Code-switched inspired losses for generic spoken dialog representations

Emile Chapuis, Pierre Colombo, Matthieu Labeau, Chloé Clavel

Code-switched inspired losses for generic spoken dialog representations

Emile Chapuis[1]*, **Pierre Colombo**[1,2]*,
Matthieu Labeau[1], **Chloe Clavel**[1]
[1]LTCI, Telecom Paris, Institut Polytechnique de Paris,
[2]IBM GBS France
[1]firstname.lastname@telecom-paris.fr,
[2]pierre.colombo@ibm.com

Abstract

Spoken dialog systems need to be able to handle both multiple languages and multilinguality inside a conversation (*e.g* in case of code-switching). In this work, we introduce new pretraining losses tailored to learn multilingual spoken dialog representations. The goal of these losses is to expose the model to code-switched language. To scale up training, we automatically build a pretraining corpus composed of multilingual conversations in five different languages (French, Italian, English, German and Spanish) from `OpenSubtitles`, a huge multilingual corpus composed of 24.3G tokens. We test the generic representations on `MIAM`, a new benchmark composed of five dialog act corpora on the same aforementioned languages as well as on two novel multilingual downstream tasks (*i.e* multilingual mask utterance retrieval and multilingual inconsistency identification). Our experiments show that our new code switched-inspired losses achieve a better performance in both monolingual and multilingual settings.

1 Introduction

A crucial step in conversational AI is the identification of underlying information of the user's utterance (*e.g* communicative intent or dialog acts, and emotions). This requires modeling utterance-level information (Mitkov, 2014; Williams et al., 2014), to capture immediate nuances of the user utterance; and discourse-level features (Thornbury and Slade, 2006), to capture patterns over long ranges of the conversation. An added difficulty to this modeling problem is that most people in the world are bilingual (Grosjean and Li, 2013): therefore, progress on these systems is limited by their inability to process more than one language (English being the most frequent). For example, many people use English as a "workplace" language but seamlessly switch to their native language when the conditions are favorable (Heredia and Altarriba, 2001). Thus, there is a growing need for understanding dialogs in a multilingual fashion (Ipsic et al., 1999; Joshi et al., 2020; Ruder et al., 2019). Additionally, when speakers share more than one language, they inevitably will engage in code-switching (Sankoff and Poplack, 1981; Gumperz, 1982; Milroy et al., 1995; Auer, 2013; Parekh et al., 2020): switching between two different languages. Thus, spoken dialog systems need to be cross lingual (*i.e* able to handle different languages) but also need to model multilinguality inside a conversation (Ahn et al., 2020).

In this paper, we focus on building generic representations for dialog systems that satisfy the aforementioned requirements. Generic representations have led to strong improvements on numerous natural language understanding tasks, and can be fine-tuned when only small labelled datasets are available for the desired downstream task (Mikolov et al., 2013; Devlin et al., 2018; Lan et al., 2019; Liu et al., 2019; Yang et al., 2019). While there has been a growing interest in pretraining for dialog (Mehri et al., 2019; Zhang et al., 2019d), the focus has mainly been on English datasets. Thus, these works can not be directly applied to our multilingual setting. Additionally, available multilingual pretraining objectives (Lample and Conneau, 2019; Liu et al., 2020; Xue et al., 2020; Qi et al., 2021) face two main limitations when applied to dialog modeling: (1) they are a generalization of monolingual objectives that use flat input text, whereas hierarchy has been shown to be a powerful prior for dialog modeling. This is a reflection of a dialog itself, for example, context plays an essential role in the labeling of dialog acts. (2) The pretraining objectives are applied separately to each language considered, which does not expose the (possible) multilinguality inside a conversation (as it is the

*stands for equal contribution

case for code-switching) (Winata et al., 2021)[1].

Our main contributions are as follows:
1. *We introduce a set of code-switched inspired losses as well as a new method to automatically obtain several million of conversations with multilingual input context in different languages*. There has been limited work on proposing corpora with a sufficient amount of conversations that have multilingual input context. Most of this work focuses on social media, or on corpora of limited size. Hence, to test our new losses and scale up our pretraining, we automatically build a pretraining corpus of multilingual conversations, each of which comprises several languages, by leveraging the alignments available in OpenSubtitles (OPS).
2. *We showcase the relevance of the aforementioned losses and demonstrate that it leads to better performances on downstream tasks, that involve both monolingual conversations and multilingual input conversations.* For monolingual evaluation, we introduce the **M**ultilingual d**I**alog**A**ct bench**M**ark (MIAM): composed of five datasets in five different languages annotated with dialog acts. Following Mehri et al. (2019); Lowe et al. (2016), we complete this task with both contextual inconsistency detection and next utterance retrieval in these five languages. For multilingual evaluation, due to the lack of code-switching corpora for spoken dialog, we create two new tasks: contextual inconsistency detection and next utterance retrieval with multilingual input context. The datasets used for these tasks are unseen during training and automatically built from OPS.

In this work, we follow the recent trend (Lan et al., 2019; Jiao et al., 2019) in the NLP community that aims at using models of limited size that can both be pretrained with limited computational power and achieve good performance on multiple downstream tasks. The languages we choose to work on are English, Spanish, German, French and Italian.[2]. MIAM is available in Datasets (Wolf et al., 2020) https://huggingface.co/datasets/miam.

[1]We refer to code-switching at the utterance level, although it is more commonly studied at the word or span level (Poplack, 1980; Banerjee et al., 2018; Bawa et al., 2020; Fairchild and Van Hell, 2017)

[2]Although our pretraining can be easily generalised to 62 languages, we use a limited number of languages to avoid exposure to the so-called "curse of multilinguality" (Conneau et al., 2019)

2 Model and training objectives

Notations We start by introducing the notations. We have a set D of contexts (*i.e* truncated conversations), i.e., $D = (C_1, C_2, \ldots, C_{|D|})$. Each context C_i is composed of utterances u, i.e $C_i = (u_1^{L_1}, u_2^{L_2}, \ldots, u_{|C_i|}^{L_{|C_i|}})$ where L_i is the language of utterance u_i[3]. At the lowest level, each utterance u_i can be seen as a sequence of tokens, i.e $u_i^{L_i} = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$. For DA classification y_i is the unique dialog act tag associated to u_i. In our setting, we work with a shared vocabulary $\mathcal{V}$ thus $\omega_j^i \in \mathcal{V}$ and $\mathcal{V}$ is language independent.

2.1 Related work

Multilingual pretraining. Over the last few years, there has been a move towards pretraining objectives, allowing models to produce general multilingual representations that are useful for many tasks. However, they focus on the word level (Gouws et al., 2015; Mikolov et al., 2013; Faruqui and Dyer, 2014) or the utterance level (Devlin et al., 2018; Lample and Conneau, 2019; Eriguchi et al., 2018). Winata et al. (2021) shows that these models obtain poor performances in presence of code-switched data.

