Hierarchical Pre-training for Sequence Labelling in Spoken Dialog

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, Chloé Clavel

Hierarchical Pre-training for Sequence Labelling in Spoken Dialog

Emile Chapuis[1]*, **Pierre Colombo**[1,2]*, **Matteo Manica**[3]
Matthieu Labeau[1], **Chloe Clavel**[1]
[1]LTCI, Telecom Paris, Institut Polytechnique de Paris, [2]IBM GBS France, [3]IBM Research Zurich
[1]firstname.lastname@telecom-paris.fr, [3]tte@zurich.ibm.com

Abstract

Sequence labelling tasks like Dialog Act and Emotion/Sentiment identification are a key component of spoken dialog systems. In this work, we propose a new approach to learn generic representations adapted to spoken dialog, which we evaluate on a new benchmark we call Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE benchmark (SILICONE). SILICONE[1] is model-agnostic and contains 10 different datasets of various sizes. We obtain our representations with a hierarchical encoder based on transformer architectures, for which we extend two well-known pre-training objectives. Pre-training is performed on OpenSubtitles: a large corpus of spoken dialog containing over 2.3 billion of tokens. We demonstrate how hierarchical encoders achieve competitive results with consistently fewer parameters compared to state-of-the-art models and we show their importance for both pre-training and fine-tuning.

1 Introduction

The identification of both Dialog Acts (DA) and Emotion/Sentiment (E/S) in spoken language is an important step toward improving model performances on spontaneous dialogue task. Especially, it is essential to avoid the generic response problem, i.e., having an automatic dialog system generate an unspecific response — that can be an answer to a very large number of user utterances (Yi et al., 2019; Colombo et al., 2019). DA and emotion identification (Witon et al., 2018; Jalalzai et al., 2020) are done through sequence labelling systems that are usually trained on large corpora (with over $100k$ labelled utterances) such as Switchboard (Godfrey et al., 1992), MRDA (Shriberg et al., 2004) or Daily Dialog Act (Li et al., 2017). Even though large corpora enable learning complex models from scratch (e.g., seq2seq (Colombo et al., 2020)), those models are very specific to the labelling scheme employed. Adapting them to different sets of emotions or dialog acts would require more annotated data.

Generic representations (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019) have been shown to be an effective way to adapt models across different sets of labels. Those representations are usually trained on large written corpora such as OSCAR (Suárez et al., 2019), Book Corpus (Zhu et al., 2015) or Wikipedia (Denoyer and Gallinari, 2006). Although achieving state-of-the-art (SOTA) results on written benchmarks (Wang et al., 2018), they are not tailored to spoken dialog (SD). Indeed, Tran et al. (2019) have suggested that training a parser on conversational speech data can improve results, due to the discrepancy between spoken and written language (e.g., disfluencies (Stolcke and Shriberg, 1996), fillers (Shriberg, 1999; Dinkar et al., 2020), different data distribution). Furthermore, capturing discourse-level features, which distinguish dialog from other types of text (Thornbury and Slade, 2006), e.g., capturing multi-utterance dependencies, is key to embed dialog that is not explicitly present in pre-training objectives (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019), as they often treat sentences as a simple stream of tokens.

The goal of this work is to train on SD data a generic dialog encoder capturing discourse-level features that produce representations adapted

*stands for equal contribution

[1]Benchmark can be found in the dataset library from HuggingFace (Wolf et al., 2020) at https://huggingface.co/datasets/silicone

to spoken dialog. We evaluate these representations on both DA and E/S labelling through a new benchmark SILICONE (Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE) composed of datasets of varying sizes using different sets of labels. We place ourselves in the general trend of using smaller models to obtain lightweight representations (Jiao et al., 2019; Lan et al., 2019) that can be trained without a costly computation infrastructure while achieving good performance on several downstream tasks (Henderson et al., 2020). Concretely, since hierarchy is an inherent characteristic of dialog (Thornbury and Slade, 2006), we propose the first hierarchical generic multi-utterance encoder based on a hierarchy of transformers. This allows us to factorise the model parameters, getting rid of long term dependencies and enabling training on a reduced number of GPUs. Based on this hierarchical structure, we generalise two existing pre-training objectives. As embeddings highly depend on data quality (Le et al., 2019) and volume (Liu et al., 2019), we preprocess OpenSubtitles (Lison et al., 2019): a large corpus of spoken dialog from movies. This corpora is an order of magnitude bigger than corpora (Budzianowski et al., 2018b; Lowe et al., 2015; Danescu-Niculescu-Mizil and Lee, 2011) used in previous works (Mehri et al., 2019; Hazarika et al., 2019). Lastly, we evaluate our encoder along with other baselines on SILICONE, which lets us draw finer conclusions of the generalisation capability of our models[2].

2 Method

We start by formally defining the Sequence Labelling Problem. At the highest level, we have a set D of conversations composed of utterances, i.e., $D = (C_1, C_2, \ldots, C_{|D|})$ with $Y = (Y_1, Y_2, \ldots, Y_{|D|})$ being the corresponding set of labels (e.g., DA, E/S). At a lower level each conversation C_i is composed of utterances u, i.e $C_i = (u_1, u_2, \ldots, u_{|C_i|})$ with $Y_i = (y_1, y_2, \ldots, y_{|C_i|})$ being the corresponding sequence of labels: each u_i is associated with a unique label y_i. At the lowest level, each utterance u_i can be seen as a sequence of words, i.e $u_i = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$. Concrete examples with dialog act can be found in Table 1.

Utterances	DA
How long does that take you to get to work?	qw
Uh, about forty-five, fifty minutes.	sd
How does that work, work out with, uh, storing your bike and showering and all that?	qw
Yeah ,	b
It can be a pain .	sd
It's, it's nice riding to school because it's all along a canal path, uh,	sd
Because it's just, it's along the Erie Canal up here.	sd
So, what school is it?	qw
Uh, University of Rochester.	sd
Oh, okay.	bk

Table 1: Examples of dialogs labelled with DA taken from SwDA. The labels qw, sd, b, bk respectively correspond to wh-question, statement-non-opinion, backchannel and response acknowledgement.

2.1 Pre-training Objectives

Our work builds upon existing objectives designed to pre-train encoders: the Masked Language Model (MLM) from Devlin et al. (2018); Liu et al. (2019); Lan et al. (2019); Zhang et al. (2019a) and the Generalized Autoregressive Pre-training (GAP) from Yang et al. (2019).

