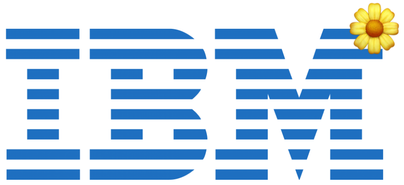
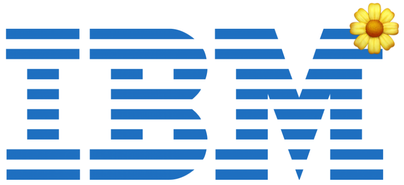


Hierarchical Pre-training for Sequence Labelling in SpokenDialog

Oral Presentation at WACAI 2021

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

* Equal contribution



Hierarchical Pre-training for Sequence Labelling in SpokenDialog

Oral Presentation at WACAI 2021

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

* Equal contribution

Sequence Labelling

Sequence Labelling

What’d you do, Prison Mike ?

I stole. And I robbed.

Euhhhh !

That is quite the rap sheet, Prison Mike.

And I never got caught neither!

Well, you are in prison…

Sequence Labelling

Dialog/Speech Act

What'd you do, Prison Mike ?

Question

I stole. And I robbed.

Answer

Euhhhh !

Back-channel

That is quite the rap sheet, Prison Mike.

Statement opinion

And I never got caught neither!

Statement

Well, you are in prison…

Appreciation

Importance of Sequence labelling

Importance of Sequence labelling

Types of label considered:

Dialog Act

DIT++ / DAMSL / DiAML

Emotion and Sentiment

Polarity (+/-)

6 emotions (+neutrals)

Importance of Sequence labelling

Types of label considered:

Dialog Act

DIT++ / DAMSL / DiAML

Emotion and Sentiment

Polarity (+/-)

6 emotions (+neutrals)

Why are sequence labelling tasks useful?

Speaker modelling

Dialog State Tracking

Avoid generic response problem

Importance of Sequence labelling

Types of label considered:

Dialog Act

DIT++ / DAMSL / DiAML

Emotion and Sentiment

Polarity (+/-)

6 emotions (+neutrals)

Why are sequence labelling tasks useful?

Speaker modelling

Dialog State Tracking

Avoid generic response problem

Generic response problem

What'd you do, Prison Mike ?

I don't know

Didn't you rob someone?

I don't know

Limitation of current systems

Limitation of current systems

Current state-of-the art considers deep learning models

Limitation of current systems

Current state-of-the art considers deep learning models

Recurrent encoder + MLP

Limitation of current systems

Current state-of-the art considers deep learning models

- Recurrent encoder + MLP
- Contextual recurrent encoder + MLP

Limitation of current systems

Current state-of-the art considers deep learning models

- Recurrent encoder + MLP
- Contextual recurrent encoder + MLP
- Contextual recurrent encoder + CRF/Recurrent decoder

Limitation of current systems

Current state-of-the art considers deep learning models

- Recurrent encoder + MLP
- Contextual recurrent encoder + MLP
- Contextual recurrent encoder + CRF/Recurrent decoder

Middle/High size corpora

- Switchboard Dialog Act (100k + utterances)
- MRDA (110k + utterances)
- Dialy Dialog Act (100k + utterances)

Limitation of current systems

Current state-of-the art considers deep learning models

- Recurrent encoder + MLP
- Contextual recurrent encoder + MLP
- Contextual recurrent encoder + CRF/Recurrent decoder

Middle/High size corpora

- Switchboard Dialog Act (100k + utterances)
- MRDA (110k + utterances)
- Dialy Dialog Act (100k + utterances)

Not practical but deep learning is data hungry!

