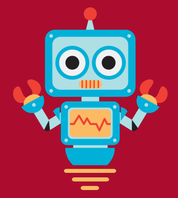


Learning generic dialog embeddings for sequence labelling tasks

Pierre Colombo

pierre.colombo@centralesupelec.fr

Conversational AI

Google Home

Apple Siri

Amazon Alexa

Conversational AI

Google Home

Apple Siri

Amazon Alexa

Definition

A spoken dialog system is a computer agent that interacts with human by understanding and producing spoken language in a coherent way. [Gatt et al., 2018]

Conversational AI

Google Home

Apple Siri

Amazon Alexa

Definition

A spoken dialog system is a computer agent that interacts with human by understanding and producing spoken language in a coherent way. [Gatt et al., 2018]

Dialog vs Written Text

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 .

ω_1^3 ω_2^3 ω_3^3 ω_4^3 ω_5^3 ω_6^3 ω_7^3 ω_8^3 ω_9^3 ω_{10}^3

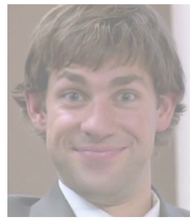
And I never got caught neither!

 u_4

u_5 

Well, you are in prison…

Dialog vs Written Text

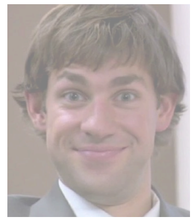
u_1 

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

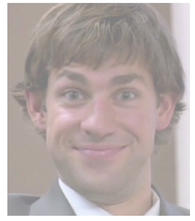
u_3 

That is quite the rap sheet , Prison Mike .

ω_1^3 ω_2^3 ω_3^3 ω_4^3 ω_5^3 ω_6^3 ω_7^3 ω_8^3 ω_9^3 ω_{10}^3

And I never got caught neither!

 u_4

u_5 

Well, you are in prison…

Hierarchy is an important feature in dialog! It is not flat as written text !

Example of Spoken Dialog

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…

Example of Spoken Dialog

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!

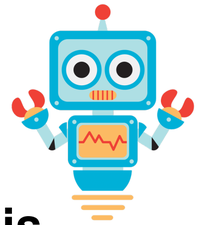
Well, you are in prison…

Dialog can be disfluent, less word diversity, grammar mistakes….

Sequence Labelling Tasks

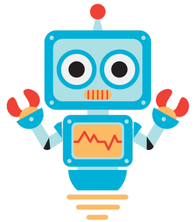
I'm worried about something.

What's that?



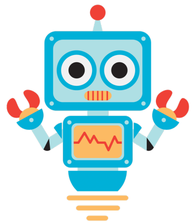
Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.

That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.



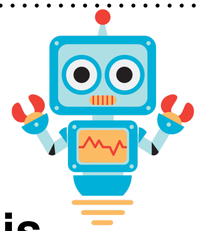
Ok, I'll try that.

Is there anything else bothering you?



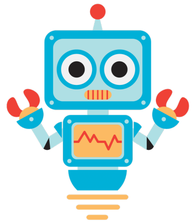
Sequence Labelling Tasks

I'm worried about something. y_1



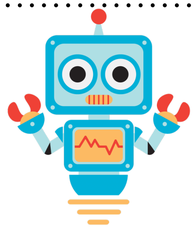
What's that? y_2

Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic. y_3



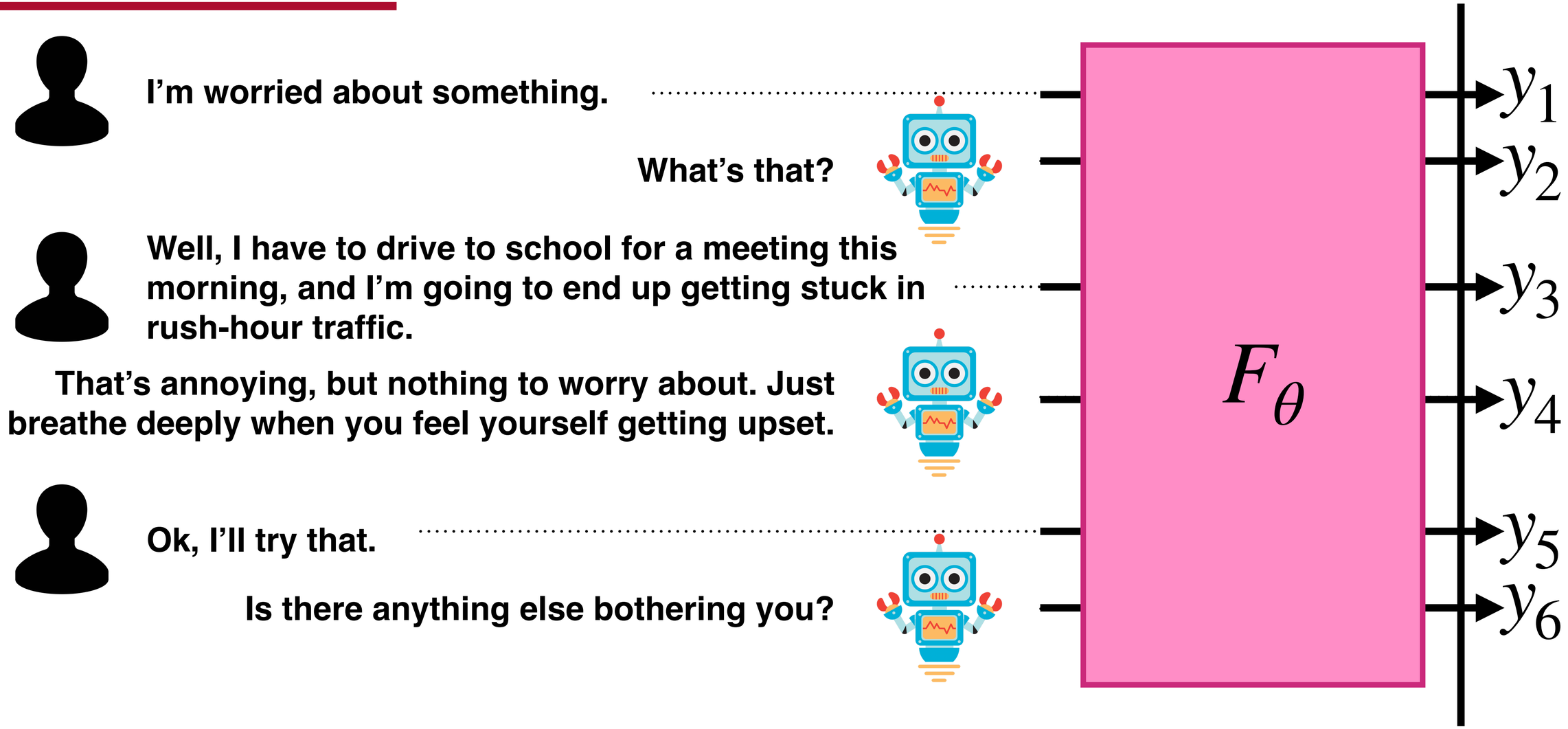
That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset. y_4

Ok, I'll try that. y_5



Is there anything else bothering you? y_6

Sequence Labelling Tasks



Emotions/Sentiments Labels (E/S)

	Emotion	Sentiment
Okay, look at this one. This is my favourite.	Joy	Positive
Oh, that is so sweet !	Joy	Positive
I know ! Phoebe is gonna love dressing them in these !	Joy	Positive
Huh. Except, Phoebe's not gonna be the one that gets to dress them.	Neutral	Neutral
Because she's not gonna get to keep the babies.	Sadness	Negative

Dialog Act Labels (DA)

DAs are semantic labels associated with each utterance in a conversational dialogue that indicate the speaker's intention.