MLM Loss: The MLM loss corrupts sequences (or in our case, utterances) by masking a proportion p_ω of tokens. The model learns bidirectional representations by predicting the original identities of the masked-out tokens. Formally, for an utterance u_i, a random set of indexed positions m^{u_i} is selected and the associated tokens are replaced by a masked token [MASK] to obtain a corrupted utterance u_i^{masked}. The set of parameters θ is learnt by maximizing :

$$\mathcal{L}^u_{\text{MLM}}(\theta, u_i) = \mathbb{E}\left[\sum_{t \in m^{u_i}} \log(p_\theta(\omega_t^i | \tilde{u}_i))\right] \quad (1)$$

where $\tilde{u}_i$ is the corrupted utterance, $m_j^{u_i} \sim unif\{1, |u_i|\}\ \forall\ j \in [1, p_\omega]$ and p_ω is the proportion of masked tokens.

GAP Loss: the GAP loss consists in computing a classic language modelling loss across different factorisation orders of the tokens. In this way, the model will learn to gather information across all possible positions from both directions. The set of parameters θ is learnt by

[2]Upon publication, we will release the code, models and especially the preprocessing scripts to replicate our results.

maximising:

$$\mathcal{L}^u_{\texttt{GAP}}(\theta, u_i) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_{|u_i|}}\left[\sum_t \log p_\theta(\omega^i_{z_t}|u_i^{\mathbf{z}<t})\right]\right] \quad (2)$$

where $\mathbb{Z}_{|u_i|}$ is the set of permutations of length $|u_i|$ and $u_i^{\mathbf{z}<t}$ represent the first t tokens of u_i when permuting the sequence according to $\mathbf{z} \in \mathbb{Z}_{|u_i|}$.

2.2 Hierarchical Encoding

Capturing dependencies at different granularity levels is key for dialog embedding. Thus, we choose a hierarchical encoder (Chen et al., 2018b; Li et al., 2018a). It is composed of two functions f^u and f^c, satisfying:

$$\mathcal{E}_{u_i} = f^u_\theta(\omega_1, \ldots, \omega_{|u_i|}) \quad (3)$$

$$\mathcal{E}_{C_j} = f^d_\theta(\mathcal{E}_{u_1}, \ldots, \mathcal{E}_{C_j}) \quad (4)$$

where $\mathcal{E}_{u_i} \in \mathbb{R}^{d_u}$ is the embedding of u_i and $\mathcal{E}_{C_j} \in \mathbb{R}^{d_d}$ the embedding of C_j. The structure of the hierarchical encoder is depicted in Figure 1.

2.3 Hierarchical Pre-training

2.3.1 General Motivation

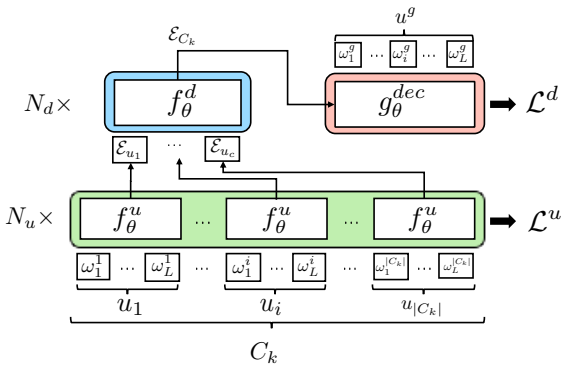


Figure 1: General structure of our proposed hierarchical dialog encoder, with a decoder: f^u_θ, f^d_θ and the sequence label decoder (g^{dec}_θ) are colored respectively in green, blue and red.

Current self-supervised pre-training objectives such as MLM and GAP are trained at the sequence level, which for us translates to only learning f^u_θ. In this section, we extend both the MLM and GAP losses at the dialog level in order to pre-train f^d_θ. Following previous work on both multi-task learning (Argyriou et al., 2007; Ruder, 2017) and hierarchical supervision (Garcia et al., 2019; Sanh et al., 2019), we argue that optimising simultaneously at both levels rather than separately improves the quality of the resulting embeddings. Thus, we write our global hierarchical loss as:

$$\mathcal{L}(\theta) = \lambda_u * \mathcal{L}^u(\theta) + \lambda_d * \mathcal{L}^d(\theta) \quad (5)$$

where $\mathcal{L}^u(\theta)$ is either the MLM or GAP loss at the utterance level and $\mathcal{L}^d(\theta)$ is its generalisation at the dialog level.

2.3.2 MLM Loss

The MLM loss at the utterance level is defined in Equation 1. Our generalisation at the dialog level masks a proportion $p_\mathcal{C}$ of utterances and generates the sequences of masked tokens (a concrete example can be found in Appendix B). Thus, at the dialog level the MLM loss is defined as:

$$\mathcal{L}^d_{\texttt{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j\in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega^j_i|\tilde{C}_k))\right] \quad (6)$$

where $m_j^{C_k} \sim unif\{1, |C_k|\}\ \forall\ j \in [1, p_\mathcal{C}]$ is the set of positions of masked utterances in the context C_k, $\tilde{C}_k$ is the corrupted context, and $p_\mathcal{C}$ is the proportion of masked utterances.

2.3.3 GAP Loss

The GAP loss at the utterance level is defined in Equation 2. A possible generalisation of the GAP at the dialog level is to compute the loss of the generated utterance across all factorization orders of the context utterances. Formally, the GAP loss is defined at the dialog level as:

$$\mathcal{L}^d_{\texttt{GAP}}(\theta, C_k) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_T}\left[\sum_{t=1}^{|C_k|}\sum_{i=1}^{|u_{z_t}|} \log p_\theta(\omega^{z_t}_i|C_k^{\mathbf{z}<t})\right]\right] \quad (7)$$

where $\omega^{z_t}_i$ denotes the first i-th tokens of the permuted t-th utterance when permuting the context according to $\mathbf{z} \in \mathbb{Z}_T$ and $C_k^{\mathbf{z}<t}$ the first t utterances of C_k when permuting the context according to $\mathbf{z}$.

2.4 Architecture

Commonly, The functions f^u_θ and f^d_θ are either modelled with recurrent cells (Serban et al., 2015) or Transformer blocks (Vaswani et al.,

2017). Transformer blocks are more parallelizable, offering shorter paths for the forward and backward signals and requiring significantly less time to train compared to recurrent layers. To the best of our knowledge this is the first attempt to pre-train a hierarchical encoder based only on transformers[3].