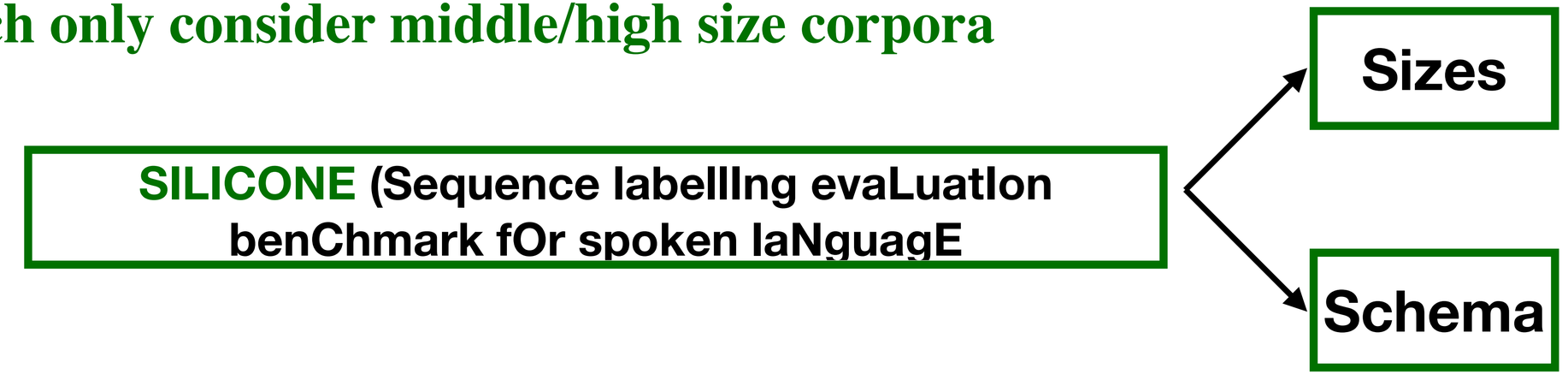
Contributions

Contributions

Research only consider middle/high size corpora

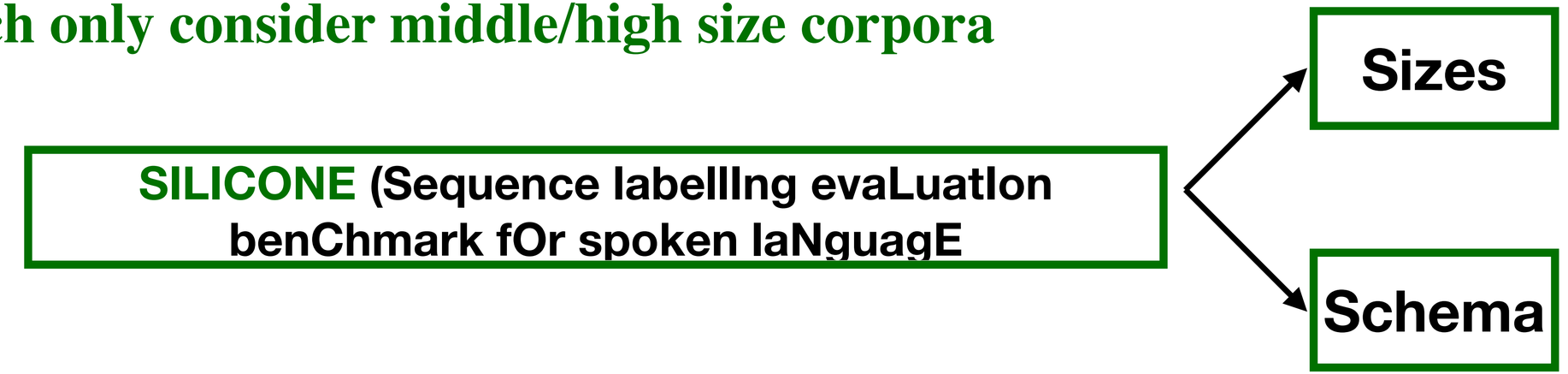
Contributions

Research only consider middle/high size corpora



Contributions

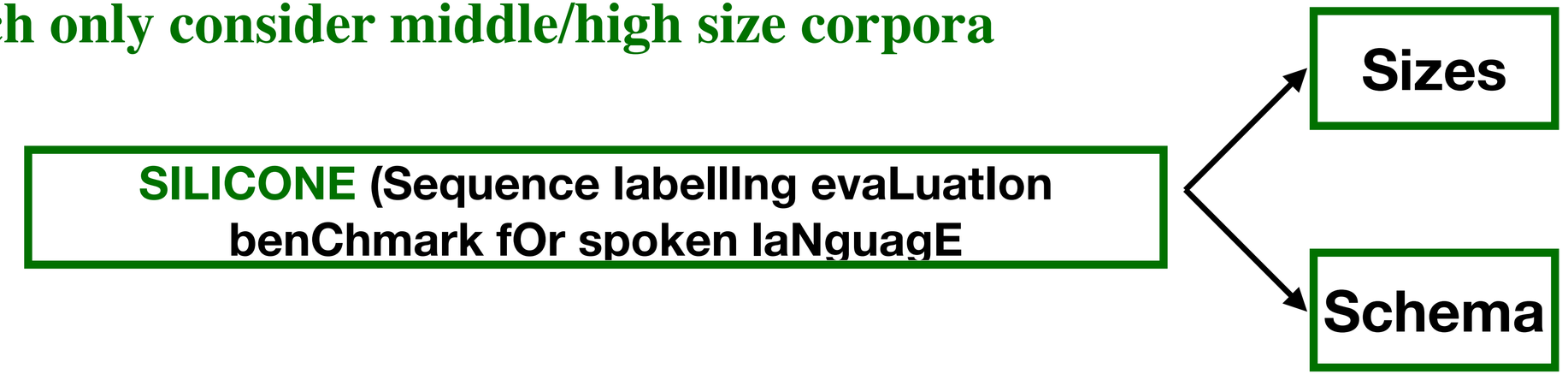
Research only consider middle/high size corpora



We propose a new model

Contributions

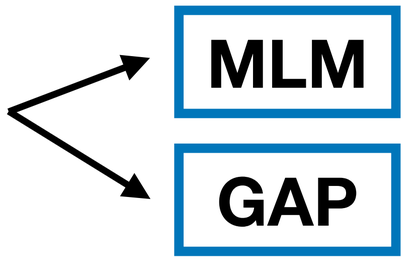
Research only consider middle/high size corpora



We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives



Contributions

Research only consider middle/high size corpora

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE

- Sizes
- Schema

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives

- MLM
- GAP

Pretraining Copora

Contributions

Research only consider middle/high size corpora

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE

- Sizes
- Schema

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives

- MLM
- GAP

Pretraining Copora

OpenSubtitles

- Spoken dialogs
- Large Scale
- Pretraining

Formalisation

Formalisation

What’d you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's son and held him for ransom.

That is quite the rap sheet, Prison Mike.

And I never got caught neither!

Well, you are in prison…

Formalisation

u_1  What'd you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's son and held him for ransom.

 u_2

u_3  That is quite the rap sheet, Prison Mike.

And I never got caught neither!

 u_4

u_5  Well, you are in prison…

Formalisation

Conversation C

u_1  What'd you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's son and held him for ransom.  u_2

u_3  That is quite the rap sheet, Prison Mike.

And I never got caught neither!  u_4

u_5  Well, you are in prison…

$$C_i = (u_1, u_2, \ldots, u_{|C_i|})$$

Formalisation

Formalisation

u_1 

Qu’est ce que tu as fait, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's son and held him for ransom.

 u_2

u_3 

That is quite the rap sheet , Prison Mike .

ω_1^4 ω_2^4 ω_3^4 ω_4^4 ω_5^4 ω_6^4 ω_7^4 ω_8^4 ω_9^4 ω_{10}^4

Et on ne m’a jamais chopé non plus!

 u_4

u_5 

Well, you are in prison…

Formalisation

u_1 

Qu’est ce que tu as fait, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's son and held him for ransom.

 u_2

u_3 

That	is	quite	the	rap	sheet	,	Prison	Mike	.
ω_1^4	ω_2^4	ω_3^4	ω_4^4	ω_5^4	ω_6^4	ω_7^4	ω_8^4	ω_9^4	ω_{10}^4

Et on ne m’a jamais chopé non plus!

 u_4

u_5 

Well, you are in prison…

$$u_i^{L_i} = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$$

Models

Models

Hierarchical Encoder composed of two functions: f^u **and** f^d

Models

Hierarchical Encoder composed of two functions: f^u **and** f^d

$$\mathscr{E}_{u_i} = f^u_\theta(\omega^i_1, \ldots, \omega^i_{|u_i|}),$$

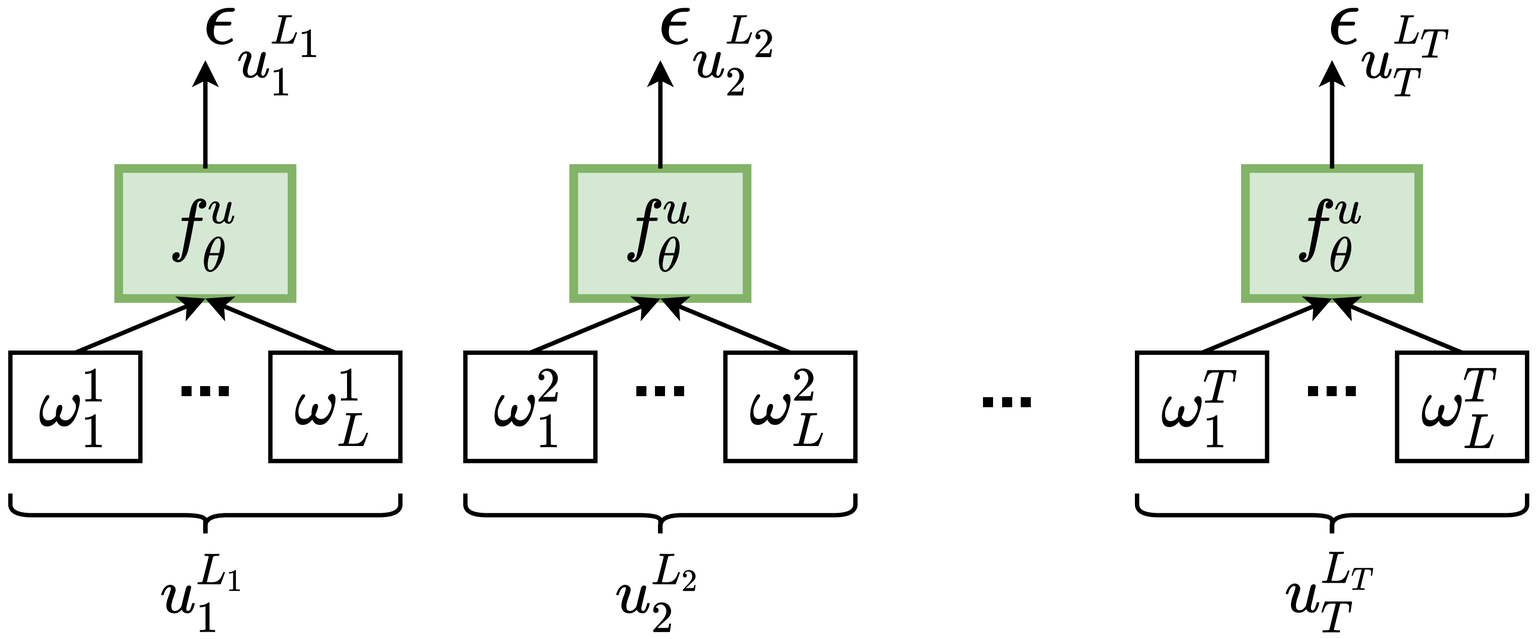
$$\mathscr{E}_{C_j} = f^d_\theta(\mathscr{E}_{u_1}, \ldots, \mathscr{E}_{u_{|C_j|}}),$$