 Um, what did you do this weekend?	Question
 Well, uh, pretty much spent most of my time in the yard.	Statement
 Uh-Huh.	Backchannel
 What do you have planned for your yard?	Question
 Well, we're in the process of, revitalizing it.	Statement

Data Specificities & Constraints

1) Transcripts of spoken dialogues


i think, i think the stand they have, or, or the way the command respect, i, i support that.


i think that is a, a positive thing for them after, um, uh, thousands of years,


they have to, uh, they ha,


i think they in,


when they be, became a country they more than, or, more or less decided they were n't going to take it anymore


and, uh.

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

Switchboard Corpus: 200k labelled utterances  Supervised Learning

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

Switchboard Corpus: 200k labelled utterances → Supervised Learning

SEMAINE: 5.6k labelled utterances → Direct approach may be suboptimal

Data Specificities & Constraints

1) Transcripts of spoken conversations

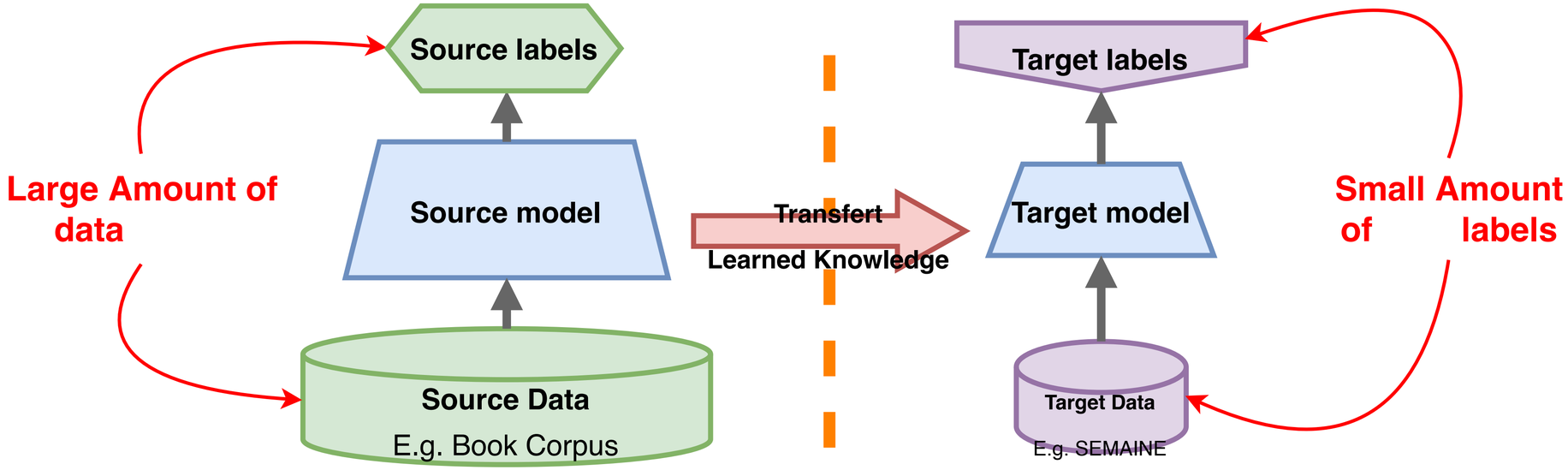
2) Amount of labelled data

Switchboard Corpus: 200k labelled utterances → **Supervised Learning**

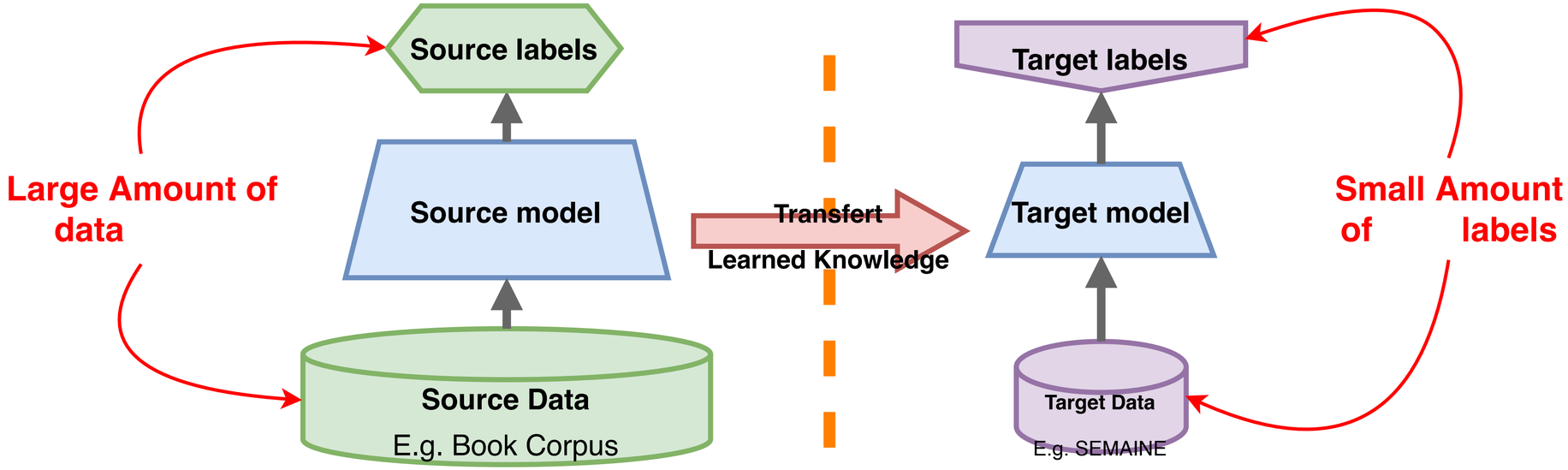
SEMAINE: 5.6k labelled utterances → **Direct approach may be suboptimal**