Pretraining to learn dialog representation. Current research efforts made towards learning dialog representation are mainly limited to the English language (Henderson et al., 2019; Mehri et al., 2019; Chapuis et al., 2020) and introduce objectives at the dialog level such as next-utterance retrieval, next-utterance generation, masked-utterance retrieval, inconsistency identification or generalisation of the cloze task (Taylor, 1953). To the best of our knowledge, this is the first work to pretrain representations for spoken dialog in a multilingual setting.

Hierarchical pretraining As we are interested in capturing information at different granularities, we follow the hierarchical approach of Chapuis et al. (2020) and decompose the pretraining objective in two terms: the first one for the utterance level and the second one to capture discourse level dependencies. Formally, the global hierarchical loss can be expressed as:

$$\mathcal{L}(\theta) = \underbrace{\lambda_u \times \mathcal{L}^u(\theta)}_{\text{utterance level}} + \underbrace{\lambda_d \times \mathcal{L}^d(\theta)}_{\text{dialog level}}. \quad (1)$$

These losses rely on a hierarchical encoder (Chen et al., 2018a; Li et al., 2018) composed of two

[3]In practice, we follow (Sankar et al., 2019a) and set the context length to 5 consecutive utterances.

functions f^u and f^d:

$$\mathcal{E}_{u_i^{L_i}} = f_\theta^u(\omega_1^i, \ldots, \omega_{|u_i|}^i), \quad (2)$$

$$\mathcal{E}_{C_j} = f_\theta^d(\mathcal{E}_{u_1^{L_1}}, \ldots, \mathcal{E}_{u_{|C_j|}^{L_{|C_j|}}}), \quad (3)$$

where $\mathcal{E}_{u_i^{L_i}} \in \mathbb{R}^{d_u}$ is the embedding of $u_i^{L_i}$ and $\mathcal{E}_{C_j} \in \mathbb{R}^{d_d}$ the embedding of C_j. The encoder is built on transformer layers.

2.2 Utterance level pretraining

To train the first level of hierarchy (*i.e* f_θ^u), we use a Masked Utterance Modelling (MUM) loss (Devlin et al., 2018). Let $u_i^{L_i}$ be an input utterance and $\tilde{u}_i^{L_i}$ its corrupted version, obtained after masking a proportion p_ω of tokens, the set of masked indices is denoted $\mathcal{M}_\omega$. The set of masked tokens is denoted Ω. The probability of the masked token given $\tilde{u}_i^{L_i}$ is given by:

$$p(\Omega|\tilde{u}_i^{L_i}) = \prod_{t \in \mathcal{M}_\omega} p_\theta(\omega_t^i|\tilde{u}_i^{L_i}). \quad (4)$$

2.3 Dialog level pretraining

The goal of the dialog level pretraining is to ensure that the model learns dialog level dependencies (through f_θ^d), *i.e* the ability to handle multi-lingual input context.

Generic framework Given C_k an input context, a proportion $p_\mathcal{C}$ of utterances is masked to obtained the corrupted version $\tilde{C}_k$. The set of masked utterances is denoted $\mathcal{U}$ and the set of corresponding masked indices $\mathcal{M}_u$. The probability of $\mathcal{U}$ given $\tilde{C}_k$ is:

$$p(\mathcal{U}|\tilde{C}_k) = \prod_{t \in \mathcal{M}_\mathcal{U}} \prod_{j=0}^{|u_t|-1} p_\theta(\omega_j^t|\omega_{1:j-1}^t, \tilde{C}_k). \quad (5)$$

As shown in Eq. 5, a masked sequence is predicted one word per step. As an example, at the j-th step, the prediction of ω_j^t is made given $(\omega_{1:j-1}^t, \tilde{C}_k)$ where $\omega_{1:j-1}^t = (\omega_1^t, \cdots, \omega_{1:j-1}^t)$. In the following, we describe different procedures to build $\mathcal{M}_\mathcal{U}$ and $\tilde{C}_k$ used in Eq. 5.

2.3.1 Masked utterance generation (MUG)

The MUG loss aims at predicting the masked utterance from a monolingual input context. As the vocabulary is shared, this loss will improve the alignment of conversations at the dialog level. This loss ensures that the model will be able to handle monolingual conversations in different languages.

Training Loss We rely on Eq. 5 for MUG. The input context is composed of utterances in the same language, *i.e* $\forall k, C_k = (u_1^{L_k}, \cdots, u_{|C_k|}^{L_k})$. The mask is randomly chosen among all the positions.

Example Given the monolingual input context given in Tab. 1, a random mask (*e.g* $[0, 3]$) is chosen among the positions $[0, 1, 2, 3, 4]$. The masked utterances are replaced by `[MASK]` tokens to obtain $\tilde{C}_k$ and a decoder attempts to generate them.

2.3.2 Translation masked utterance generation (TMUG)

The previous objectives are self-supervised and cannot be employed with parallel data when available. In addition, these losses do not expose the model to multilinguality inside the conversation. The TMUG loss addresses this limitation using a translation mechanism: the model learns to translate the masked utterance in a new language.

Training Loss We use Eq. 5 for TMUG with a bilingual input context C_k. C_k contains two different languages (*i.e* L and L') $\forall k, C_k = (u_1^{L_1}, \cdots, u_{|C_k|}^{L_k})$ with $L_i \in \{L, L'\}$. The masked positions $\mathcal{M}_u$ are all the utterances in language L'. Thus $\tilde{C}_k$ is a monolingual context.

Example Given the multilingual input context given in Tab. 1, the positions $[3, 4]$ are masked with sequences of `[MASK]` and the decoder will generate them in French. See ssec. 8.1 for more details on the generative pretraining.

2.3.3 Multilingual masked utterance generation (MMUG)

In the previous objectives, the model is exposed to monolingual input only. MMUG aims at relaxing this constraint by considering multilingual input context and generating the set of masked utterances in any possible target language.

Training Loss Given a multi-lingual input context $C_k = (u_1^{L_1}, \cdots, u_{|C_k|}^{L_{|C_k|}})$. A random set of indexes is chosen and the associated utterances are masked. The goal remains to generate the masked utterances.

Example In Tab. 1, the positions $[2, 3]$ are randomly selected from the available positions $[0, 1, 2, 3, 4]$. Given these masked utterances the model will generate 2 in Italian and 3 in Spanish. MMUG is closely related to code-switching as it exposes the model to multilingual context and the generation can be carried out in any language.

Index	Speaker	Monolingual Input	Multilingual Input
0	A	Good afternoon.	Good afternoon.
1	A	I'm here to see Assistant Director Harold Cooper.	Je suis ici pour voir l'assistant directeurr Harold Cooper.
2	B	Do you have an appointment?	Do you have an appointment?
3	A	I do not.	Non.
4	A	Tell him it's Raymond Reddington.	Dites lui que c'est Raymond Reddington.