Models

Hierarchical Encoder composed of two functions: f^u **and** f^d

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

$$\mathscr{E}_{C_j} = f_\theta^d(\mathscr{E}_{u_1}, \ldots, \mathscr{E}_{u_{|C_j|}}),$$

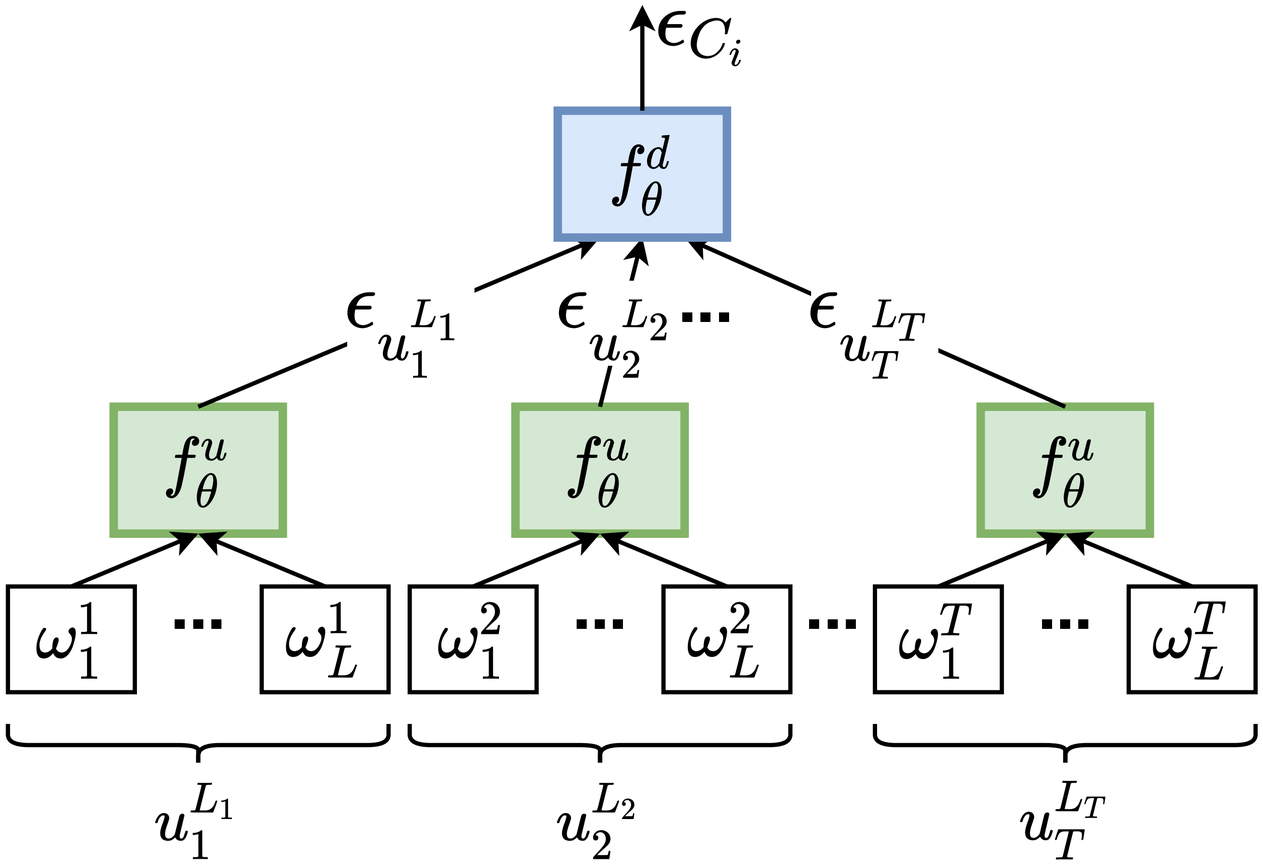


Models

Hierarchical Encoder composed of two functions: f^u **and** f^d

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

$$\mathscr{E}_{C_j} = f_\theta^d(\mathscr{E}_{u_1}, \ldots, \mathscr{E}_{u_{|C_j|}}),$$



Training Objective

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

Models

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \, .$$

Models

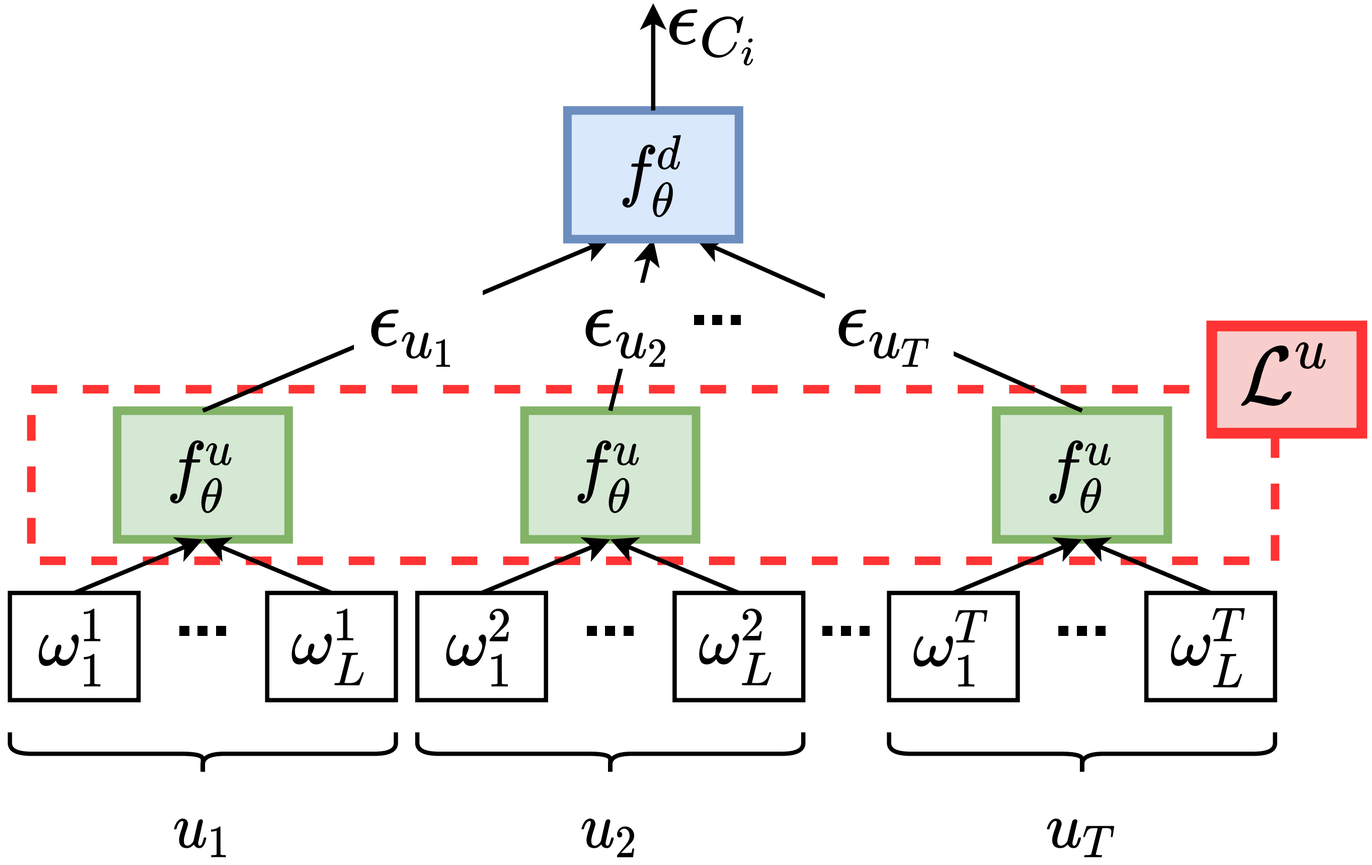
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \, .$$