The structure of the model can be found in Figure 1. In order to optimize dialog level losses as described in Equation 5, we generate (through g_θ^{dec}) the sequence with a Transformer Decoder ($\mathcal{T}_{dec}$). For downstream tasks, the context embedding $\mathcal{E}_{C_k}$ is fed to a simple MLP (simple classification), or to a CRF/GRU/LSTM (sequential prediction) — see Appendix B for more details. In the rest of the paper, we will name our hierarchical transformer-based encoder $\mathcal{HT}$ and the hierarchical RNN-based encoder $\mathcal{HR}$. We use θ_y^x to refer to the set of model parameters learnt using the pre-training objective y (either `MLM` or `GAP`) at the level x[4].

2.5 Pre-training Datasets

Datasets used to pre-train dialog encoders (Hazarika et al., 2019; Mehri et al., 2019) are often medium-sized (e.g. Cornell Movie Corpus (Danescu-Niculescu-Mizil and Lee, 2011), Ubuntu (Lowe et al., 2015), MultiWOz (Budzianowski et al., 2018a)). In our work, we focus on OpenSubtitles (Lison and Tiedemann, 2016)[5] because (1) it contains spoken language, contrarily to the Ubuntu corpus (Lowe et al., 2015) based on logs; (2) as Wizard of Oz (Budzianowski et al., 2018a) and Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil and Lee, 2011), it is a multi-party dataset; and (3) OpenSubtitles is an order of magnitude larger than any other spoken language dataset used in previous work. We segment OpenSubtitles by considering the duration of the silence between two consecutive utterances. Two consecutive utterances belong to the same conversation if the silence is shorter than δ_T[6]. Conversations shorter than the context size T are dropped[7]. After preprocessing, Opensubtitles contains subtitles from 446520 movies or series which represent 54642424 conversations and over 2.3 billion of words.

2.6 Baseline Encoder

We compare the different methods we presented with two different types of baseline encoders: pre-trained encoders, and hierarchical encoders based on recurrent cells. The latter, achieve current SOTA performance in many sequence labelling tasks (Li et al., 2018a; Colombo et al., 2020; Lin et al., 2017).

Pre-trained Encoder Models. We use BERT (Devlin et al., 2018) through the pytorch implementation provided by the Hugging Face transformers library (Wolf et al., 2019). The pre-trained model is fed with a concatenation of the utterances. Formally given an input context $C_k = (u_1, \dots u_T)$ the concatenation $[u_1, \dots, u_T]$ is fed to BERT.

Hierarchical Recurrent Encoders. In this work we rely on our own implementation of the model based on $\mathcal{HR}$. Hyperparameters are described in Appendix B.

3 Evaluation of Sequence Labelling

3.1 Related Work

Sequence labelling tasks for spoken dialog mainly involve two different types of labels: `DA` and `E/S`. Early work has tackled the sequence labelling problem as an independent classification of each utterance. Deep neural network models that currently achieve the best results (Keizer et al., 2002; Surendran and Levow, 2006; Stolcke et al., 2000) model both contextual dependencies between utterances (Colombo et al., 2020; Li et al., 2018b) and labels (Chen et al., 2018b; Kumar et al., 2018; Li et al., 2018c).

The aforementioned methods require large corpora to train models from scratch, such as: Switchboard Dialog Act (`SwDA`) (Godfrey et al., 1992), Meeting Recorder Dia-

[3] Although it is possible to relax the fixed size imposed by transformers (Dai et al., 2019) in this paper we follow (Colombo et al., 2020) and fix the context size to 5 and the max utterance length to 50 — these choices are made to work with OpenSubtitles, since the number of available dialogs drops when considering a number of utterances greater than 5.

[4] if $x = u$ solely utterance level training is used, if $x = d$ solely dialog level is used and if $x = u, d$ multi level supervision is used ($\lambda_u, \lambda_d \in \{0, 1\}^2$ according to the case.)

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

[6] We choose $\delta_T = 6s$

[7] Using pre-training method based on the next utterance proposed by Mehri et al. (2019) requires dropping conversation shorter than $T + 1$ leading to a non-negligible loss in the preprocessing stage.

log Act (MRDA) (Shriberg et al., 2004), Daily Dialog Act (Li et al., 2017), HCRC Map Task Corpus (MT) (Thompson et al., 1993). This makes harder their adoption to smaller datasets, such as: Loqui human-human dialogue corpus (Loqui) (Passonneau and Sachar., 2014), BT Oasis Corpus (Oasis) (Leech and Weisser, 2003), Multimodal Multi-Party Dataset (MELD) (Poria et al., 2018a), Interactive emotional dyadic motion capture database (IEMO), SEMAINE database (SEM) (Mckeown et al., 2013).

3.2 Presentation of SILICONE

Despite the similarity between methods usually employed to tackle DA and E/S sequential classification, studies usually rely on a single type of label. Moreover, despite the variety of small or medium-sized labelled datasets, evaluation is usually done on the largest available corpora (e.g., SwDA, MRDA). We introduce SILICONE, a collection of sequence labelling tasks, gathering both DA and E/S annotated datasets. SILICONE is built upon preexisting datasets which have been considered by the community as challenging and interesting. Any model that is able to process multiple sequences as inputs and predict the corresponding labels can be evaluated on SILICONE. We especially include small-sized datasets, as we believe it will ensure that well-performing models are able to both distil substantial knowledge and adapt to different sets of labels without relying on a large number of examples. The description of the datasets composing the benchmark can be found in the following sections, while corpora statistics are gathered in Table 2.

3.2.1 DA Datasets

Switchboard Dialog Act Corpus (SwDA) is a telephone speech corpus consisting of two-sided telephone conversations with provided topics. This dataset includes additional features such as speaker id and topic information. The SOTA model, based on a seq2seq architecture with guided attention, reports an accuracy of 85.5% (Colombo et al., 2020) on the official split.

ICSI MRDA Corpus (MRDA) has been introduced by Shriberg et al. (2004). It contains transcripts of multi-party meetings hand-annotated with DA. It is the second biggest dataset with around $110k$ utterances. The SOTA model reaches an accuracy of 92.2% (Li et al., 2018a) and uses Bi-LSTMs with attention as encoder as well as additional features, such as the topic of the transcript.