Table 1: Example of automatically built input context from OPS.

2.4 Pretraining corpora

There is no large corpora freely available that contains a large number of transcripts of well segmented multilingual spoken conversation[4] with code switching phenomenon. Collecting our pretraining corpus involves two steps: the first step consists of segmenting the corpus into conversations, in the second step, we obtain aligned conversations.

Conversation segmentation Ideal pretraining corpora should contain multilingual spoken language with dialog structure. In our work, we focus on OPS (Lison and Tiedemann, 2016)[5] because it is the only free multilingual dialog corpus (62 different languages). After preprocessing, OPS contains around $50M$ of conversations and approximately 8 billion of words from the five different languages (*i.e* English, Spanish, German, French and Italian). Tab. 2 gathers statistics on the considered multilingual version of OPS. To obtain conversations from OPS, we consider that two consecutive utterances are part of the same conversation if the inter-pausal unit (Koiso et al., 1998) (*i.e* silence between them) is shorter than $\delta_T = 6s$. If a conversation is shorter than the context size T, they are dropped and utterance are trimmed to 50 (for justification see Fig. 1).

Obtaining aligned conversations We take advantage of the alignment files provided in OPS. They provide an alignment between utterances written in two different languages. It allows us to build aligned conversations with limited noise (solely high confidence alignments are kept). Statistics concerning the aligned conversations can be found in Tab. 3 and an example of automatically aligned context can be found in Tab. 1. The use of more advanced methods to obtain more fine-grained alignment (*e.g* word level alignment, span alignment inside an utterance) is left as future work.

[4]Specific phenomena appear (*e.g.* disfluencies (Dinkar et al., 2018), filler words (Dinkar et al., 2020)) when working with spoken language, as opposed to written text.

[5]http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php

	de	en	es	fr	it
# movies	46.5K	446.5K	234.4K	127.2K	134.7K
# conversations	1.8M	18.2M	10.0M	5.2M	4.2M
# tokens	363.6M	3.7G	1.9G	1.0G	994.7M

Table 2: Statistics of the processed version of OPS.

	de-en	de-es	de-fr	de-it	en-es
# utt.	23.4M	19.9M	17.1M	14.1M	63.5M
# tokens.	217.3M	194.1M	167.0M	139.5M	590.9M
	en-fr	en-it	es-fr	es-it	fr-it
# utt.	44.2M	36.7M	37.9M	31.4M	23.8M
# tokens.	413.7M	347.1M	362.1M	304.6M	248.5M

Table 3: Statistics of the processed version of the alignment files from OPS.

3 Evaluation framework

This section presents our evaluation protocol. It involves two different types of evaluation depending on the input context. The first group of experiences consists in multilingual evaluations with monolingual input context and follows classical downstream tasks (Finch and Choi, 2020; Dziri et al., 2019) including sequence labeling (Colombo et al., 2020), utterance retrieval (Mehri et al., 2019) or inconsistency detection. The second group focuses on multilingual evaluations with multilingual context.

3.1 Dialog representations evaluation

3.1.1 Monolingual context

Sequence labeling tasks. The ability to efficiently detect and model discourse structure is an important step toward modeling spontaneous conversations. A useful first level of analysis involves the identification of dialog act (DA) (Stolcke et al., 2000a) thus DA tagging is commonly used to evaluate dialog representations. However, due to the difficulty to gather language-specific labelled datasets, multilingual sequence labeling such as DA labeling remains overlooked.

Next-utterance retrieval (NUR) The utterance retrieval task (Duplessis et al., 2017; Saraclar and Sproat, 2004) focuses on evaluating the ability of an encoder to model contextual dependencies. Lowe

et al. (2016) suggests that NUR is a good indicator of how well context is modeled.

Inconsistency Identification (II) Inconsistency identification is the task of finding inconsistent utterances within a dialog context (Sankar et al., 2019b). The perturbation is as follow: one utterance is randomly replaced, the model is trained to find the inconsistent utterance.[6]

3.1.2 Multilingual context

To the best of our knowledge, we are the first to probe representation for multi-lingual spoken dialog with multilingual input context. As there is no labeled code-switching datasets for spoken dialog (research focuses on on synthetic data (Stymne et al., 2020), social media (Pratapa et al., 2018) or written text (Khanuja et al., 2020; Tan and Joty, 2021) rather than spoken dialog). Thus we introduce two new downstream tasks with automatically built datasets: Multilingual Next Utterance Retrieval (mNUR) and Multilingual Inconsistency Identification (mII). To best assess the quality of representations, for both mII and mNUR we choose to work with train/test/validation datasets of 5k conversations. The datasets, unseen during training, are built using the procedure described in ssec. 2.4.

Multilingual next utterance retrieval. mNUR consists of finding the most probable next utterance based on an input conversation. The evaluation dataset is built as follow: for each conversation in language L composed of T utterances, a proportion $p_{L'}$ of utterances is replaced by utterances in language L'. D utterances that we call distractors[7] in language L or L' from the same movie. For testing, we frame the task as a ranking problem and report the recall at N (R@N) (Schatzmann et al., 2005).

Multilingual inconsistency identification. The task of mII consists of identifying the index of the inconsistent sentences introduced in the conversation. Similarly to the previous task: for each conversation in language L composed of T utterances, a proportion $p_{L'}$ is replaced by utterances in language L', a random index is sampled from $[1, T]$ and the corresponding utterance is replaced by a negative utterance taken from the same movie.

[6] To ensure fair comparison, contrarily to Mehri et al. (2019) the pretraining is different from the evaluation tasks.

[7] D is set to 9 according to (Lowe et al., 2015)

3.2 Multilingual dialog act benchmark

DAs are semantic labels associated with each utterance in a conversational dialog that indicate the speaker's intention (examples are provided in Tab. 9). A plethora of freely available dialog act dataset (Godfrey et al., 1992; Shriberg et al., 2004; Li et al., 2017)) has been proposed to evaluate DA labeling systems in English. However, constituting a multilingual dialog act benchmark is challenging (Ribeiro et al., 2019b). We introduce **M**ultilingual d**I**alogue **A**ct bench**M**ark (in short MIAM). This benchmark gathers five free corpora that have been validated by the community, in five different European languages (*i.e.* English, German, Italian, French and Spanish). We believe that this new benchmark is challenging as it requires the model to perform well along different evaluation axis and validates the cross-lingual generalization capacity of the representations across different annotation schemes and different sizes of corpora.

DA for English For English, we choose to work on the MapTask corpus. It consists of conversations where the goal of the first speaker is to reproduce a route drawn only on the second speaker's map, with only vocal indications. We choose this corpus for its small size that will favor transfer learning approaches (27k utterances).