Large unlabelled corpus → **Self Supervised Learning + Transfert Learning**

Transfert Learning

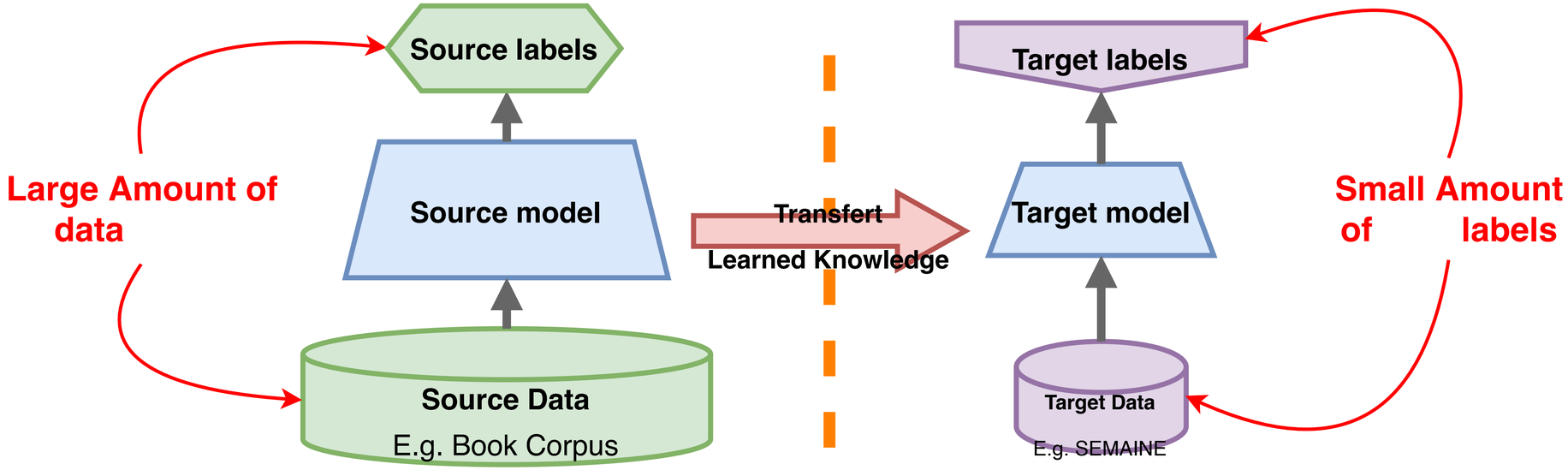


Transfert Learning



Leverage large unlabelled corpus with Self Supervised Learning

Transfert Learning



Leverage large unlabelled corpus with Self Supervised Learning

Pre-trained models

BERT [Devlin and al., 2019]

RoBERTa [Liu and al., 2019]

XLNet [Z Yang et al. 2019]

BART [Lewis and al., 2019]

GPT1 [Radford and Narasimhan, 2018]

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

Monolingual dialogue

Um, what did you do this weekend?

Well, uh, pretty much spent most of my time in the yard.

Uh-Huh.

What do you have planned for your yard?

Well, we’re in the process of, revitalizing it.

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

Monolingual dialogue

Um, what did you do this weekend?

Well, uh, pretty much spent most of my time in the yard.

Uh-Huh.

What do you have planned for your yard?

Well, we're in the process of, revitalizing it.

Code switching

Um, qu'est ce que tu as fait ce weekend dude?

Well, uh, pretty much spent most of my time in the yard.

Uh-Huh.

Qu'as tu prévu pour ton yard?

Well, we're in the process of, revitalizing it.

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

4) Spoken conversations are multimodal

Data Specificities & Constraints

1) Transcripts of spoken conversations
2) Amount of labelled data
3) Language used in the data
4) **Spoken conversations are multimodal**

« What you say »

Verbal

I hate this game!

Data Specificities & Constraints

1) Transcripts of spoken conversations
2) Amount of labelled data
3) Language used in the data
4) **Spoken conversations are multimodal**

« What you say »

Verbal

I hate this game!

Human Communication is multimodal.

Data Specificities & Constraints

1) Transcripts of spoken conversations
2) Amount of labelled data
3) Language used in the data
4) **Spoken conversations are multimodal**

Human Communication is multimodal.

« What you say »

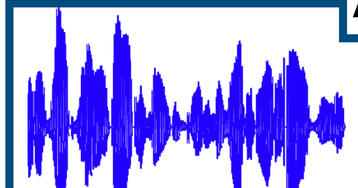
« How you say it »

Verbal

I hate this game!

Visual

Audio



Goals and Approach

Goals and Approach

Input Conversation

What'd you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's

That is quite the rap sheet, Prison Mike.

And I never got caught neither!

Well, you are in prison…

Goals and Approach

Input Conversation


What'd you do, Prison Mike ?

I stole. And I robbed.


And I kidnapped the president's

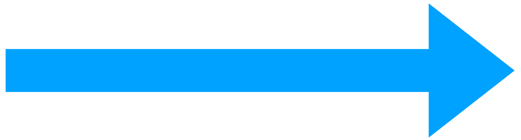

That is quite the rap sheet, Prison Mike.

And I never got caught neither!



Well, you are in prison…

Pretrained Encoder



Multi-purpose embedding

0.9
-0.1
-0.2
1.0
…
0

Goals and Approach

Input Conversation


What'd you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's


That is quite the rap sheet, Prison Mike.

And I never got caught neither!


Well, you are in prison…

Pretrained Encoder

Multi-purpose embedding

0.9
-0.1
-0.2
1.0
…
0

Goal: Learn a multi-purpose dialog embedding

Goals and Approach

Input Conversation


What'd you do, Prison Mike ?

I stole. And I robbed.

And I kidnapped the president's


That is quite the rap sheet, Prison Mike.

And I never got caught neither!


Well, you are in prison…

Pretrained Encoder

Multi-purpose embedding

0.9
-0.1
-0.2
1.0
…
0

Goal: Learn a multi-purpose dialog embedding

Propose new pretaining losses that are better suited for spoken conversations.

How to learning generic dialog embeddings for sequence labelling tasks?

Outline

Outline

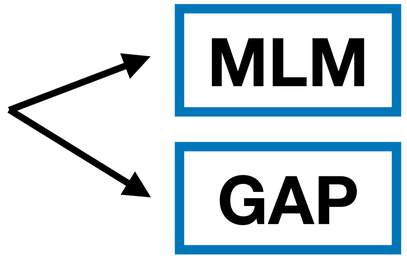
We propose a new model

Outline

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives

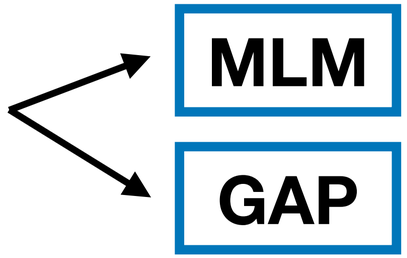


Outline

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives



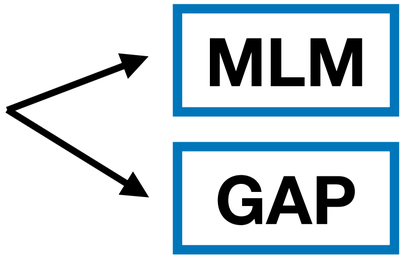
Pretraining Copora

Outline

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives



Pretraining Copora

OpenSubtitles

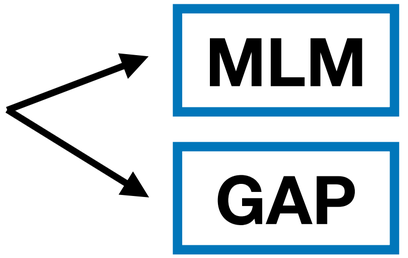
Spoken dialogs | **Large Scale** | **Pretraining**