DailyDialog Act Corpus (DyDA_{a}) has been produced by Li et al. (2017). It contains multi-turn dialogues, supposed to reflect daily communication by covering topics about daily life. The dataset is manually labelled with dialog act and emotions. It is the third biggest corpus of SILICONE with $102k$ utterances. The SOTA model reports an accuracy of 88.1% (Li et al., 2018a), using Bi-LSTMs with attention as well as additional features. We follow the official split introduced by the authors.

HCRC MapTask Corpus (MT) has been introduced by (Thompson et al., 1993). To build this corpus, participants were asked to collaborate verbally by describing a route from a first participant's map by using the map of another participant. This corpus is small ($27k$ utterances). As there is no standard train/dev/test split[8] performances depends on the split. Tran et al. (2017) make use of a Hierarchical LSTM encoder with a GRU decoder layer and achieves an accuracy of 65.9%.

Bt Oasis Corpus (Oasis) contains the transcripts of live calls made to the BT and operator services. This corpus has been introduced by (Leech and Weisser, 2003) and is rather small ($15k$ utterances). There is no standard train/dev/test split [9] and few studies use this dataset.

3.2.2 S/E Datasets

In S/E recognition for spoken language, there is no consensus on the choice the evaluation metric (e.g., Ghosal et al. (2019); Poria et al. (2018b) use a weighted F-score while Zhang et al. (2019b) report accuracy). For SILICONE, we choose to stay consistent with the DA research and thus follow Zhang et al. (2019b) by reporting the accuracy. Additionally, emotion/sentiment labels are neither merged nor prepossessed[10].

[8] We split according to the code in https://github.com/NathanDuran/Maptask-Corpus.

[9] We use a random split from https://github.com/NathanDuran/BT-Oasis-Corpus.

[10] Comparison with concurrent work is more difficult as system performance heavily depends on the number of classes and label processing varies across studies

DailyDialog Emotion Corpus (DyDA$_e$) has been previously introduced and contains eleven emotional labels. The SOTA model (De Bruyne et al., 2019) is based on BERT with additional Valence Arousal and Dominance features and reaches an accuracy of 85% on the official split.

Multimodal EmotionLines Dataset (MELD) has been created by enhancing and extending EmotionLines dataset (Chen et al., 2018a) where multiple speakers participated in the dialogues. There are two types of annotations MELD$_s$ and MELD$_e$: three sentiments (positive, negative and neutral) and seven emotions (anger, disgust, fear, joy,neutral, sadness and surprise). The SOTA model with text only is proposed by Zhang et al. (2019b) and is inspired by quantum physics. On the official split, it is compared with a hierarchical bi-LSTM, which it beats with an accuracy of 61.9% (MELD$_s$) and 67.9% (MELD$_e$) against 60.8% and 65.2.

IEMOCAP database (IEMO) is a multimodal database of ten speakers. It consists of dyadic sessions where actors perform improvisations or scripted scenarios. Emotion categories are: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. There is no official split on this dataset. One proposed model is built with bi-LSTMs and achieves 35.1%, with text only (Zhang et al., 2019b).

SEMAINE database (SEM) comes from the Sustained Emotionally coloured Machine human Interaction using Nonverbal Expression project (Mckeown et al., 2013). This dataset has been annotated on three sentiments labels: positive, negative and neutral by Barriere et al. (2018). It is built on Multimodal Wizard of Oz experiment where participants held conversations with an operator who adopted various roles designed to evoke emotional reactions. There is no official split on this dataset.

4 Results on SILICONE

This section gathers experiments performed on the SILICONE benchmark. We first analyse an appropriate choice for the decoder, which is selected over a set of experiments on our baseline encoders: a pre-trained BERT model and a hierarchical RNN-based encoder ($\mathcal{HR}$). Since we focus on small-sized pre-trained representations, we limit the sizes of our pre-trained models to TINY and SMALL (see Table 7). We then study the results of the baselines and our hierarchical transformer encoders ($\mathcal{HT}$) on SILICONE along three axes: the accuracy of the models, the difference in performance between the E/S and the DA corpora, and the importance of pre-training. As we aim to obtain robust representations, we do not perform an exhaustive grid search on the downstream tasks.

4.1 Decoder Choice

Current research efforts focus on single label prediction, as it seems to be a natural choice for sequence labelling problems (subsection 2.1). Sequence labelling is usually performed with CRFs (Chen et al., 2018b; Kumar et al., 2018) and GRU decoding (Colombo et al., 2020), however, it is not clear to what extent inter-label dependencies are already captured by the contextualised encoders, and whether a plain MLP decoder could achieve competitive results. As can be seen in Table 3, we found that in the case of E/S prediction there is no clear difference between CRFs and MLPs, while GRU decoders exhibit poor performance, probably due to a lack of training data. It is also important to notice, that training a sequential decoder usually requires thorough hyper-parameter fine-tuning. As our goal is to learn and evaluate general representations that are decoder agnostic, in the following, we will use a plain MLP decoder for all the models compared.

4.2 General Performance Analysis

Table 4 provides an exhaustive comparison of the different encoders over the SILICONE benchmark. As previously discussed, we adopt a plain MLP as a decoder to compare the different encoders. We show that SILICONE covers a set of challenging tasks as the best performing model achieves an average accuracy of 74.3. Moreover, we observe that despite having half the parameters of a BERT model, our proposed model achieves an average result that is 2% higher on the benchmark. SILICONE covers two different sequence labelling tasks: DA and E/S. In Table 4 and Table 3, we can see that all models exhibit a consistently higher

(Clavel and Callejas, 2015).

Corpus	$\|Train\|$	$\|Val\|$	$\|Test\|$	Utt.	$\|Labels\|$	Task	Utt./$\|Labels\|$
SwDA*	1k	100	11	200k	42	DA	4.8k
MRDA*	56	6	12	110k	5	DA	2.6k
DyDA_{a}	11k	1k	1k	102k	4	DA	25.5k
MT*	121	22	25	36k	12	DA	3k
Oasis*	508	64	64	15k	42	DA	357
DyDA_{e}	11k	1k	1k	102k	7	E	2.2k
$\text{MELD}^{\star}_{\text{s}}$	934	104	280	13k	3	S	4.3k
$\text{MELD}^{\star}_{\text{e}}$	934	104	280	13k	7	S	1.8k
IEMO	108	12	31	10k	6	E	1.7k
SEM	62	7	10	5,6k	3	S	1.9k

Table 2: Statistics of datasets composing SILICONE. E stands for emotion label and S for sentiment label; * stands for datasets with available official split. Sizes of Train, Val and Test are given in number of conversations.