Utterance Level Pretraining

Models

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

Utterance Level Pretraining



Losses

Utterance Level Pretraining

$$\mathcal{L}(\theta) = \lambda_u \mathcal{L}^u(\theta) + \lambda_d \mathcal{L}^d(\theta) \,.$$

Losses

Utterance Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Utterance Modelling (MUM)

Losses

Utterance Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Masked Utterance Modelling (MUM)

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

Losses

Utterance Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Masked Utterance Modelling (MUM)

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

Set of masked tokens

Set of masked indices

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

Set of masked tokens

Set of masked indices

u_3

That	is	quite	the	rap	sheet	,	Prison	Mike	.
ω_1^4	ω_2^4	ω_3^4	ω_4^4	ω_5^4	ω_6^4	ω_7^4	ω_8^4	ω_9^4	ω_{10}^4

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

Set of masked tokens

Set of masked indices

u_3

That	is	[MASK]	the	rap	[MASK]	,	Prison	Mike	.
ω_1^4	ω_2^4	ω_3^4	ω_4^4	ω_5^4	ω_6^4	ω_7^4	ω_8^4	ω_9^4	ω_{10}^4

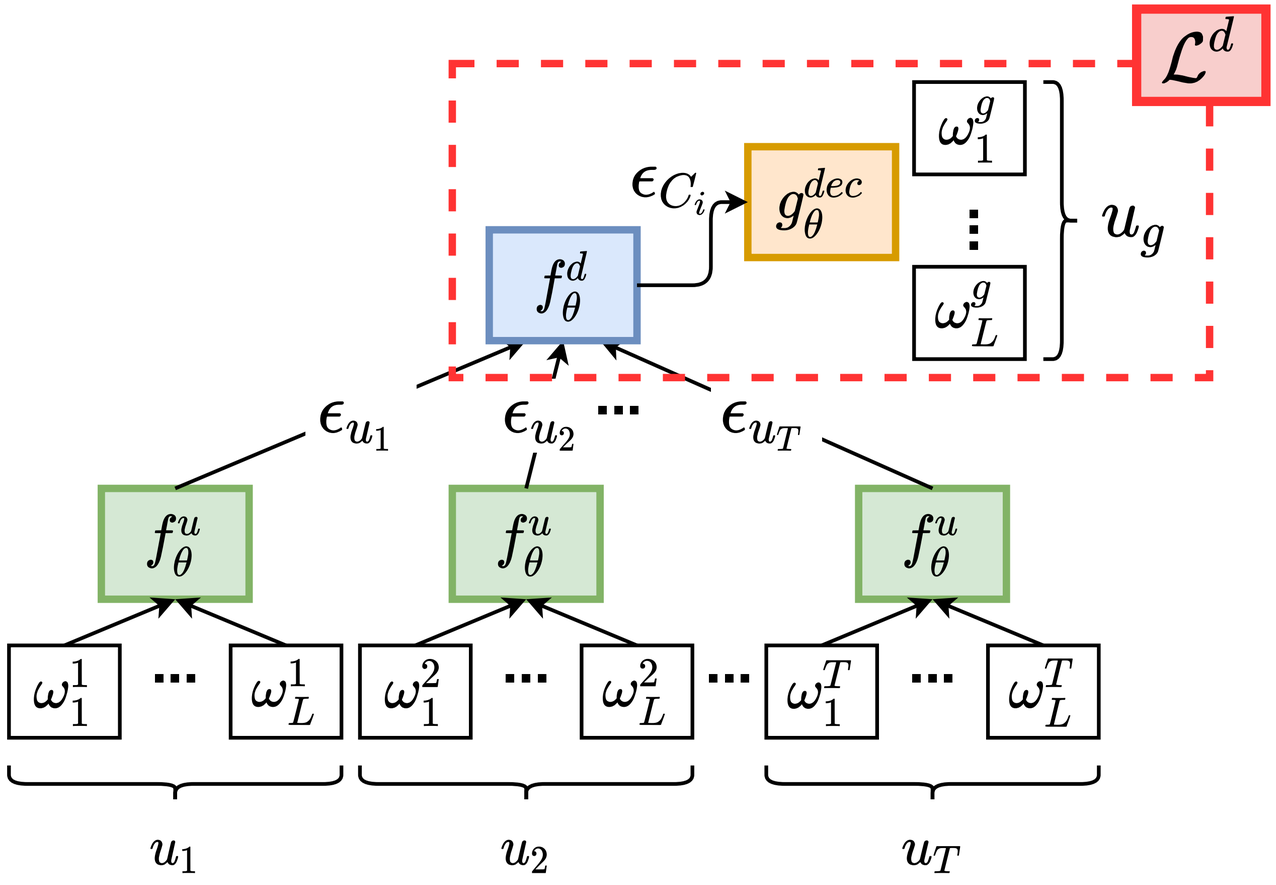
Predict → quite sheet

Models

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \, .$$