Outline

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives



Pretraining Copora

OpenSubtitles

Spoken dialogs

Large Scale

Pretraining

Research only consider middle/high size corpora

Outline

We propose a new model

Hierarchical Transformer encoder

New Pretraining Objectives

- MLM
- GAP

Pretraining Copora

OpenSubtitles

Spoken dialogs | **Large Scale** | **Pretraining**

Research only consider middle/high size corpora

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE

- Sizes
- Schema

Models

Goal: **Learn a multi-purpose dialog embedding**

Hierarchical Transformer Encoder: f_θ^u **and** f_θ^d

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

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

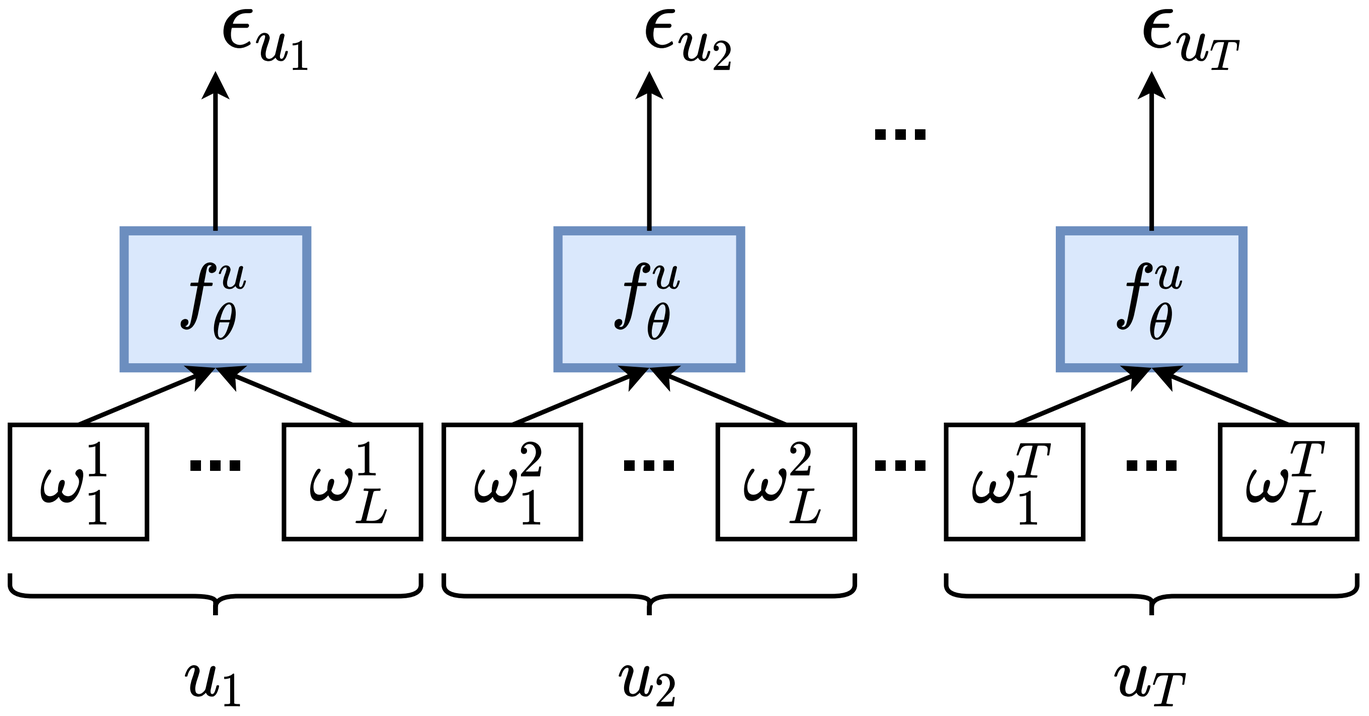
Models

Goal: **Learn a multi-purpose dialog embedding**