	Avg	Avg DA	Avg E/S
BERT (+MLP)	72,8	81.5	64.0
BERT (+GRU)	69.9	80.4	59.3
BERT (+CRF)	72.8	81.5	64.1
$\mathcal{HR}$ (+MLP)	69.8	79.1	60.4
$\mathcal{HR}$ (+GRU)	67.6	79.4	55.7
$\mathcal{HR}$ (+CRF)	70.5	80.3	60.7

Table 3: Experiments comparing decoder performances. Results are given on SILICONE for two types of baseline encoders (pre-trained BERT models and hierarchical recurrent encoders $\mathcal{HR}$).

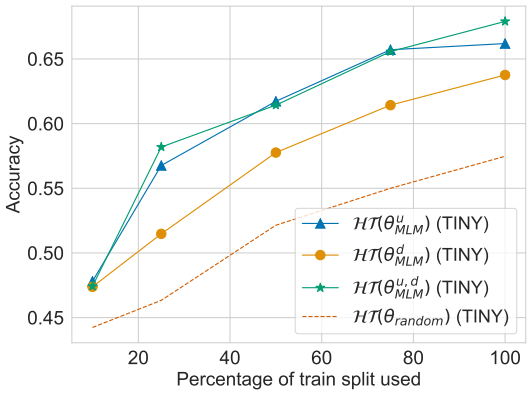


Figure 2: A comparison of pre-trained encoders being fine-tuned on different percentage the training set of SEM. Validation and test set are fixed over all experiments, reported scores are averaged over 10 different random split.

average accuracy (up to 14%) on DA tagging compared to E/S prediction. This performance drop could be explained by the different sizes of the corpora (see Table 2). Despite having a larger number of utterances per label (u/l), E/S tasks seem generally harder to tackle for the models. For example, on Oasis, where the u/l is inferior than those of most E/S datasets (MELD_{s}, MELD_{e}, IEMO and SEM), models consistently achieve better results.

4.3 Importance of Pre-training for SILICONE

Results reported in Table 4 and Table 3 show that pre-trained transformer-based encoders achieve consistently higher accuracy on SILICONE, even when they are not explicitly considering the hierarchical structure. This difference can be observed both in small-sized datasets (e.g. MELD and SEM) and in medium/large size datasets (e.g SwDA and MRDA). To validate the importance of pre-training in a regime of low data, we train different $\mathcal{HT}$ (with random initialisation) on different portions of SEM and MELD_{s}. Results shown in Figure 2 illustrate the importance of pre-trained representations.

5 Model Analysis

In this section, we dissect our hierarchical pre-trained models in order to better understand the relative importance of each component. We show how a hierarchical encoder allows us to obtain a light and efficient model. Additional experiments can be found in Appendix C.

5.1 Pre-training on Spoken vs Written Data

First, we explore the differences in training representations on spoken and written corpora. Experimentally, we compare the predictions on SILICONE made by $\mathcal{HT}(\theta^{u}_{MLM})$ and the one made by $\mathcal{HT}(\theta_{BERT-2layers})$. The latter is a

	Avg	SwDA	MRDA	DyDA_{DA}	MT	Oasis	DyDA_{e}	MELD_{s}	MELD_{e}	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
BERT	72.8	79.2	90.7	**82.6**	88.2	66.9	91.9	59.3	**61.4**	**45.0**	62.7
$\mathcal{HR}$	69.8	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HT}(\theta^{u,d}_{MLM})_{\texttt{(TINY)}}$	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{GAP})_{\texttt{(TINY)}}$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{HT}(\theta^{u,d}_{MLM})_{\texttt{(SMALL)}}$	**74.3**	79.2	**92.4**	81.5	**90.6**	**69.4**	**92.7**	**64.1**	60.1	**45.0**	**68.2**

Table 4: Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size. MT stands for Map Task, IEM for IEMOCAP and Sem for Semaine.

	Avg DA	Avg E/S
BERT (4 layers)	80.5	60.2
$\mathcal{HT}(\theta_{BERT-2layers})$	80.5	61.1
$\mathcal{HT}$ (θ^{u}_{MLM})	**80.8**	**64.0**

Table 5: Results of ablation studies on SILICONE

hierarchical encoder where utterance embeddings are obtained with the hidden vector representing the first token [CLS] (see (Devlin et al., 2018)) of the second layer of BERT. In both cases, predictions are performed using an MLP[11]. Results in Table 5 show higher accuracy when the pre-training is performed on spoken data. Since SILICONE is a spoken language benchmark, this result might be due to the specific features of colloquial speech (e.g. disfluencies, sentence length, vocabulary, word frequencies).

5.2 Hierarchy and Multi-Level Supervision

We study the relative importance of three aspects of our hierarchical pre-training with multi-level supervision. We first show that accounting for the hierarchy increases the performance of fine-tuned encoders, even without our specific pre-training procedure. We then compare our two proposed hierarchical pre-training procedures based on the GAP or MLM loss. Lastly, we look at the contribution of the possible levels of supervision on reduced training data from SEM.

[11]We consider the two first layer for a fair comparison based on the number of model parameters.

5.2.1 Importance of hierarchical fine-tuning

We compare the performance of BERT-4layers with the $\mathcal{HT}(\theta_{BERT-2layer})$ previously described. Results reported in Table 5 demonstrate that fine-tuning on downstream tasks with a hierarchical encoder yields to higher accuracy, with fewer parameters, even when using already pre-trained representations.

5.2.2 MLM vs GAP

In this experiment, we compare the different pre-training objectives at utterance and dialog level. As a reminder $\mathcal{HT}(\theta^{u}_{MLM})$ and $\mathcal{HT}(\theta^{u}_{GAP})$ are respectively trained using the standard MLM loss (Devlin et al., 2018) and the standard GAP loss (Yang et al., 2019). In Table 6 we report the different pre-training objective results. We observe that pre-training at the dialog level achieves comparable results to the utterance level pre-training for MLM and slightly worse for GAP. Interestingly, we observe that $\mathcal{HT}(\theta^{u}_{GAP})$ compared to $\mathcal{HT}(\theta^{u}_{MLM})$ achieves worse results, which is not consistent with the performance observed on other benchmarks, such as GLUE (Wang et al., 2018). The lower accuracy of the models trained using a GAP-based loss could be due to several factors (e.g., model size, pre-training using the GAP loss could require a finer choice of hyperparameters). Finally, we see that supervising at both dialog and utterance level helps for MLM[12].