Models

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

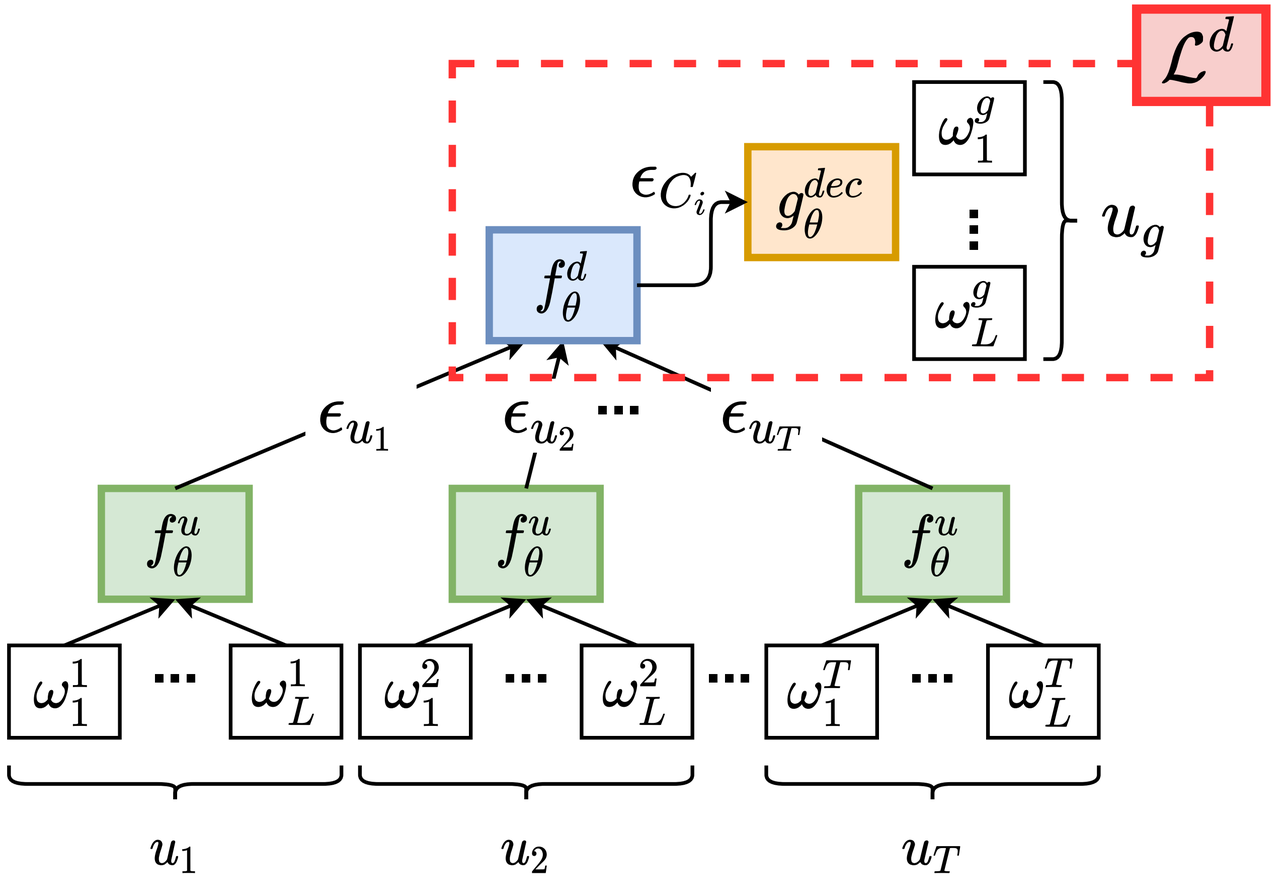


Models

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

Dialog Level Pretraining:

- **Masked Sequence Generation**
- **Add a *causal decoder***



Models

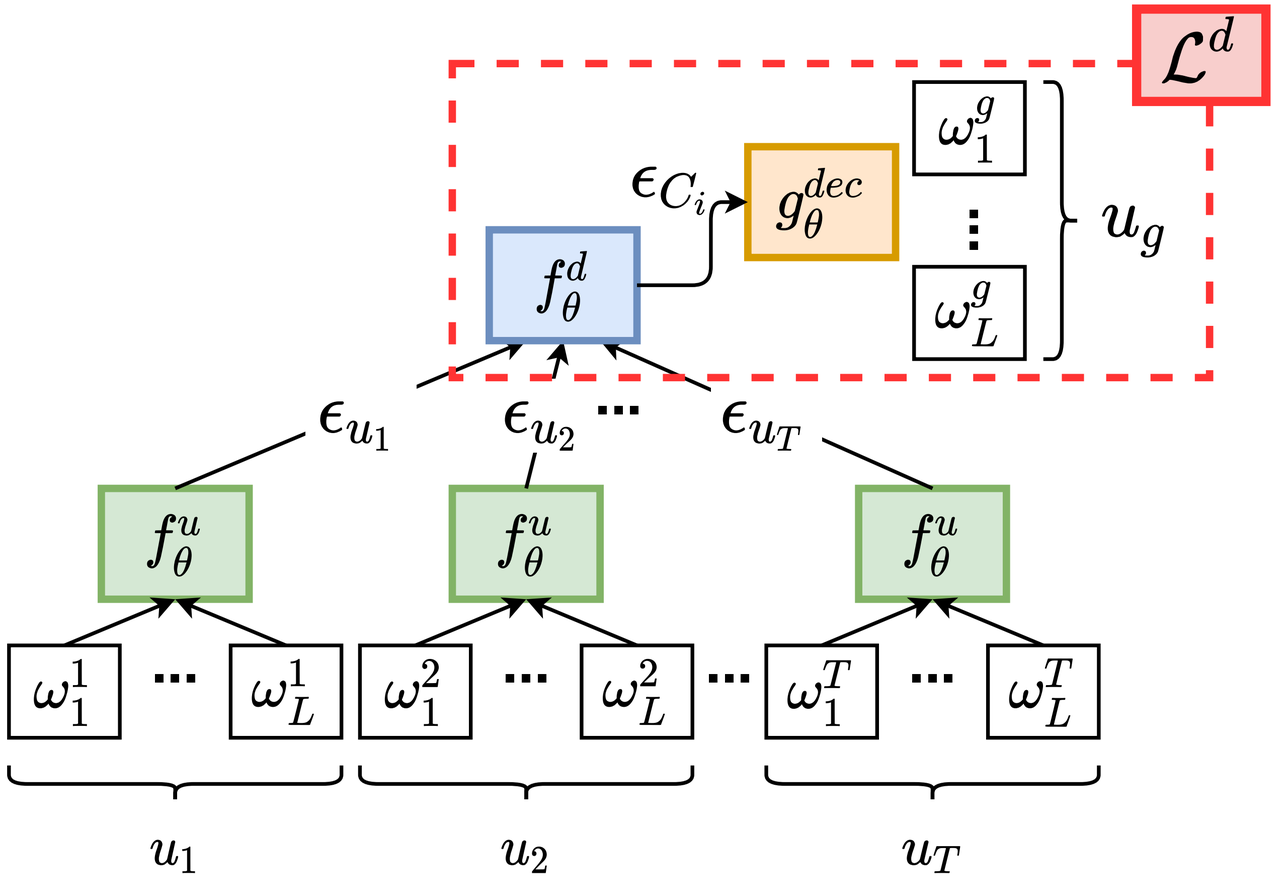
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

Dialog Level Pretraining:

- **Masked Sequence Generation**
- **Add a *causal decoder***

Goal:

- **Learnt the inter-utterance dependancies**
- **Train the 2nd level of the Transformer**



Utterance Level MLM Loss

Dialog Level Pretraining $\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)$.

Utterance Level MLM Loss

Dialog Level Pretraining $\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

Utterance Level MLM Loss

Dialog Level Pretraining $\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

Utterance Level MLM Loss

Dialog Level Pretraining $\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

Utterance Level MLM Loss

Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

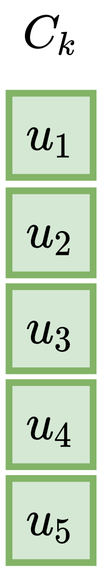
$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

Utterance Level MLM Loss

Dialog Level Pretraining $\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)$.