Hierarchical Transformer Encoder: f_θ^u **and** f_θ^d

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

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



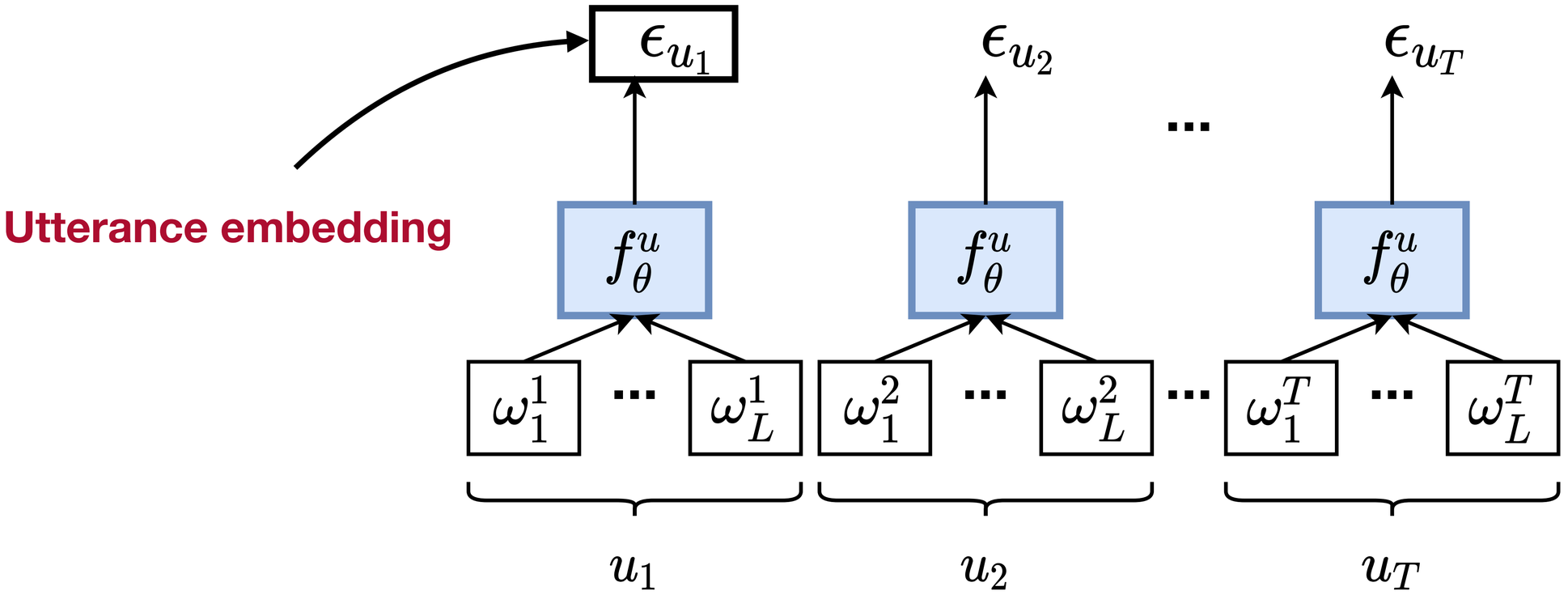
Models

Goal: **Learn a multi-purpose dialog embedding**

Hierarchical Transformer Encoder: f_θ^u **and** f_θ^d

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

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



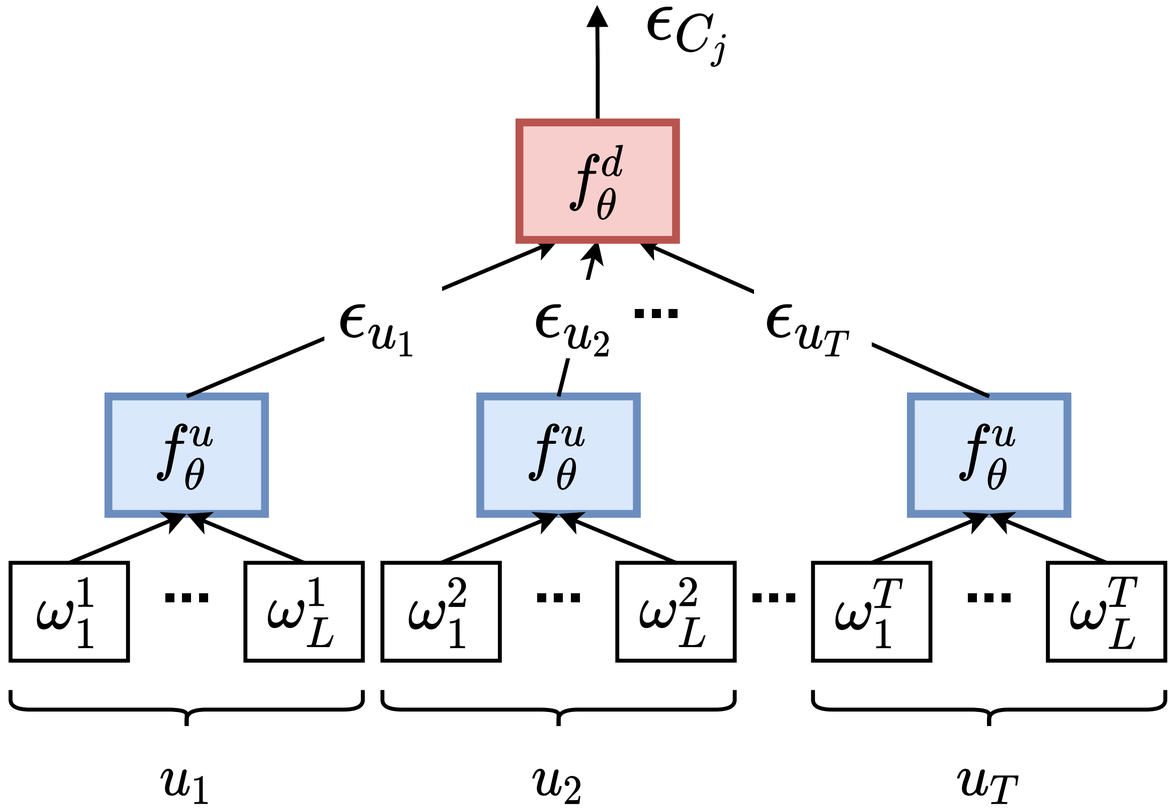
Models

Goal: **Learn a multi-purpose dialog embedding**

Hierarchical Transformer Encoder: f_θ^u **and** f_θ^d

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

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



Models

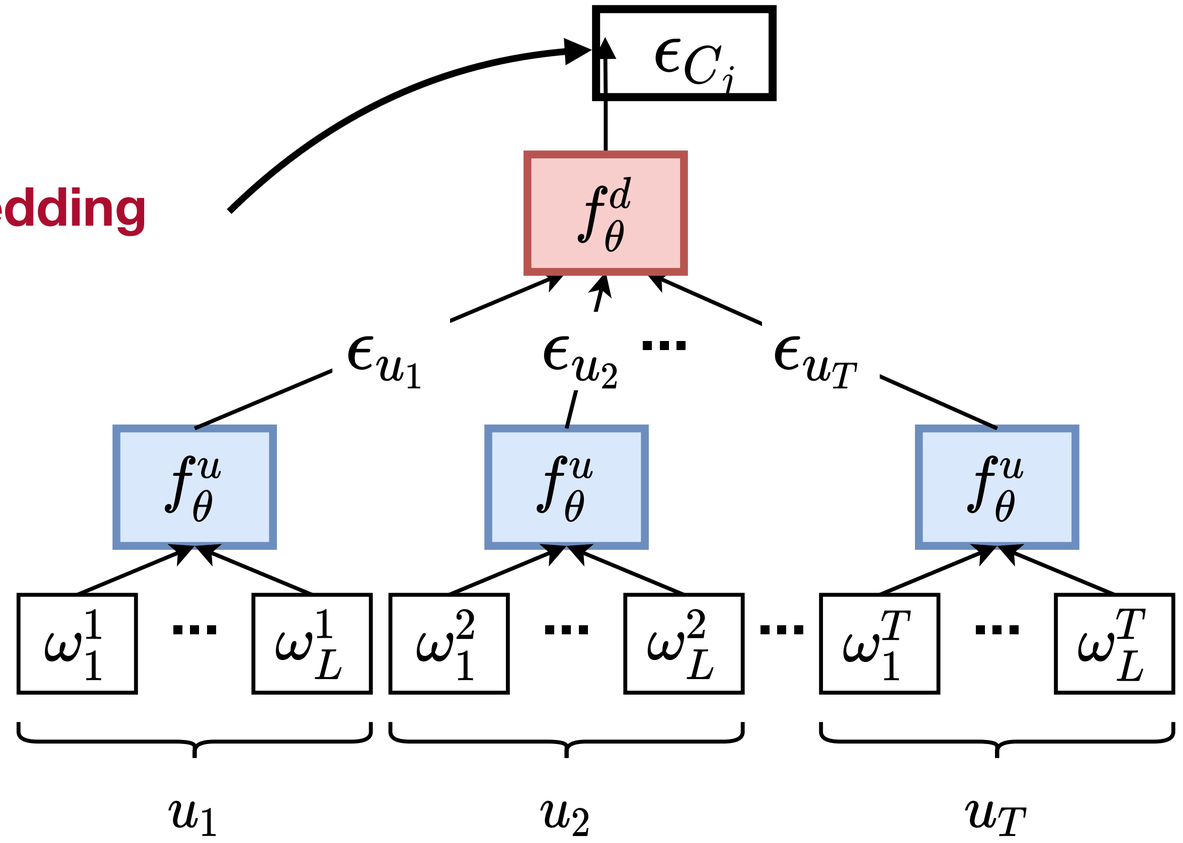
Goal: Learn a multi-purpose dialog embedding