DA for Spanish Spanish research on DA recognition mainly focuses on three different datasets Dihana, CallHome Spanish (Post et al., 2013) and DIME (Coria and Pineda, 2005; Olguin and Cortés, 2006). Dihana is the only available corpora that contains free DA annotation (Ribeiro et al., 2019a). It is a spontaneous speech corpora (Benedı et al., 2006) composed of 900 dialogs from 225 users. Its acquisition was carried out using a Wizard of Oz setting (Fraser and Gilbert, 1991). For this dataset, we focus on the first level of labels which is dedicated to the task-independent DA.

DA for German For German, we rely on the VERBMOBIL (VM2) dataset (Kay et al., 1992). This dataset was collected in two phases: first, multiple dialogs were recorded in an appointment scheduling scenario, then each utterance was annotated with DA using 31 domain-dependent labels. The three most common labels (*i.e.* inform, suggest and feedback) are highly related to the planning nature of the data.

DA for French Freely available to academic and nonprofit research datasets are limited in the french language as most available datasets are privately

owned. We rely on the french dataset from the Loria Team (Barahona et al., 2012) (LORIA) where the collected data consists of approximately 1250 dialogs and 10454 utterances. The tagset is composed of 31 tags.

DA for Italian For Italian, we rely on the Ilisten corpora (Basile and Novielli, 2018). The corpus was collected in a Wizard of Oz setting and contains a total of 60 dialogs transcripts, 1,576 user dialog turns and 1,611 system turns. The tag set is composed of 15 tags.

Metrics: There is no consensus on the evaluation metric for DA labelling (e.g., Ghosal et al. (2019); Poria et al. (2018) use a weighted F-score while Zhang et al. (2019c) report accuracy). We follow Chapuis et al. (2020) and report accuracy.

3.3 Baseline encoders for downstream tasks

The encoders that will serve as baselines can be divided into two different categories: hierarchical encoders based on GRU layers ($\mathcal{HR}$) and pretrained encoders based on Transformer cells (Vaswani et al., 2017). The first group achieve SOTA results on several sequence labelling tasks (Lin et al., 2017; Li et al., 2018). The second group can be further divided in two groups: language specific ($BERT$) and multilingual BERT ($mBERT$) [8] and pretrained hierarchical transformers from (Zhang et al., 2019b) ($\mathcal{HT}$) are used as a common architecture to test the various pretraining losses.

Tokenizer We will work with both language specific and multlingual tokenizer. Model with multilingual tokenizer will be referred with a m (*e.g* $mBERT$ as opposed to $BERT$).

4 Numerical results

In this section, we empirically demonstrate the effectiveness of our code-switched inspired pretraining on downstream tasks involving both monolingual and multilingual input context.

4.1 Monolingual input context

4.1.1 DA labeling

Global analysis. Tab. 5 reports the results of the different models on MIAM. Tab. 5 is composed of two distinct groups of models: *language specifics models* (with language-specific tokenizers) and *multilingual models* (with a multilingual tokenizer denoted with a m before the model name). Overall, we observe that $mMUG$ augmented with both $TMUG$ and $MMUG$ gets a boost in performance (1.8% compared to $mMUG$ and 2.6% compared to a mBERT model with a similar number of parameters). This result shows that the model benefits from being exposed to aligned bilingual conversations and that our proposed losses (*i.e.* $TMUG$ and $MMUG$) are useful to help the model to better catch contextual information for DA labeling.

Language-specific v.s. multilingual models. By comparing the performances of $\mathcal{HR}$ (with either a CRF or MLP decoder), we can notice that for these models on DA labelling it is better to use a multilingual tokenizer. As multilingual tokenizers are not tailored for a specific language and have roughly twice as many tokens than their language-specific counterparts, one would expect that models trained from scratch using language-specific tokenizers would achieve better results. We believe this result is related to the spoken nature of MIAM and further investigations are left as future work. Recent work (Rust et al., 2020) has demonstrated that pretrained language models with language-specific tokenizers achieve better results than those using multilingual tokenizers. This result could explain the higher accuracy achieved by the language-specific versions of MUG compared to $mMUG$.

We additionally observe that some language-specific versions of BERT achieve lower results (*e.g* Dihana, Loria) than the multilingual version which could suggest that these pretrained BERT might be less carefully trained than the multilingual one; in the next part of the analysis we will only use multilingual tokenizers.

Overall, pretrained models achieve better results. Contrarily to what can be observed in some syntactic tagging tasks (Zhang and Bowman, 2018), for DA tagging pretrained models achieve consistently better results on the full benchmark. This result of multilingual models confirms what is observed with monolingual data (see Mehri et al. (2019)): pretraining is an efficient method to build accurate dialog sequence labellers.

Comparison of pretraining losses In Tab. 5 we dissect the relative improvement brought by the different parts of the code-switched inspired losses and the architecture to better understand the relative importance of each component. Similarly to Chapuis et al. (2020), we see that the hierarchical pretraining on spoken data (see $mMUG$)

[8] Details of language specific BERT and on baseline models can be found in ssec. 8.1 and in ssec. 8.3 respectively.

improves over the $mBERT$ model. Interestingly, we observe that the monolingual pretraining works slightly better compared to the multilingual pretraining when training using the same loss. This result surprising results might be attributed to the limited size of our models (Karthikeyan et al., 2019). We see that in both cases, introducing a loss with aligned multilingual conversations ($MMUG$ or $TMUG$) induces a performance gain (+1.5%). This suggests that our pretraining with the new losses better captures the data distribution. By comparing the results of $mMUG + TMUG$ with $mMUG$, we observe that the addition of cross-lingual generation during pretraining helps. A marginal gain is induced when using $MMUG$ over $TMUG$, thus we believe that the improvement of $mMUG + MMUG$ over $mMUG$ can mainly be attributed to the cross-lingual generation part. Interestingly, we observe that the combination of all losses out-performs the other models which suggests that different losses model different patterns present in the data.

4.1.2 Inconsistency Identification

In this section, we follow Mehri et al. (2019) and evaluate our pretrained representations on `II` with a monolingual context. A random guess identifies the inconsistency by randomly selecting an index in $[1, T]$ which corresponds to an accuracy of 20% (as we have set $T = 5$). Tab. 4 gathers the results. Similarly conclusion than in sssec. 4.1.3 can be drawn: pretrained models achieve better results and the best performing model is obtained with $mMUG$+$MMUG$+$TMUG$.

4.1.3 Next utterance retrieval

In this section, we evaluate our representations on `NUR` using a monolingual input context. As we use 9 distractors, a random classifier would achieve 0.10 for R@1, 0.20 for R@2 and 0.50 for R@5. The results are presented in Tab. 7. When comparing the accuracy obtained by the baselines models (*e.g* mBERT, mBERT (4-layers) and $\mathcal{HR}$) and our model using the contextual losses at the context level for pretraining (*i.e* MUG, $TMUG$ and $MMUG$) we observe a consistent improvement.

