our study that should be taken into consideration.

Firstly, our research lacks external validity since we were unable to extend the protocol to other languages available in Miam due to time constraints. As a result, the generalizability of our findings to other languages may be limited.

Furthermore, we were not able to fully exploit the training process, as Figure 4 shows that there is still room for further learning. This is partly due to our limited computing resources, which only allowed us to train on a tiny fraction of the OpenSubtitle content (0.2%).

Additionally, our study highlights the importance of selecting the appropriate optimizer and its meta-parameters, such as learning rate and weight decay for AdamW. However, we were unable to explore alternative options due to time constraints. Future research could investigate the performance of other optimizers and their associated parameters to identify the most effective approach for training language models.

In conclusion, while our study provides valuable insights into training language models on a small subset of OpenSubtitle content using various optimizers, it is essential to consider the limitations of our research when interpreting the findings. Further research is needed to address these limitations and advance the development of effective language models.

Other extensions may include:

1. using multi-tasking learning in the second stage in order to fine-tune a multilingual model on multilingual intent classification (producing one single model), where we have restricted ourselves to separate downstream fine-tuning (producing several language-specific models)

2. tackling class imbalance

3. collecting additional intent classification tasks

4. working on better generation algorithms to include dialog act labels in the response (Colombo et al., 2022, 2021)

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

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*.

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. *CoRR*, abs/1907.11692.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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. *arXiv preprint arXiv:2009.11152*.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *arXiv preprint arXiv:2009.11340*.

Zhigang Dai, Jinhua Fu, Qile Zhu, Hengbin Cui, Yuan Qi, et al. 2020. Local contextual attention with hierarchical structure for dialogue act recognition. *arXiv preprint arXiv:2003.06044*.

Emile Chapuis, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2021. Code-switched inspired losses for generic spoken dialog representations. *CoRR*, abs/2108.12465.

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. 2021. A novel estimator of mutual information for learning to disentangle textual representations. *() ACL 2021*.

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