[12]We investigate a similar setting for GAP which lead to poor results, the loss hit a plateau suggesting that objectives are competing against each other. More advanced optimisations techniques (Sener and Koltun, 2018) are left for future work.

	Avg DA	Avg E/S
$\mathcal{HT}(\theta^{u}_{MLM})$	80.8	64.0
$\mathcal{HT}(\theta^{d}_{MLM})$	80.8	64.0
$\mathcal{HT}(\theta^{u}_{GAP})$	80.7	62.0
$\mathcal{HT}(\theta^{d}_{GAP})$	80.4	62.8
$\mathcal{HT}(\theta^{u,d}_{MLM})$	**81.9**	**64.7**

Table 6: Comparison of GAP and MLM with a comparable number of parameters. For all models a MLP decoder is used on top of a TINY pre-trained encoder.

	Emb.	Word	Seq	Total
BERT		87		110
BERT (4-layer)		43		66
HMLP	23	8.6	7.8	40
(TINY)		2.9	2.8	28.7
(SMALL)		10.6	10.6	45

Table 7: Number of parameters for the encoders. Sizes are given in million of parameters.

5.2.3 Multi level Supervision for pre-training

In this section, we illustrate the advantages of learning using several levels of supervision on small datasets. We fine-tune different model on SEM using different size of the training set. Results are shown in Figure 2. Overall we see that introducing sequence level supervision induces a consistent improvement on SEM. Results on MELD_s are provided in Appendix C.

5.3 Other advantages of hierarchy

Introducing a hierarchical design in the encoder allows to break dialog into utterances and to consider inputs of size T instead of size 512. First, it allows parameters sharing, reducing the number of model parameters. The different model sizes are reported in Table 7. Our TINY model contains half the parameters of BERT (4-layers). Furthermore, modelling long-range dependencies hierarchically makes learning faster and allows to get rid of learning tricks (e.g., partial order prediction (Yang et al., 2019), two-stage pre-training based on sequence length (Devlin et al., 2018)) required for non-hierarchical encoders. Lastly, original BERT and XLNET are pre-trained using respectively 16 and 512 TPUs. Pre-training lasts several days with over $500K$ iterations. Our TINY hierarchical models are pre-trained during $180K$ iterations (1.5 days) on 4 NVIDIA V100.

6 Conclusions

In this paper, we propose a hierarchical transformer-based encoder tailored for spoken dialog. We extend two well-known pre-training objectives to adapt them to a hierarchical setting and use OpenSubtitles, the largest spoken language dataset available, for encoder pre-training. Additionally, we provide an evaluation benchmark dedicated to comparing sequence labelling systems for the NLP community, SILICONE, on which we compare our models and pre-training procedures with previous approaches. By conducting ablation studies, we demonstrate the importance of using a hierarchical structure for the encoder, both for pre-training and fine-tuning. Finally, we find that our approach is a powerful method to learn generic representations on spoken dialog, with less parameters than state-of-the-art transformer models.

These results open new future research directions: (1) to investigate new pre-training objectives leveraging the hierarchical framework in order to achieve better results on SILICONE while keeping light models (2) to provide multilingual models using the whole pre-training corpus (OpenSubtitles) available in 62 languages, (3) investigate robust methods (Staerman et al., 2020a) and the application of our embedding to different anomaly detection settings (Staerman et al., 2019, 2020b). We hope that the SILICONE benchmark, experimental results, and publicly available code encourage further research to build stronger sequence labelling systems for NLP.

Acknowledgement

This work was supported by a grant overseen from the French National Research Agency (ANR-17-MAOI).

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2007. Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48.

Valentin Barriere, Chloé Clavel, and Slim Essid. 2018. Attitude classification in adjacency pairs of a human-agent interaction with hidden conditional random fields. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4949–4953. IEEE.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018a. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018b. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

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

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.

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, 7(1):74–93.

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

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*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2019. Joint emotion label space modelling for affect lexica. *arXiv preprint arXiv:1911.08782*.

Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 12–19. Springer.

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*.

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. *arXiv preprint arXiv:1908.11216*.

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.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2019. Emotion recognition in conversations with transfer learning from generative conversation modeling. *arXiv preprint arXiv:1910.04980*.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *arXiv preprint arXiv:2002.05651*.

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

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*.

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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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*.

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

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

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

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2018c. A dual-attention hierarchical recurrent neural network for dialogue act classification. *arXiv preprint arXiv: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.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

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.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee,

and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018b. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956.

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

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808.

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.

Elizabeth E Shriberg. 1999. Phonetic consequences of speech disfluency. Technical report, SRI INTERNATIONAL MENLO PARK CA.

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.

Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. 2020a. When ot meets mom: Robust estimation of wasserstein distance. *arXiv preprint arXiv:2006.10325*.

Guillaume Staerman, Pavlo Mozharovskyi, Stéphan Clémen, et al. 2020b. The area of the convex hull of sampled curves: a robust functional statistical depth measure. In *International Conference on Artificial Intelligence and Statistics*, pages 570–579.

Guillaume Staerman, Pavlo Mozharovskyi, Stephan Clémençon, and Florence d'Alché Buc. 2019. Functional isolation forest. *arXiv preprint arXiv:1904.04573*.

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

Andreas Stolcke and Elizabeth Shriberg. 1996. Statistical language modeling for speech disfluencies. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 405–408. IEEE.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.

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*.

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

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

Quan Hung Tran, Gholamreza Haffari, and Ingrid Zukerman. 2017. A generative attentional neural network model for dialogue act classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 524–529.

Trang Tran, Jiahong Yuan, Yang Liu, and Mari Ostendorf. 2019. On the role of style in parsing speech with neural models. *Proc. Interspeech 2019*, pages 4190–4194.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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*.

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.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019a. 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. 2019b. Quantum-inspired interactive networks for conversational sentiment analysis.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

	TINY	SMALL
Nbs of heads	1	6
N_d	2	4
N_u	2	4
T	50	50
C	5	5
$\mathcal{T}_d$ nbs of heads	6	6
Inner dimension	768	768
Model Dimension	768	768
Vocab length	32000	32000
$\mathcal{T}_d$: Emb. size	768	768
d_k:	64	64
d_v:	64	64

Table 8: Architecture hyperparameters used for the hierarchical pre-training.