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

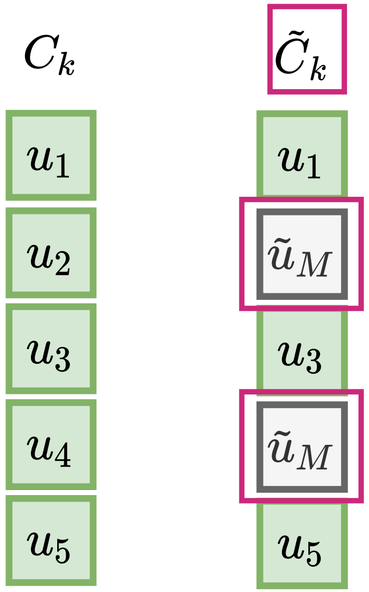


Utterance Level MLM Loss

Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)



$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$

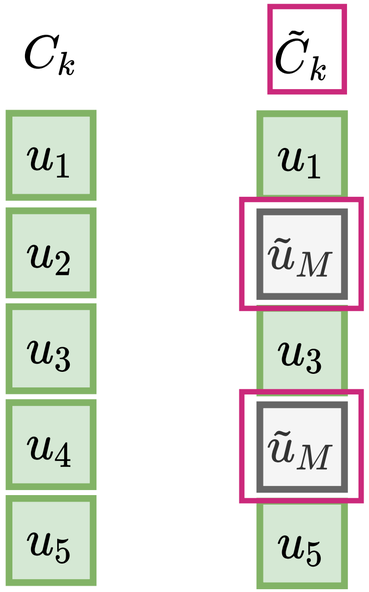
Utterance Level MLM Loss

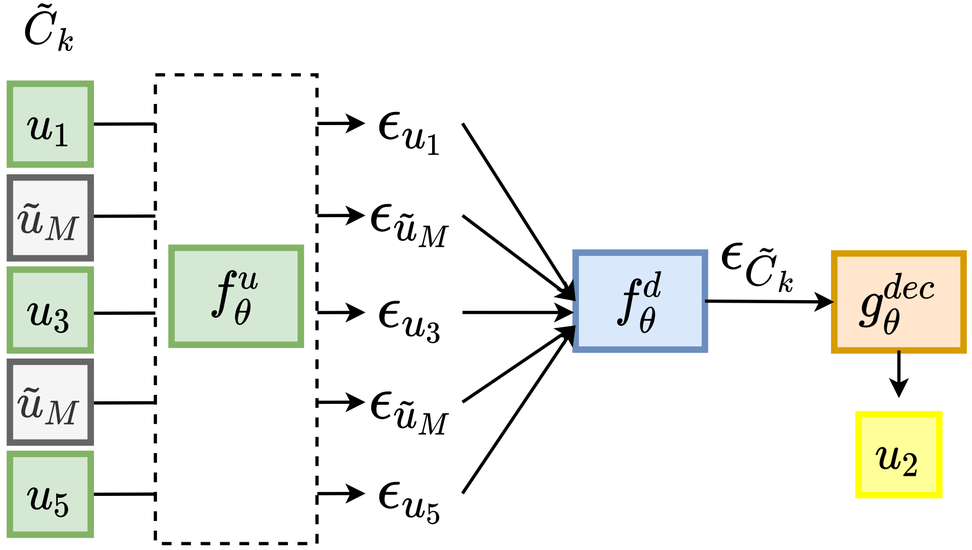
Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$





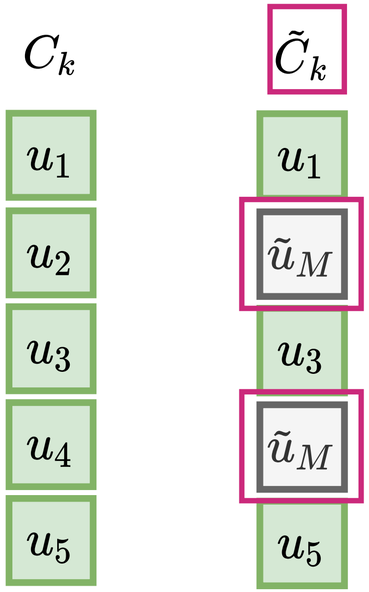
Utterance Level MLM Loss

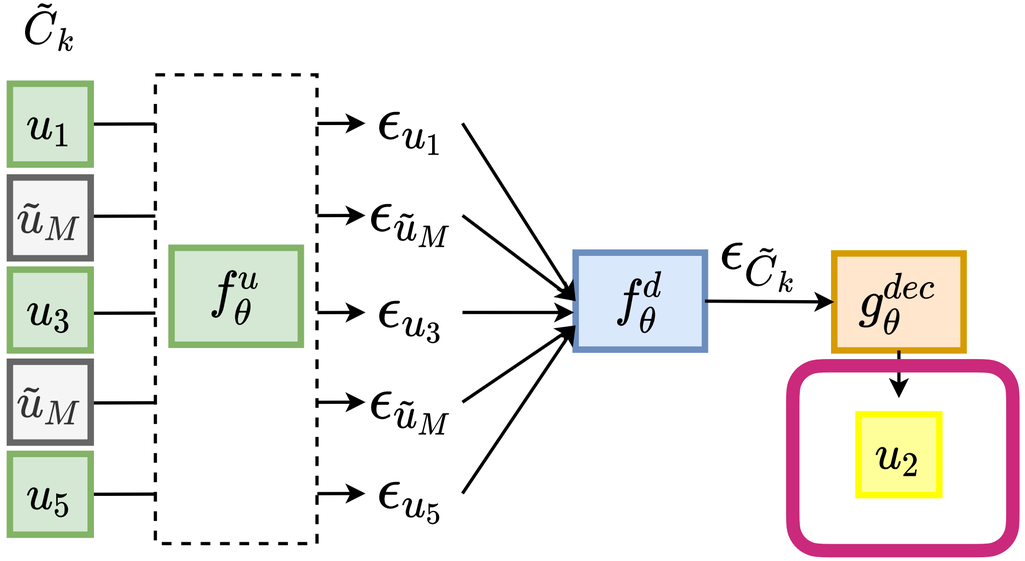
Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$





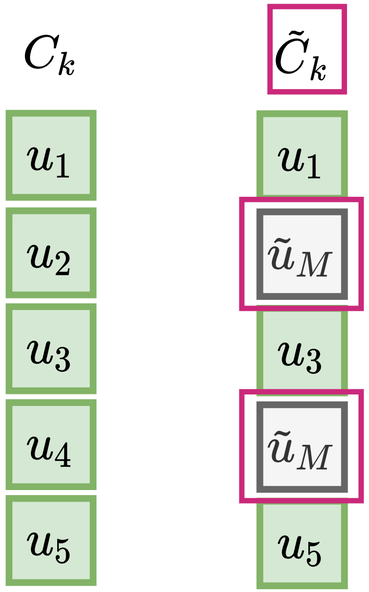
Utterance Level MLM Loss

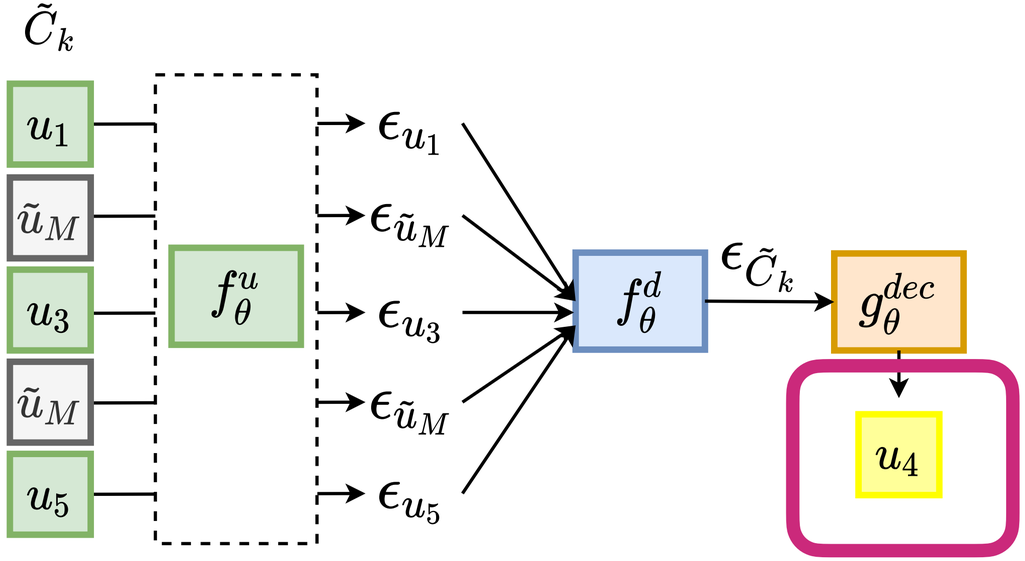
Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{MLM}}(\theta, C_k) = \mathbb{E}\left[\sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j \mid \tilde{C}_k))\right]$$





SILICONE

Research only consider middle/high size corpora

SILICONE

Research only consider middle/high size corpora

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

SILICONE

Research only consider middle/high size corpora

Dialog Acts

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

SILICONE

Research only consider middle/high size corpora

Dialog Acts

Emotions & Sentiments

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

General Performances

Results on SILICONE

General Performances

Results on SILICONE

	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**

General Performances

Results on SILICONE

	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**

General Performances

Results on SILICONE

	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**

General Performances

Results on SILICONE

	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**

Increasing model size is a good idea!