Takeaways Across all the three considered tasks, we observe that the models pretrained with our losses achieve better performances. We believe it is indicative of the validity of our pretraining.

4.2 Multilingual input context

In this section, we present the results on the downstream tasks with multilingual input context.

4.2.1 Multilingual inconsistency identification

Tab. 6 gathers the results for the `mII` with bilingual input context. As previously a random baseline would achieve an accuracy of 20%. As expected predicting inconsistency with bilingual context is more challenging than with a monolingual context: we observe a drop in performance of around 15% for all methods including the multilingual BERT. Our results confirm the observation of Winata et al. (2021): multilingual pretraining does not guarantee good performance in code switched data. However, we observe that the losses, by exposing the model with bilingual context, obtain a large boost (absolute improvement of 6% which correspond to a relative boost of more than 20%). We also observe that MUG+$MMUG$+$TMUG$ outperforms $mBERT$ on all pairs, with fewer parameters.

4.2.2 Multilingual next utterance retrieval

The results on bilingual context for `mNUR` are presented in Tab. 8. `mNUR` is more challenging than `NUR`. Overall, we observe a strong gain in performance when exposing the model to bilingual context (gain over 9% absolute point in R@5). **Takeaways**: These results show that our code-switched inspired losses help to learn better representations in a particularly effective way in the case of multilingual input context.

	de	en	es	fr	it	Avg
$mBERT$	44.6	42.9	43.7	43.5	42.3	43.4
$mBERT$ (4-layers)	44.6	42.1	43.7	42.5	41.4	42.9
$m\mathcal{HR}$	44.1	42.0	40.4	41.3	41.2	41.8
$mMUG$	45.2	43.5	45.1	43.1	42.7	43.9
$mMUG + TMUG$	48.2	42.6	**47.7**	44.6	44.3	45.5
$mMUG + MMUG$	**49.6**	**43.8**	46.1	**46.2**	43.3	45.8
$mMUG + TMUG + MMUG$	49.1	43.4	46.2	45.9	**45.1**	**46.0**

Table 4: Results on the `II` task with monolingual input context. On this task the accuracy is reported.

5 Conclusions

In this work, we demonstrate that the new code-switched inspired losses help to learn representations for both monolingual and multilingual dialogs. This work is the first that explicitly includes code switching during pretraining to learn multilingual spoken dialog representations. In the future, we plan to further work on `OPS` to obtain fine-grained alignments (*e.g* at the span and word levels) and enrich the definition of code-switching

	Toke.	VM2	Map Task	Dihana	Loria	Ilisten	Total
BERT	lang	54.7	66.4	86.0	50.2	74.9	66.4
BERT - $4layers$	lang	52.8	66.2	85.8	55.2	76.2	67.2
$\mathcal{HR}$ + CRF	lang	49.7	63.1	85.8	73.4	75.2	69.4
$\mathcal{HR}$ + MLP	lang	51.3	63.0	85.6	58.9	75.0	66.8
MUG (Chapuis et al., 2020)	lang	54.0	66.4	99.0	79.0	74.8	74.6
mBERT	multi	53.2	66.4	98.7	76.2	74.9	73.8
mBERT - $4layers$	multi	52.7	66.2	98.0	75.1	75.0	73.4
$m\mathcal{HR}$ + CRF	multi	49.8	65.2	97.6	75.2	76.0	72.8
$m\mathcal{HR}$ + MLP	multi	51.0	65.7	97.8	75.2	76.0	73.1
$mMUG$	multi	53.0	67.3	98.3	78.5	74.0	74.2
$mMUG + TMUG$	multi	54.8	**67.4**	99.1	**80.8**	74.9	75.4
$mMUG + MMUG$	multi	**56.2**	**67.4**	99.0	78.9	**77.6**	75.8
$mMUG + TMUG + MMUG$	multi	**56.2**	66.7	**99.3**	80.7	77.0	**76.0**

Table 5: Accuracy of pretrained and baseline encoders on MIAM. Models are divided in three groups: hierarchical transformer encoders pretrained using our custom losses, baselines (see ssec. 8.3) using either multilingual or language specific tokenizer. *Toke.* stands for the type of tokenizer: *multi* and *lang* denotes a pretrained tokenizer on multilingual and language specific data respectively. When using $lang$ tokenizer, MUG pretraining and finetuning are performed on the same language.

	de-en	de-es	de-fr	de-it	en-es	en-fr	en-it	es-fr	es-it	fr-it	Avg
mBERT	31.2	28.0	28.0	27.6	28.4	33.0	32.1	35.1	31.0	28.7	30.3
mBERT (4-layers)	30.7	28.7	28.2	27.1	28.7	33.1	30.9	35.1	30.1	28.1	30.1
$m\mathcal{HR}$	28.7	27.9	26.9	27.3	25.5	25.1	30.6	34.3	30.0	26.8	28.3
$mMUG$	34.5	30.1	30.1	27.7	28.2	33.1	32.1	35.4	32.0	29.5	31.2
$mMUG + TMUG$	34.0	32.0	32.2	29.1	28.3	32.9	32.4	35.1	33.0	29.3	31.8
$mMUG + MMUG$	35.1	33.8	**34.0**	30.1	29.4	32.8	32.6	36.1	33.9	31.6	32.9
$mMUG + TMUG + MMUG$	**35.7**	**34.0**	32.5	**31.4**	**30.1**	**33.6**	**33.9**	**36.2**	**34.0**	**32.1**	**33.4**

Table 6: Results on the mII task with bilingual input context.

	de			en			es			fr			it		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	65.1	27.1	20.1	62.1	26.1	16.8	62.4	24.8	15.3	63.9	22.9	13.4	66.1	27.8	16.9
mBERT (4-layers)	65.1	27.5	20.2	61.4	25.6	15.1	62.3	24.6	15.9	63.4	22.8	12.9	65.6	27.4	15.8
$m\mathcal{HR}$	65.0	27.1	20.0	60.3	25.0	15.2	61.0	23.9	14.7	63.0	22.9	13.0	65.4	27.3	15.8
$mMUG$	66.9	28.0	20.0	65.9	26.4	16.3	66.7	26.4	16.4	66.2	25.2	17.2	68.9	28.9	17.2
$mMUG + TMUG$	67.2	**28.2**	20.1	68.3	**29.8**	17.5	69.0	26.9	17.3	67.1	**25.4**	17.3	69.9	29.4	18.6
$mMUG + MMUG$	66.9	28.1	20.7	68.1	26.7	18.0	68.7	26.9	17.5	67.2	25.2	**17.4**	69.7	29.4	18.6
$mMUG + TMUG + MMUG$	**68.3**	27.4	**21.2**	**68.9**	27.8	**18.3**	**69.3**	**27.1**	**17.9**	**67.4**	25.3	**17.4**	**70.2**	**30.0**	**18.7**