Hierarchical Transformer Encoder: f_θ^u and f_θ^d

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

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

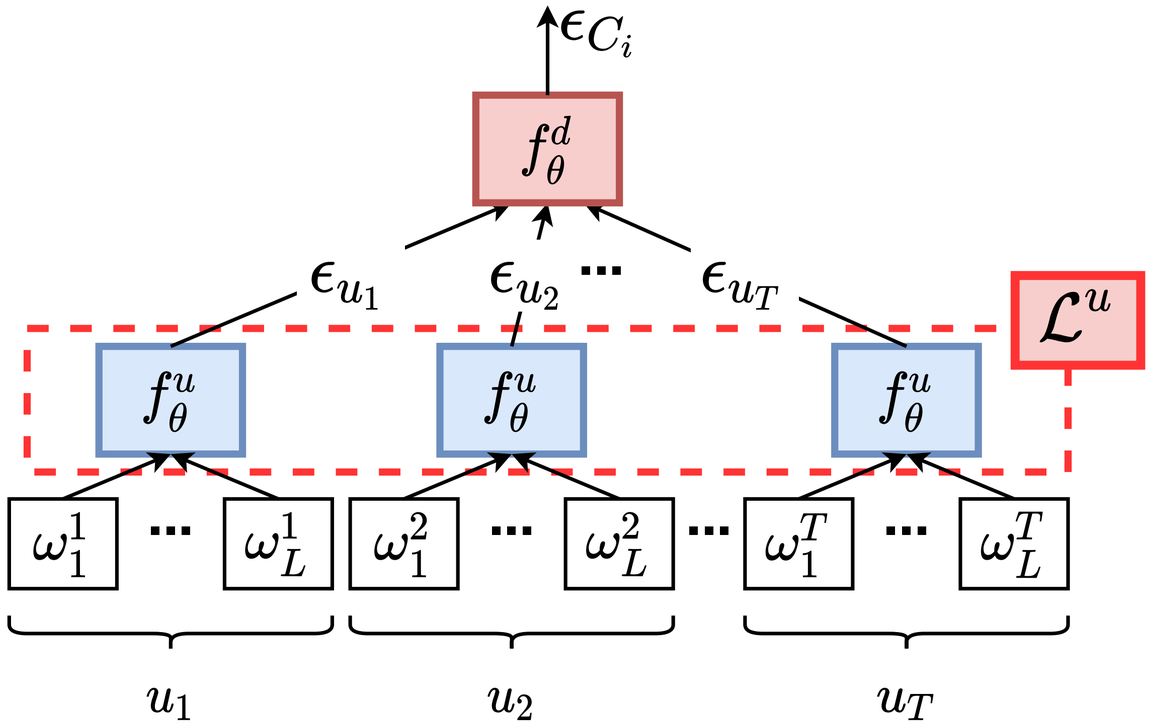
Dialog embedding



Models

Utterance Level Pretraining

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



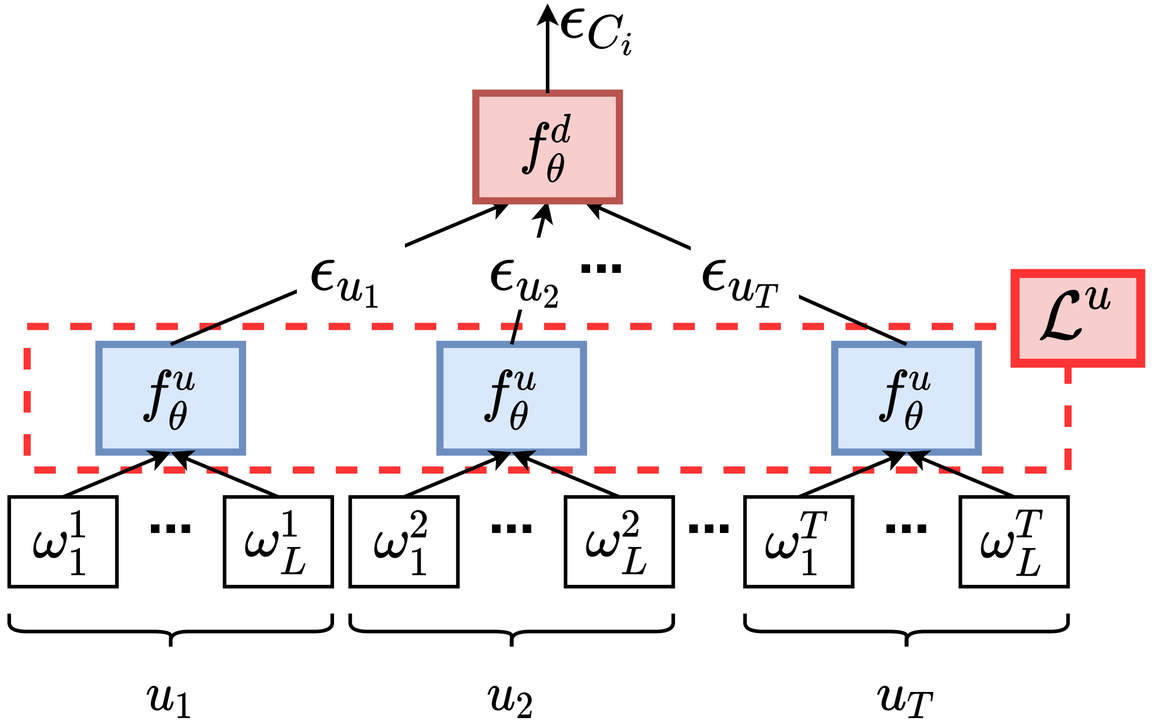
Models

Utterance Level Pretraining

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

Utterance Level Pretraining $\mathscr{L}^u(\theta)$

- **Masked Word Pretraining**



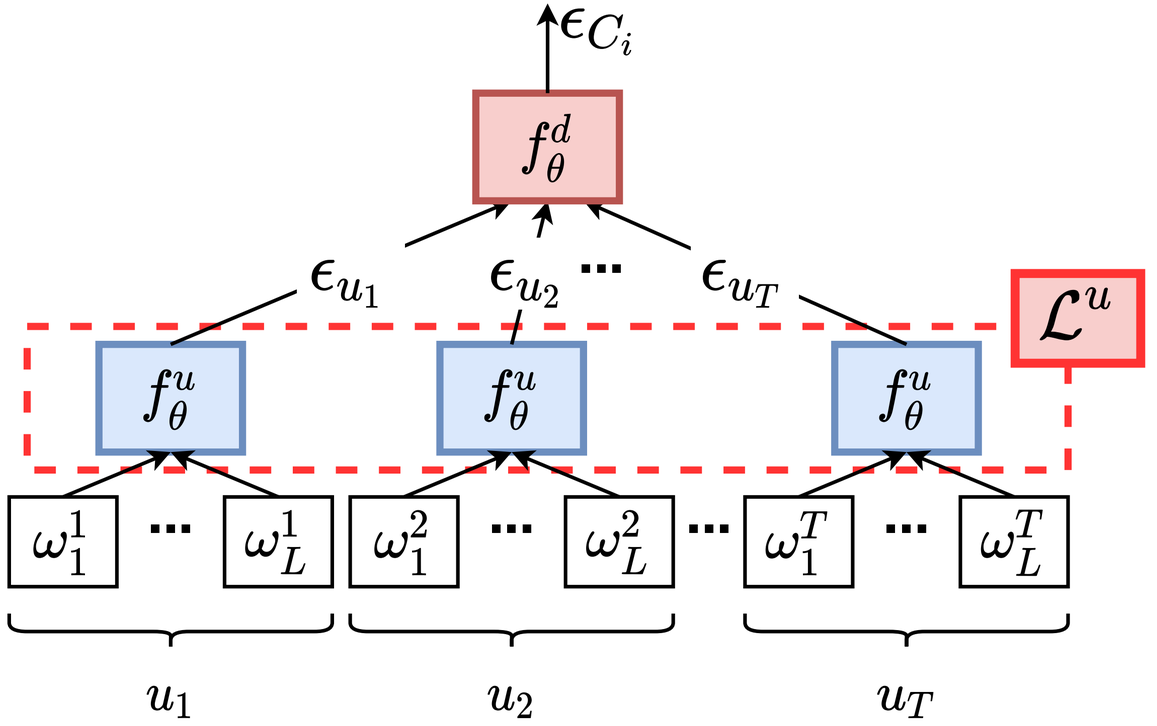
Models

Utterance Level Pretraining

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

Utterance Level Pretraining $\mathscr{L}^u(\theta)$

- Masked Word Pretraining



Models

Utterance Level Pretraining

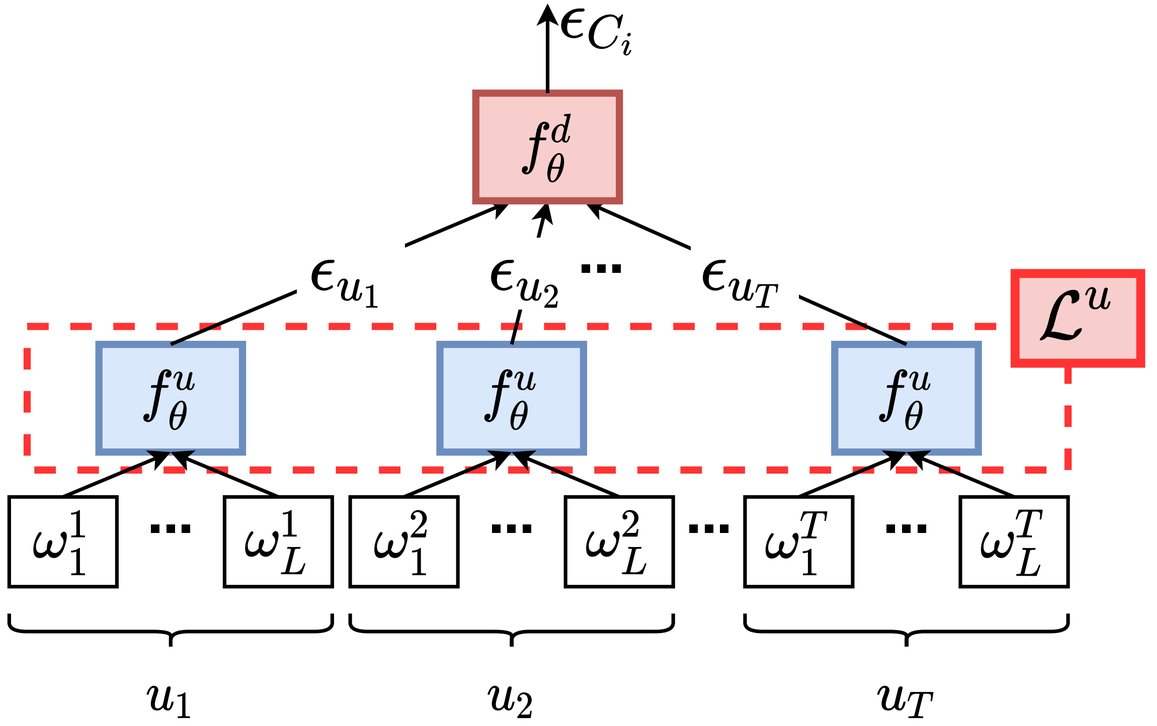
$$\mathcal{L}(\theta) = \lambda_u \mathcal{L}^u(\theta) + \lambda_d \mathcal{L}^d(\theta) \, .$$

Utterance Level Pretraining $\mathcal{L}^u(\theta)$