General Performances

Results on SILICONE

	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
$\mathcal{HT}(\theta^{u,d}_{MLM})$ (TINY)	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}(\theta^{d}_{GAP})$ (TINY)	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{HT}(\theta^{u,d}_{MLM})$ (SMALL)	**74.3**	79.2	**92.4**	81.5	**90.6**	**69.4**	**92.7**	**64.1**	60.1	**45.0**	**68.2**

Our hierarchical pertaining helps!

Increasing model size is a good idea!

Results I

Results I

Decoder Choice

Results I

Decoder Choice

	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

Results I

Decoder Choice

	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

Sequential decoder helps !

Results I

Decoder Choice

	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

Convergence Speed

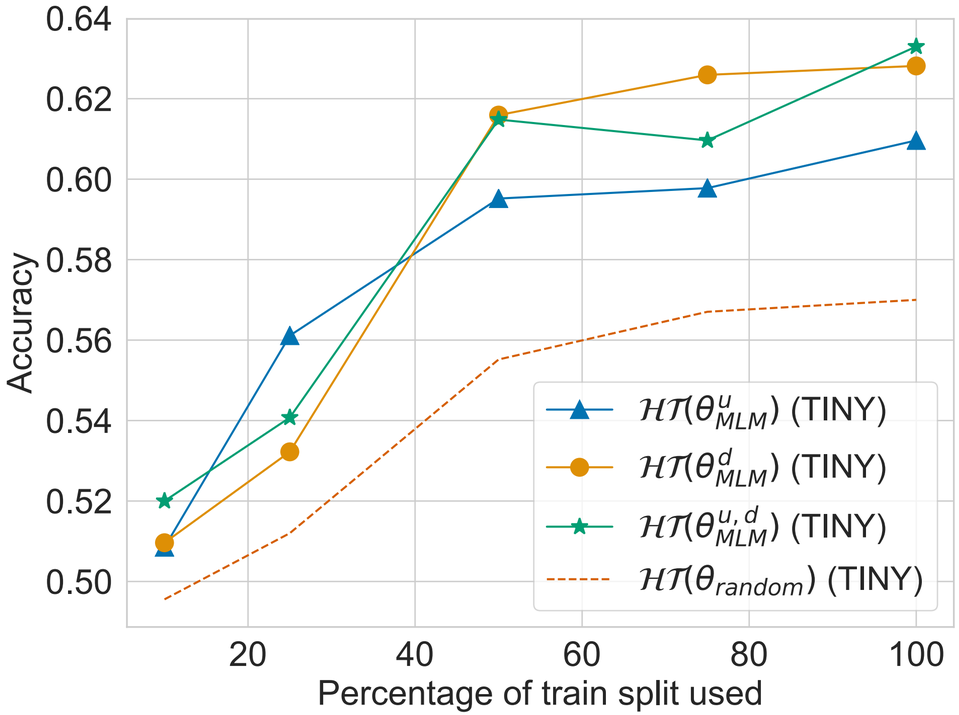
Sequential decoder helps !

Results I

Decoder Choice

	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

Convergence Speed



Sequential decoder helps !

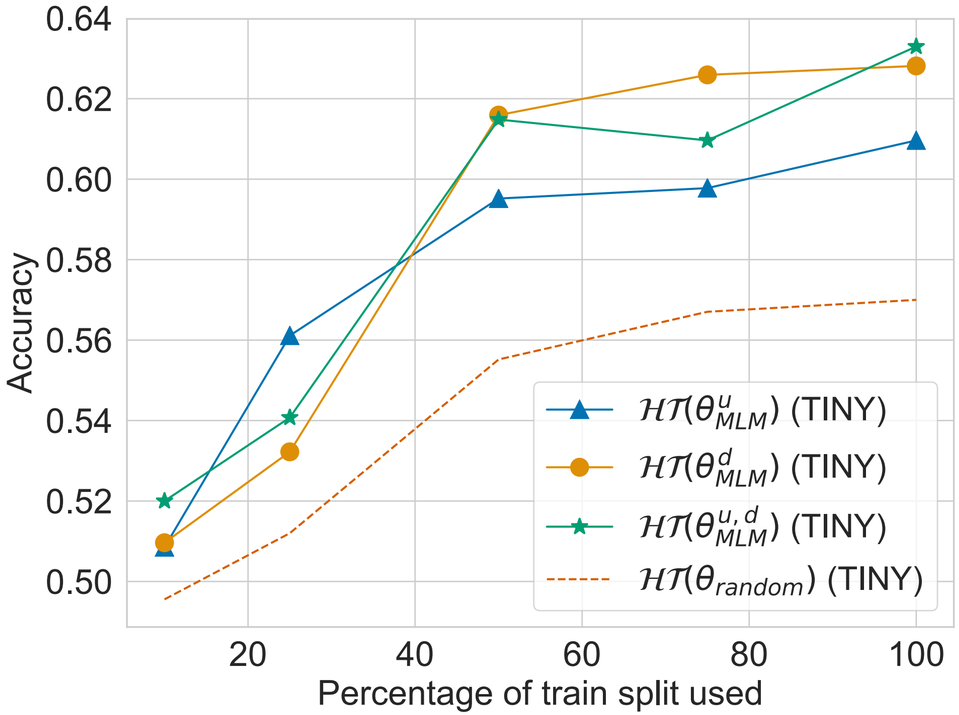
Results I

Decoder Choice

	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

Sequential decoder helps !

Convergence Speed



Pretrained Representations Learn faster!

Conclusion

What is before?

Guiding attention in Sequence-to-sequence models for Dialogue Act prediction

Pierre Colombo*, Emile Chapuis*, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel at AAAI 2020

What is next?

Code-switched inspired losses for generic spoken dialog representations

Emile Chapuis*, Pierre Colombo*, Matthieu Labeau, Chloé Clavel at EMNLP 2021

Thesis of Emile Chapuis: Methods for Spoken Dialogue Understanding

Décembre 2021

Thanks for listening

Title: **Hierarchical Pre-training for Sequence Labelling in SpokenDialog**

Link to Paper

Link to SILICONE

Corresponding Authors:

Emile Chapuis *Pierre Colombo*

Context

Context

Conversational AI or dialogue systems

Context

Conversational AI or dialogue systems

- Natural Language Processing (NLP)
- Linguistics
- Psychology
- Information Retrieval (IR)
- Machine Learning (ML)

Context

Conversational AI or dialogue systems

- Natural Language Processing (NLP)
- Linguistics
- Psychology
- Information Retrieval (IR)
- Machine Learning (ML)

Industrial applications

Context

Conversational AI or dialogue systems

- Natural Language Processing (NLP)
- Linguistics
- Psychology
- Information Retrieval (IR)
- Machine Learning (ML)

Industrial applications

Best available solutions are far from being capable of carrying open-domain conversations

Results II

Results II

Written vs Spoken

Results II

Written vs Spoken

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

Results II

Written vs Spoken

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

Training on Spoken Corpora help!