Table 7: Results on the NUR task with monolingual input context. R@N stands for recall at N.

	de-en			de-es			de-fr			de-it			en-es		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	54.4	27.0	11.6	55.9	24.8	11.9	57.9	24.2	12.9	57.5	23.9	13.0	55.4	25.6	13.0
mBERT (4-layers)	54.1	26.5	11.9	55.7	24.8	**12.4**	57.2	24.1	12.4	57.0	23.5	13.1	55.6	23.1	12.9
$m\mathcal{HR}$	52.1	25.5	12.1	54.9	14.6	10.7	56.1	22.9	11.3	56.9	24.9	13.0	53.9	23.7	12.8
$mMUG$	59.7	25.2	11.5	61.2	26.2	11.6	60.7	25.3	13.8	61.6	**26.4**	11.9	62.1	23.9	13.10
$mMUG + TMUG$	59.8	26.2	12.1	62.7	29.0	10.7	61.9	27.3	13.9	63.2	26.3	12.6	63.1	28.4	14.0
$mMUG + MMUG$	59.8	27.2	12.1	62.7	28.1	11.6	60.7	24.8	**14.4**	62.7	26.1	**13.8**	63.4	28.2	14.7
$mMUG + TMUG + MMUG$	**61.0**	**28.2**	**13.1**	**63.2**	**29.1**	11.7	**62.1**	**28.7**	14.1	**63.4**	26.3	12.9	**64.3**	**29.4**	**15.2**
	en-fr			en-it			es-fr			es-it			fr-it		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	57.9	25.4	12.3	57.1	23.5	12.1	57.8	27.9	12.2	54.2	22.1	11.2	58.1	22.9	12.5
mBERT (4-layers)	57.8	23.2	12.1	57.1	23.4	11.9	57.1	27.6	12.1	55.1	22.0	11.1	58.9	22.6	12.7
$m\mathcal{HR}$	55.9	20.9	11.6	56.8	22.9	11.8	54.9	27.0	12.0	53.9	21.0	11.6	56.1	21.9	11.4
$mMUG$	61.9	24.9	12.9	61.4	27.6	11.9	64.6	29.7	13.9	59.0	24.2	13.4	59.7	23.6	12.2
$mMUG + TMUG$	62.9	25.2	14.3	62.7	27.8	12.9	64.9	29.9	13.8	**60.1**	25.1	13.5	61.5	25.8	13.1
$mMUG + MMUG$	63.9	26.3	**14.7**	61.5	27.6	13.1	65.0	30.2	13.1	**60.1**	25.3	12.9	**63.1**	**25.9**	13.6
$mMUG + TMUG + MMUG$	**64.0**	**26.7**	14.1	**63.5**	**28.7**	**13.7**	**66.1**	**31.4**	**14.5**	**60.1**	**25.5**	**13.6**	**63.1**	**25.9**	**14.2**

Table 8: Results on the mNUR task with bilingual input context.

(currently limited at the utterance level). Lastly, when considering interactions with voice assistants and chatbots, users may not be able to express their intent in the language in which the voice assistant is programmed. Thus, we would like to strengthen our evaluation protocol by gathering a new DA benchmark with code-switched dialog to improve the multilingual evaluation. A possible future research direction includes focusing on emotion classification instead of dialog acts (Witon et al., 2018; Jalalzai et al., 2020), extend our pre-training to multimodal data (Garcia et al., 2019; Colombo et al., 2021a) and use our model to obtain better results in sequence generation tasks (*e.g* style transfer (Colombo et al., 2021b, 2019), automatic evaluation of natural language generation (Colombo et al., 2021c)).

6 Acknowledgments

The research carried out in this paper has received funding from IBM, the French National Research Agency's grant ANR-17-MAOI and the DSAIDIS chair at Telecom-Paris. This work was also granted access to the HPC resources of IDRIS under the allocation 2021-AP010611665 as well as under the project 2021-101838 made by GENCI.

References

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What code-switching strategies are effective in dialog systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 213–222.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. *arXiv preprint arXiv:1806.05997*.

Lina Maria Rojas Barahona, Alejandra Lorenzo, and Claire Gardent. 2012. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents.

Pierpaolo Basile and Nicole Novielli. 2018. Overview of the evalita 2018 italian speech act labe ling (iliste n) task. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:44.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

José-Miguel Benedı, Eduardo Lleida, Amparo Varona, Marıa-José Castro, Isabel Galiano, Raquel Justo, I López, and Antonio Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. *CoRR*, abs/1805.06280.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

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.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018a. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018b. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234. ACM.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922*.

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. *arXiv preprint arXiv:2002.08801*.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *arXiv preprint arXiv:2105.02685*.

Pierre Colombo, Guillaume Staerman, Chloe Clavel, and Pablo Piantanida. 2021c. Automatic text evaluation through the lens of wasserstein barycenters. *arXiv preprint arXiv:2108.12463*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

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. *arXiv preprint arXiv:1911.02116*.

S Coria and L Pineda. 2005. Predicting obligation dialogue acts from prosodic and speaker infomation. *Research on Computing Science (ISSN 1665-9899), Centro de Investigacion en Computacion, Instituto Politecnico Nacional, Mexico City*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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*.

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*.

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, and Chloé Clavel. 2018. Disfluencies and teaching strategies in social interactions between a pedagogical agent and a student: Background and challenges. In *SEMDIAL 2018 (AixDial), The 22nd workshop on the Semantics and Pragmatics of Dialogue*, pages 188–191. Laurent Prévot, Magalie Ochs and Benoît Favre.

Guillaume Dubuisson Duplessis, Franck Charras, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. 2017. Utterance retrieval based on recurrent surface text patterns. In *European Conference on Information Retrieval*, pages 199–211. Springer.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Carlos Escolano, Marta R Costa-jussà, José AR Fonollosa, and Mikel Artetxe. 2020. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *arXiv preprint arXiv:2006.01594*.

Sarah Fairchild and Janet G Van Hell. 2017. Determiner-noun code-switching in spanish heritage speakers. *Bilingualism: Language and Cognition*, 20(1):150–161.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Sarah E Finch and Jinho D Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. *arXiv preprint arXiv:2006.06110*.

Norman M Fraser and G Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99.

Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. 2019. 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. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

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.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments.

François Grosjean and Ping Li. 2013. *The psycholinguistics of bilingualism*. John Wiley & Sons.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ivo Ipsic, Nikola Pavesic, France Mihelic, and Elmar Noth. 1999. Multilingual spoken dialog system. In *ISIE'99. Proceedings of the IEEE International Symposium on Industrial Electronics (Cat. No. 99TH8465)*, volume 1, pages 183–187. IEEE.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *arXiv preprint arXiv:2003.11593*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Martin Kay, Peter Norvig, and Mark Gawron. 1992. *Verbmobil: A translation system for face-to-face dialog*. University of Chicago Press.