A Additional Details on data composing SILICONE

In this section, we illustrate the diversity of the dataset composing SILICONE. In Figure 3, we plot two histograms representing the different utterance lengths for DA and E/S. As expected, for spoken dialog, lengths are shorter than for written benchmarks (e.g., GLUE).

B Additional Details for Models

In this section we report model hyper-parameters and as well as additional descriptions of our baselines. For all models we use a tokenizer based on WordPiece (Wu et al., 2016).
We also provide a concrete example of corrupted context for the MLM Loss.

B.1 Hierarchical pre-training

We report in Table 8 the main hyper-parameters used fo our model pre-training. We used GELU (Hendrycks and Gimpel, 2016) activations and the dropout rate (Srivastava et al., 2014) is set to 0.1.

B.2 MLM Loss example

In this section we propose a visual illustration of the corrupted context Figure 4 by the MLM Loss.

B.3 Experimental Hyper-parameters for SILICONE

For all models, we use a batch size of 64 and automatically select the best model on the validation set according to its loss. We do not perform exhaustive grid search either on the learning rate (that is set to 10^{-4}), nor on other hyper-parameters to perform a fair comparison between all the models. We use ADAMW (Kingma and Ba, 2014; Loshchilov and Hutter, 2017) with a linear scheduler on the learning rate and the number of warm-up steps is set to 100.

B.4 Additional Details on Baselines

A representation for all the baselines can be found in Figure 5. For all models, both hidden dimension and embedding dimension is set to 768 to ensure fair comparison with the proposed model. The MLP used for decoding contains 3 layers of sizes $(768, 348, 192)$. We use RELU (Agarap, 2018) to introduce non linearity inside our architecture.

C Additional Experimental Results

In this section we report the detailed results on SILICONE, including the ones presented in Table 4. We report results on two new experiments: importance of pre-training time for both a TINY and SMALL model, we report the convergence time of a TINY model and finally we extend subsubsection 5.2.3 by reporting results on IEMO.

C.1 Detailed Results on SILICONE

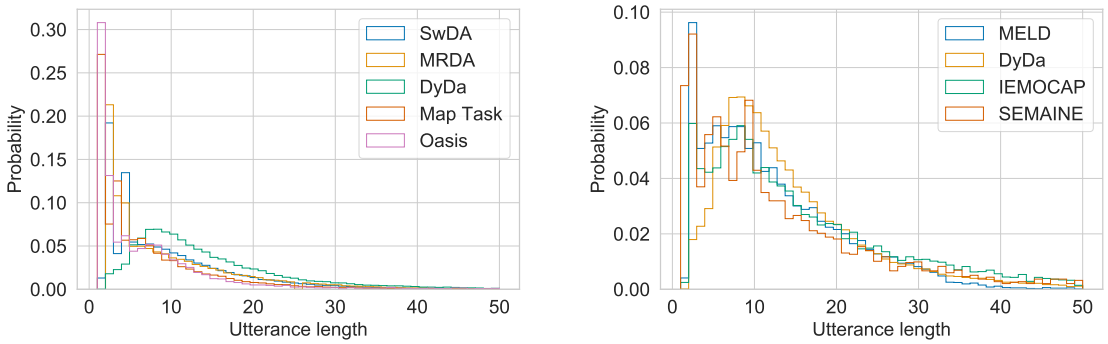
We show in Table 9 the results on the SILICONE benchmark for all the models mentioned in the paper.

C.2 Improvement over pre-training

In this experiment we illustrate how pre-training improves performance on SEM (see Figure 6). As expected accuracy improves when pre-training.

C.3 Multi level Supervision for pre-training MELD

In this experiment we report results of the experiment mentioned in subsubsection 5.2.3. In this experiment we see that the training process seems to be noisier for fractions lower than 40%. For larger percentages, we observe that including higher supervision (at the dialog level) during pre-training leads to a consistent improvement.



(a) SILICONE DA (b) SILICONE S/E

Figure 3: Histograms showing the utterance length for each dataset of SILICONE.

(a) Initial context composed by 5 utterances.

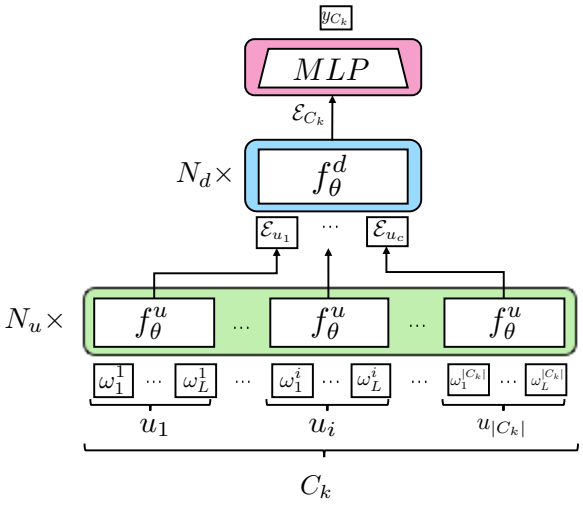
(b) u_1 is chosen to be masked.

(c) Corrupted context with utterance u_1 masked.

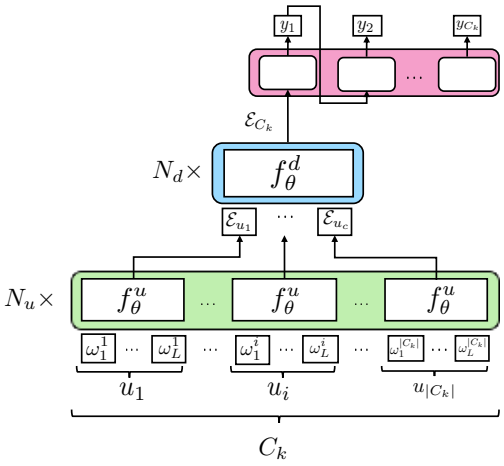
(d) u_4 is chosen to be masked.

(e) Corrupted context with utterance u_4 masked.