Results II

Written vs Spoken

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

Size Matters

Training on Spoken Corpora help!

Results II

Written vs Spoken

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

Size Matters

	Emb.	Word	Seq	Total
BERT	5*23	87		110
BERT (4-layer)		43		66
HMLP		8.6	7.8	40
(TINY)		2.9	2.8	28.7
(SMALL)		10.6	10.6	45

Training on Spoken Corpora help!

Results II

Written vs Spoken

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

Size Matters

	Emb.	Word	Seq	Total
BERT	5*23	87		110
BERT (4-layer)		43		66
HMLP		8.6	7.8	40
(TINY)		2.9	2.8	28.7
(SMALL)		10.6	10.6	45

Training on Spoken Corpora help!

Hierarchy factorize parameters:

- Less GPUs
- Faster Models
- Easier to learn

Utterance Level GAP Loss

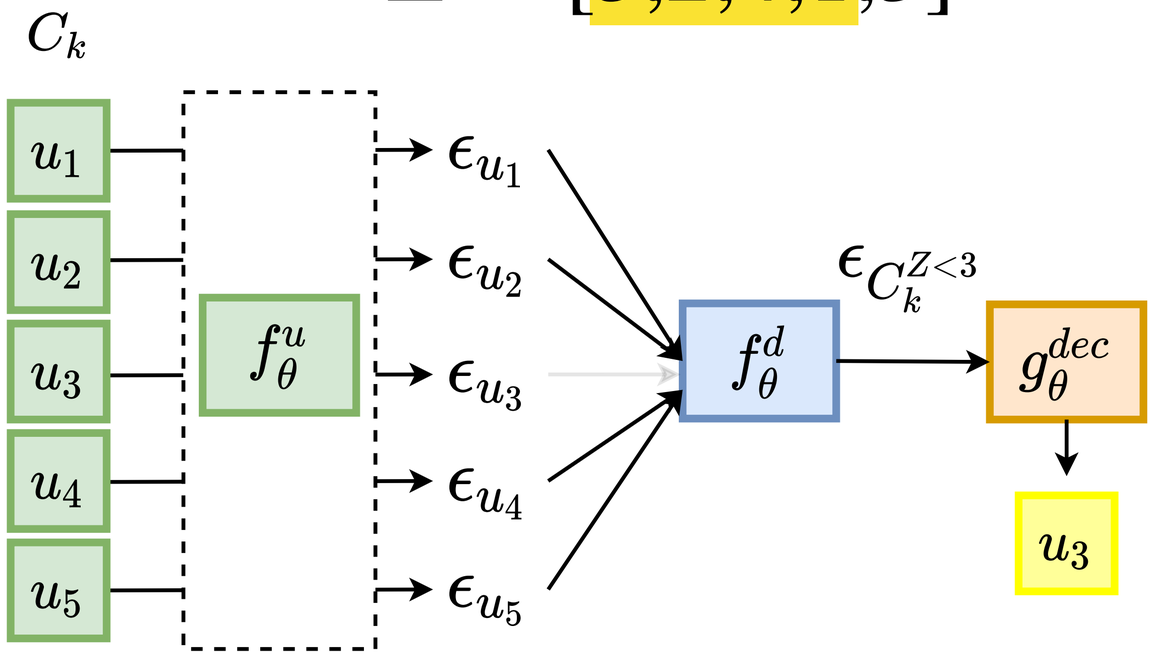
Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Generalized Autoregressive Pre-training (GAP)

$$\mathscr{L}^d_{\textbf{GAP}}(\theta, C_k) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_T}\left[\sum_{t=1}^{|C_k|}\sum_{i=1}^{|u_{z_t}|}\log p_\theta(\omega_i^{z_t} \mid C_k^{\mathbf{z}<t})\right]\right]$$

$$\mathbf{Z} = [5,2,4,1,3]$$



Utterance Level GAP Loss

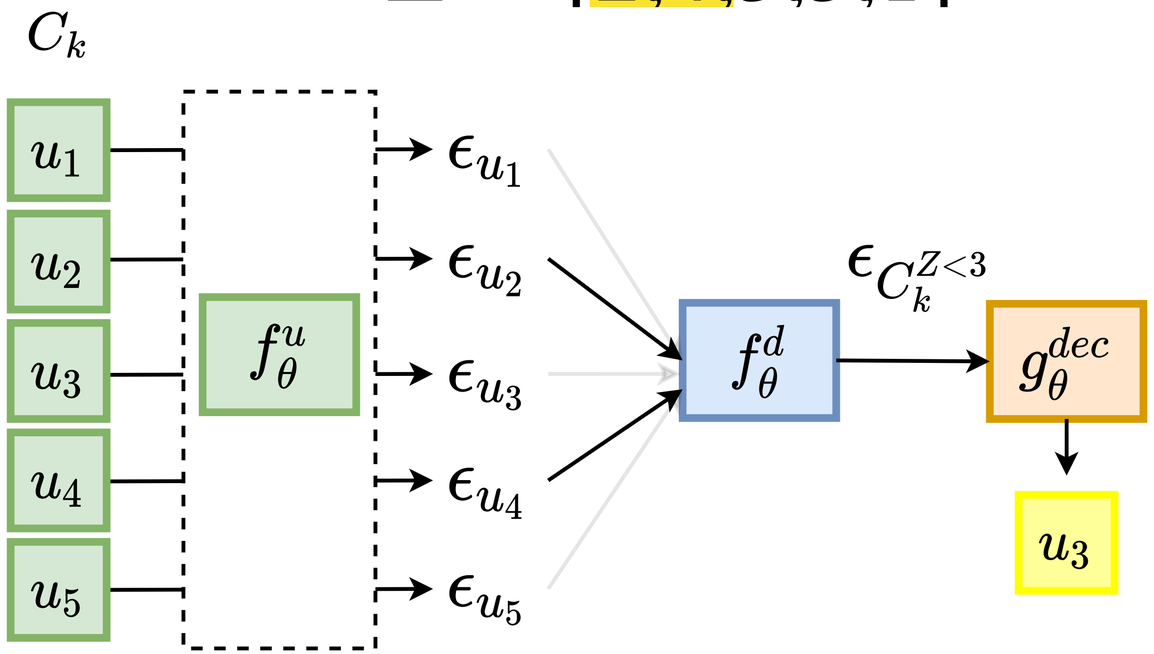
Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{GAP}}(\theta, C_k) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_T}\left[\sum_{t=1}^{|C_k|}\sum_{i=1}^{|u_{z_t}|}\log p_\theta(\omega_i^{z_t} \mid C_k^{\mathbf{z}<t})\right]\right]$$

$$\mathbf{Z} = [2,4,3,5,1]$$



Utterance Level GAP Loss

Dialog Level Pretraining

$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta)\,.$$

Masked Language Model Loss (MLM)

$$\mathscr{L}^d_{\textbf{GAP}}(\theta, C_k) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_T}\left[\sum_{t=1}^{|C_k|}\sum_{i=1}^{|u_{z_t}|}\log p_\theta(\omega_i^{z_t} \mid C_k^{\mathbf{z}<t})\right]\right]$$

$$\mathbf{Z} = [4,3,2,5,1]$$