Simon Keizer, Rieks op den Akker, and Anton Nijholt. 2002. Dialogue act recognition with bayesian networks for dutch dialogues. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *COLING*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2018. A dual-attention hierarchical recurrent neural network for dialogue act classification. *CoRR*, abs/1810.09154.

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

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909.

Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. *arXiv preprint arXiv:1605.05414*.

Shikib Mehri, Evgeniia Razumovsakaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

James Milroy et al. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Sergio Rafael Coria Olguin and Luis Albreto Pineda Cortés. 2006. Predicting dialogue acts from prosodic information. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 355–365. Springer.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

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. *arXiv preprint arXiv:1810.02508*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3067–3072.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019a. Hierarchical multi-label dialog act recognition on spanish data. *arXiv preprint arXiv:1907.12316*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019b. A multilingual and multidomain study on dialog act recognition using character-level tokenization. *Information*, 10(3):94.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019a. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019b. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45.

Murat Saraclar and Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 129–136.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

Stefan Schweter. 2020. Italian bert and electra models.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000a. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000b. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Sara Stymne et al. 2020. Evaluating word embeddings for indonesian–english code-mixed text based on synthetic data. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 26–35.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *Ninth International Conference on Spoken Language Processing*.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. *arXiv preprint arXiv:2103.09593*.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Scott Thornbury and Diana Slade. 2006. *Conversation: From description to pedagogy*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at iest 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019d. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

7 Additional details on evaluation corpus

7.1 MIAM: examples and diversity

In this section we give more details on the MIAM benchmark. Tab. 9 shows examples extracted from the benchmark. In Fig. 1 we illustrate the diversity of the gathered corpora through the lens of utterance length.

Lang.	Utterances	DA
de	soll ich dann mit dem Hotel da dann die Buchung vereinbaren	OFFER
	ja das ist gut	FEED. POS.
	das wäre toll	ACCEPT
	dann kümmere ich mich um die Tickets	COMMIT
	wunderbar	ACCEPT
en	how far underneath the diamond mine	ASK
	it's about an inch or so	REPLY
	right okay five inches right along	ACK.
	up along to near a r– a ravine stuff thing	ASK
	no i don't have the ravine	REPLY
es	¿ Qué día desea salir ?	ASK
	El diez de noviembre .	REQUEST
	Quiere horarios de trenes a barcelona,	CONFIRM
	¿ desde zaragoza ?	CONFIRM
	Sí , por favor .	AFF.
fr	Bonjour	GREETINGS
	Bonjour , je suis Sophia l'opérateur (...).	GREETINGS
	Enchanté	GREETINGS
	Qu'est ce que je peux faire pour vous ?	ASK
	J'ai besoins des informations sur les composants de la manette.	INFORMER
it	mangio tre volte al giorno	STATEMENT
	Ti piace mangiare?	QUESTION
	abbastanza	ANSWER
	Che cosa hai mangiato per colazione?	QUESTION
	latte e biscotti	STATEMENT

Table 9: Examples of dialogs labelled with DA taken from MapTask, Dihana, VM2, Loria and Ilisten. AFF. stands for affirmation, FEED. for feedback and ACK. for acknowledgement.

7.2 Altering tasks difficulty

One of the interesting properties of II, mII, NUR, mNUR is the ability to alter the task difficulty in a controlled manner when sampling the negative utterances. For example, instead of randomly sampling the false utterances, the most similar to the true one as measured by a similarity metric (Zhang et al., 2019a; Celikyilmaz et al., 2020) could be chosen. This flexibility could allow increasing the difficulty of the task as models get better.

8 Experimental settings

8.1 Additional details on pretrained models

In this section, we gather additional details on the pretrained models (*e.g* architectures, schema, hyperparameters).

8.1.1 Pretraining losses

Fig. 2 gives graphical examples for each monolingual and multilingual losses used. **Choice of scaling factor in Eq. 1**. In the case of multi-task setting, different losses may have different scales, making the optimization perform poorly. In that case, scaling factors or more advanced techniques (Sener and Koltun, 2018) can be applied. As we did not observe such phenomena, all scaling factors are set to 1.

8.1.2 Pretraining with generation

For both TMUG and MMUG, the model needs to be aware of the target language. Thus, the first token fed to the decoder indicates the target language (*e.g* in English the corresponding id is 99, in Spanish 98). To avoid creating a discrepancy between pretraining objectives we also add this token for MUG.

8.1.3 Choice of the multilingual encoder

The two dominant approaches for multilingual systems involve either using a language-specific encoder (Escolano et al., 2020) or one shared encoder across languages (Feng et al., 2020; Artetxe and Schwenk, 2019). To reduce the number of learnt parameters, we rely on the second approach.

8.1.4 Pretraining details

Our model is pretrained on 4 NVIDIA V100 for 2 days (500k iterations) with a batch size of 256. We use AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2017) with 4000 warmups steps (Vaswani et al., 2017). During this stage, we do not perform any grid search.

8.2 Additional details on downstream task

In this section, we gather additional details on downstream tasks (*e.g* choice of pretrained encoders, choice of decoder and further details on the downstream tasks).

8.2.1 Pretrained encoders baseline

The first group of pretrained encoders are based on BERT. A concatenation of utterances is fed to the model to obtain a conversation embedding. For our language-specific models, we use the German BERT[9], the original BERT for English, BETO (Cañete et al., 2020) for Spanish, Flaubert (Le et al., 2019) for French and Italian BERT Schweter (2020) for Italian. We rely on the multilingual