- Masked Word Pretraining

Goal:

- **Learn the inter-word dependancies**
- **Train the 1st level of the Transformer**



Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

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

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

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

Losses

Utterance Level Pretraining

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

Masked Utterance Modelling (MUM)

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

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

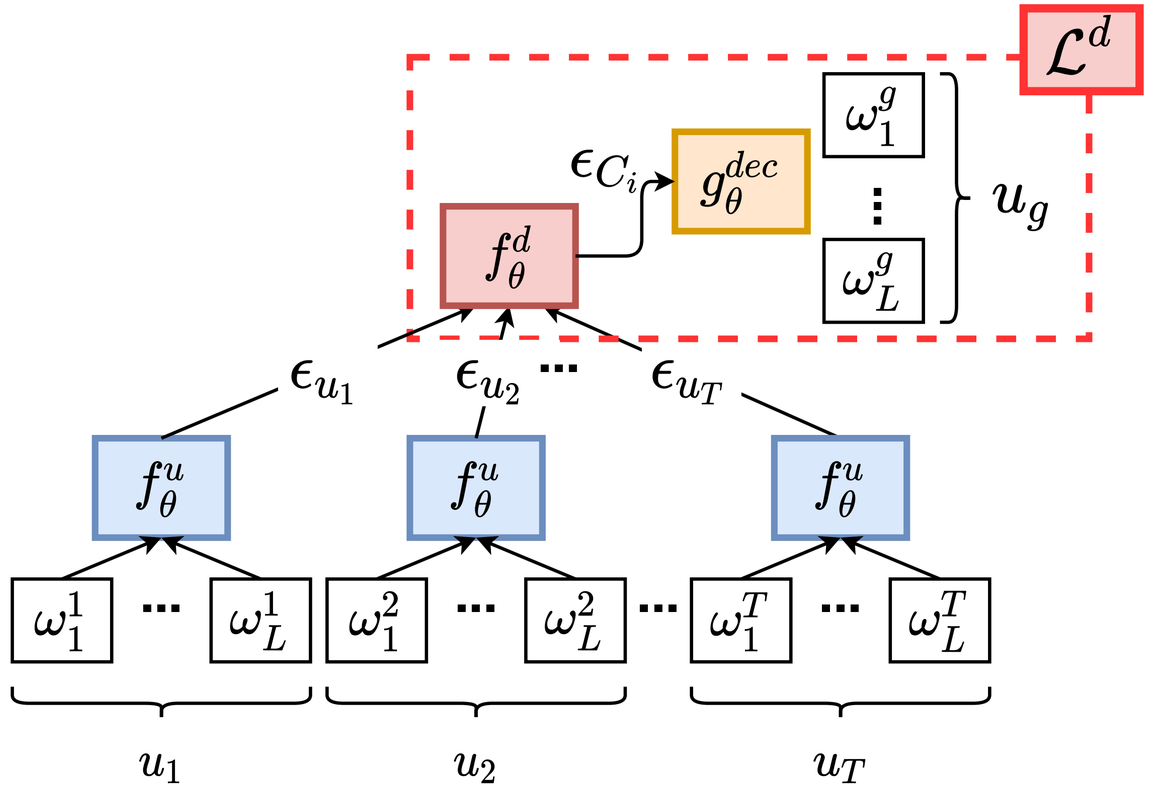
Losses

Dialog Level Pretraining

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

Dialog Level Pretraining: $\mathscr{L}^d(\theta)$

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



Losses

Dialog Level Pretraining

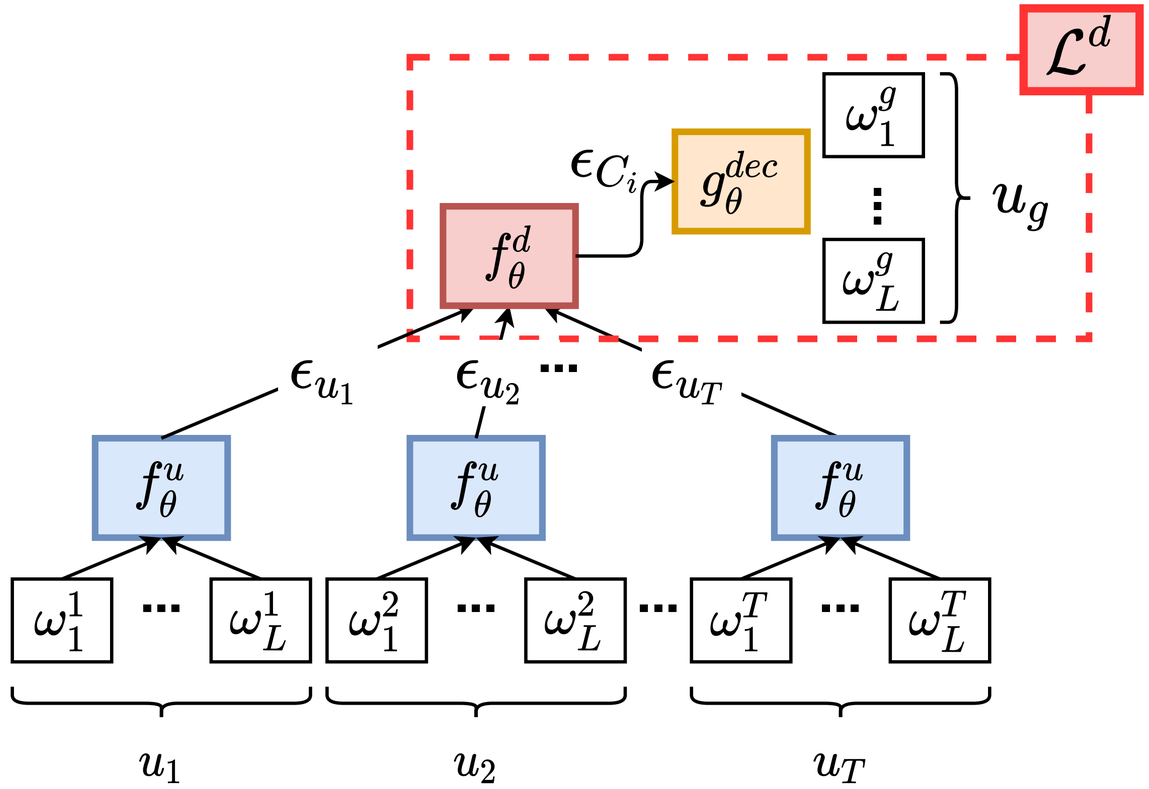
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) \,.$$