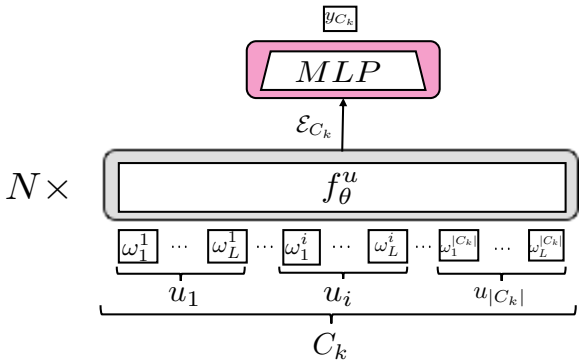
Figure 4: This figure shows an example of corrupted context. Here p_C is randmoly set to 2 meaning that two utterances will be corrupted. u_1 and u_4 are randomly picked in 4b, 4d and then masked in 4c, 4e.



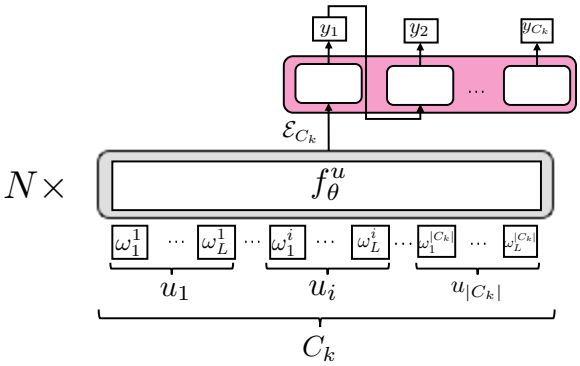
(a) Hierarchical encoder with MLP decoder performing single label prediction.



(b) Hierarchical encoder with sequential decoder (either GRU or CRF).



(c) BERT encoder with MLP decoder performing single label prediction.



(d) BERT encoder with sequential decoder (either GRU or CRF)

Figure 5: Schema of the different models evaluated on SILICONE. In this figure f_θ^u, f_θ^d and the sequence label decoder (g_θ^{dec}) are respectively colored in green, blue and red for the hierarchical encoder (see Figure 5a and Figure 5d). For BERT there is no hierarchy and embedding is performed through f_θ^u colored in grey (see Figure 5c, Figure 5d)

	Avg	SwDA	MRDA	DyDA_{DA}	MT	Oasis	DyDA_{e}	MELD_{s}	MELD_{e}	IEMO	SEM
BERT-4layers (+MLP)	69.45	77.8	90.7	79.0	88.4	66.8	90.3	49.3	50.4	43.0	58.8
BERT (+MLP)	72.79	79.2	90.7	82.6	88.2	66.9	91.9	59.3	61.4	45.0	62.7
BERT (+GRU)	69.84	78.2	90.4	80.8	88.7	63.7	90	50.4	48.9	45.0	62.3
BERT (+CRF)	72.8	79.0	90.8	88.3	67.2	81.9	91.5	59.4	61.0	44.2	61.5
$\mathcal{HR}$ (+MLP)	69.77	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HR}$ (+GRU)	67.54	78.2	90.9	79,9	84,4	63,5	91.5	50,7	50.4	35.2	50.7
$\mathcal{HR}$ (+CRF)	70.5	77.8	91,3	79,7	87,5	65,3	91,1	62,1	57,4	42.1	50.7
$\mathcal{HT}(\theta^{u,d}_{MLM})$ (TINY)	73.3	79.3	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{MLM})$ (TINY)	72.4	78.5	91.8	78.0	89.8	66.0	92.5	62.6	59.3	42.0	63.5
$\mathcal{HT}(\theta^{u}_{MLM})$ (TINY)	72.4	78.6	91.8	79.0	89.8	65.0	91.8	61.8	58.1	39.2	68.9
HBERT (w) θ_{BERT_milmil} (TINY)	70.8	77.6	91.4	79.3	88.3	65.8	91.9	58.0	56.3	40.0	59.1
$\mathcal{HT}(\theta^{u,d}_{MLM})$ (SMALL)	74.32	79.2	92.4	81.5	90.6	69.4	92.7	64.1	60.1	45.0	68.2
$\mathcal{HT}(\theta^{d}_{GAP})$ (TINY)	71.58	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{HT}(\theta^{u}_{GAP})$ (TINY)	71.52	78.5	90.9	79.0	88.9	66.3	92.0	59.2	57.5	39.9	63.0

Table 9: Performances of all mentioned model with different decoders such as MLP, GRU, CRF SILICONE. The datasets are grouped by label type (DA vs E/S) and order by decreasing size.

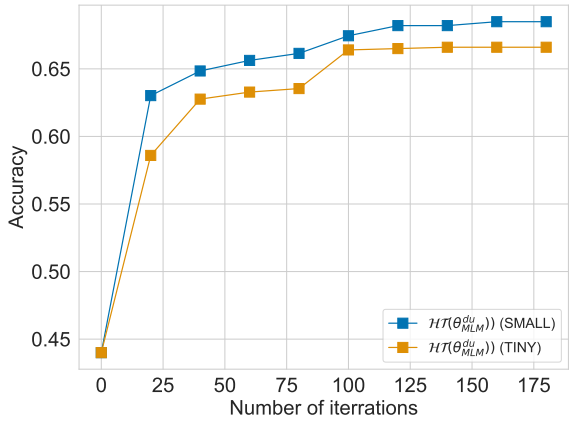


Figure 6: Illustration of improvement of accuracy during pre-training stage on SEM for both a TINY and SMALL model.

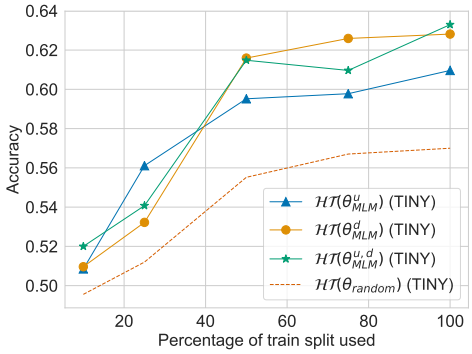


Figure 7: A comparison of different parameters initialisation on $\texttt{MELD}_\texttt{s}$. Training is performed using a different percentage of complete training set. Validation and test set are fixed over all experimentation. Each score is the averaged accuracy over 10 random runs.

D Negative Results on GAP

We briefly describe few ideas we tried to make GAP works at both the utterance and dialog level. We hypothesise that:

- giving the same weight to the utterance level and the dialog level (see Equation 3) was responsible of the observed plateau. Different combinations lead to fairly poor improvements.
- the limited model capacity was part of the issue. Larger models does not give the expected results.