[9]https://deepset.ai/

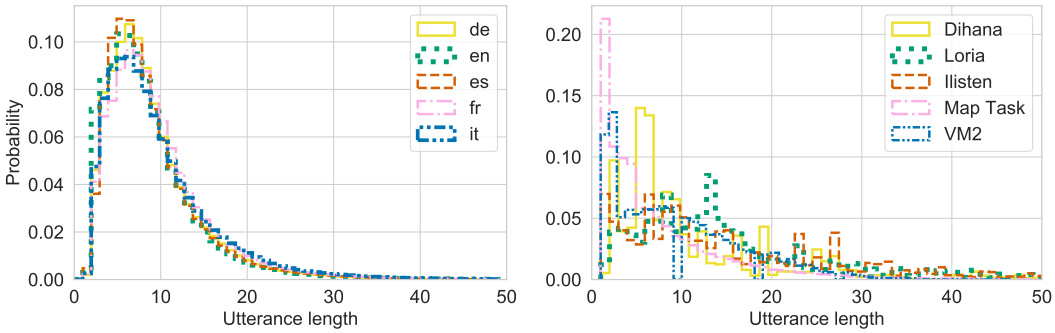
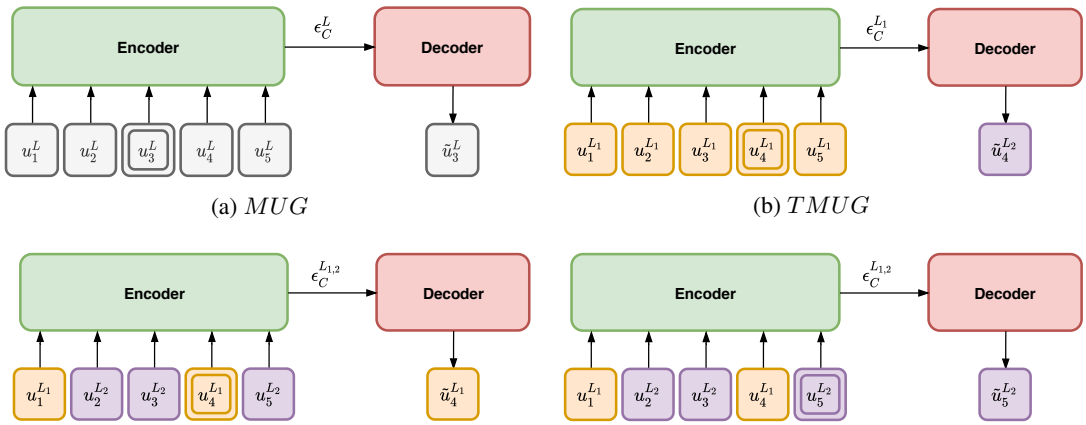


Figure 1: Histograms showing the utterance length for OPS (left) and MIAM (right).



(a) MUG (b) $TMUG$

(c) $MMUG$: prediction of u_4^{L1} (d) $MMUG$: prediction of u_5^{L3}

Figure 2: 2a and 2b illustrate pretraining losses using monolingual context. 2b and 2c show two scenarios for the $MMUG$ loss using multilingual context. Double squares on the figure indicates the randomly selected utterance to predict.

BERT (mBERT) (Devlin et al., 2018)[10] provided by the transformers library (Wolf et al., 2019) implemented using the pytorch (Paszke et al., 2017) framework. For pretrained hierarchical transformers, we rely on the work of Chapuis et al. (2020) and for each considered language, we pretrain a language-specific encoder.

8.2.2 Decoders

Given the different nature of the proposed downstream tasks, we use various type of decoders. **DA classification**: Methods to tackle sequence labelling on monolingual representations can be divided into two different classes. The first one perform classification on each utterance independently using Bayesian Networks (Keizer et al., 2002), SVMs (Surendran and Levow, 2006) or HMMs (Stolcke et al., 2000b). The second class, which achieves stronger results, leverages the adjacency utterances by using deep representations (Bothe et al., 2018; Khanpour et al., 2016). Sequence labelling can be improved when sufficiently many training points are available by modelling inter-tag dependencies using RNN-based decoders (Hochreiter and Schmidhuber, 1997; Chung et al., 2014), and CRFs (Lafferty et al., 2001; Chen et al., 2018b). Thus, in this work, we choose to experiment with a MLP, a CRF and a RNN decoder based on GRU.

II and mII : For this task, the context embedding $\mathcal{E}_{C_k}$ is fed to a MLP. Both the encoder and the MLP are trained to predict the inconsistent utterance index by minimising a cross-entropy loss. Formally, this task is formulated as a classification problem with T classes.

NUR and mNUR: For this task, we first compute the context embedding $\mathcal{E}_{C_k}$ then the candidate utterance $u_c^{L_c}$ is embedded using the either f_θ^u or a chosen encoder to obtain $\mathcal{E}_{u_c^{L_c}}$. Both representations are concatenated and given to a MLP. The architecture is trained to predict if the provided candidate utterance is a suitable next utterance by minimizing

[10] https://github.com/google-research/bert/blob/master/multilingual.md

a binary cross-entropy. This experiment is similar to the one in (Lowe et al., 2015).

8.3 Additional details on models

In this section, we describe models used as well as details on the pretraining parameters. In Tab. 10 we report the main hyper-parameters used for our model pretraining. We used GELU (Hendrycks and Gimpel, 2016) activations and the dropout rate (Srivastava et al., 2014) is set to 0.1. Although vanilla Transformers impose a fixed context size it can be relaxed (Dai et al., 2019). We follow Sankar et al. (2019a); Colombo et al. (2020) and set $T = 5$. We rely on the tokenizers provided by the HuggingFace library based on the SentencePiece (Kudo and Richardson, 2018) and WordPiece (Wu et al., 2016) algorithms. In all experiments, for our models relying on the $\mathcal{HT}$ we use the same architecture as the `SMALL` model from Chapuis et al. (2020) which contains 80 millions parameters. Original BERT has 167 millions parameters and is pretrained using 16 TPUs during several days with over 500K iterations.

	Pretrained Encoder
Nbs of heads	6
N_d	4
N_u	4
T	50
C	5
$\mathcal{T}_d$ nbs of heads	6
Inner dimension	768
Model Dimension	768
$\lvert\mathcal{V}\rvert$	105879
$\mathcal{T}_d$: Emb. size	768
d_k:	64
d_v:	64

Table 10: Architecture hyperparameters used for the hierarchical pretraining.

8.4 Training details

For each task, the model is fine-tuned and dropout (Srivastava et al., 2014) is set to 0.1. The best learning rate is found in $\{0.01, 0.001, 0.0001\}$ and chosen based on the validation loss.

9 Additional experiment: ablation study on pretraining data

We showcase the difference between pretraining with spoken and written corpora. We compare $m\mathcal{HT}(\theta_{written})$, a hierarchical encoder where each utterance is embedded using the representation of the `[CLS]` token given by the second layer of BERT, and $m\mathcal{HT}_u(\theta_{spoken})$, a model pretrained on `OPS` using $\mathcal{L}^u$ only. The prediction is performed by feeding the utterance embeddings to a simple MLP. In Tab. 11, we report the results on `MIAM`. Results demonstrate an overall higher accuracy when the pretraining is performed on spoken data. This supports the choice of `OPS` as pretraining corpora and demonstrates that the origin of the pretraining data matters.

	VM2	Map Task	Dihana	Loria	Ilisten	Total
$m\mathcal{HT}(\theta_{written})$	52.8	64.6	98.1	76.5	**74.2**	73.2
$m\mathcal{HT}_u(\theta_{spoken})$	**53.0**	**67.3**	**98.3**	**78.5**	74.0	**74.2**

Table 11: Ablation studies on pretraining data. We report the accuracy on MIAM for the $m\mathcal{HT}$. $m\mathcal{HT}_u(\theta_{spoken})$ stands for the model pretrained with the utterance level loss $m\mathcal{L}^u$ on spoken data and $m\mathcal{HT}(\theta_{written})$ stands for a hierarchical encoder where sentence embeddings is computed using a pretrained BERT encoder.