Dialog Level Pretraining: $\mathscr{L}^d(\theta)$

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

Goal:

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



Dialog Level Pretraining

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

Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

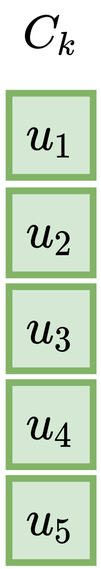
$$p(\mathscr{U} \mid \tilde{C}_k) = \prod_{t \in \mathscr{M}_{\mathscr{U}}} \prod_{j=0}^{|u_t|-1} p_\theta(\omega_j^t \mid \omega_{1:j-1}^t, \tilde{C}_k)$$

Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

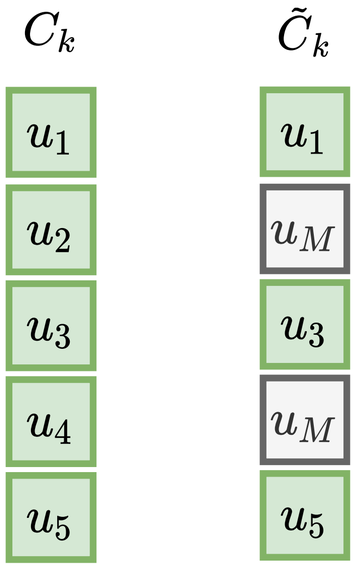


Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

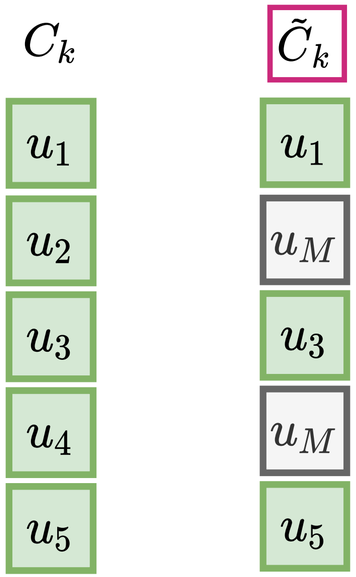


Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

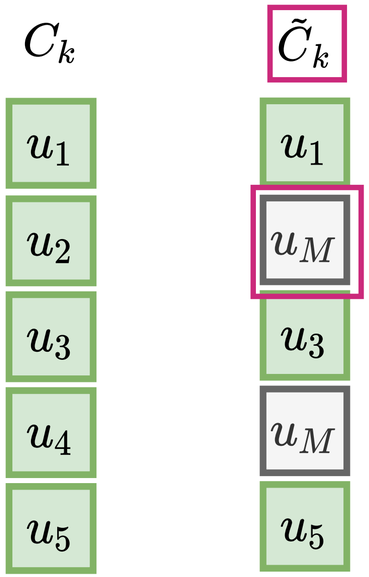


Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

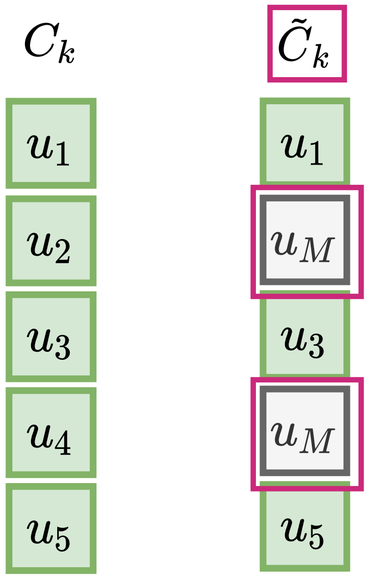


Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

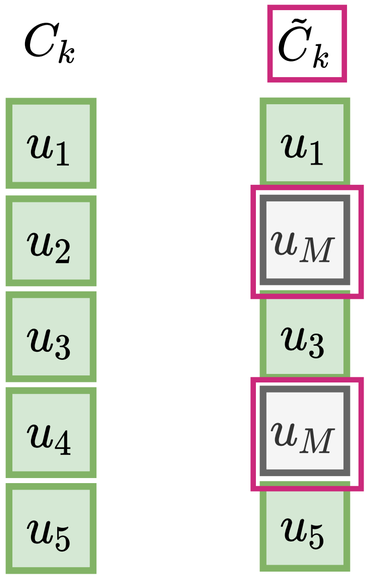


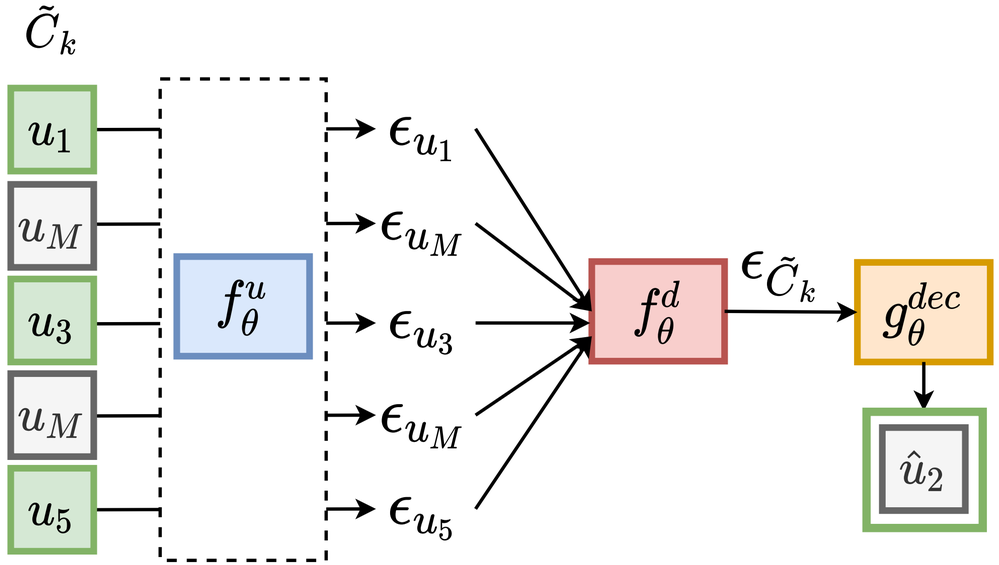
Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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



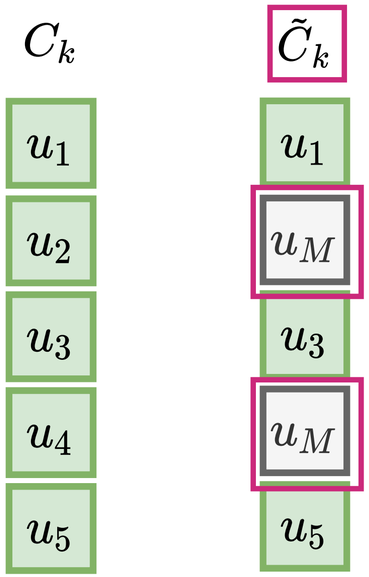


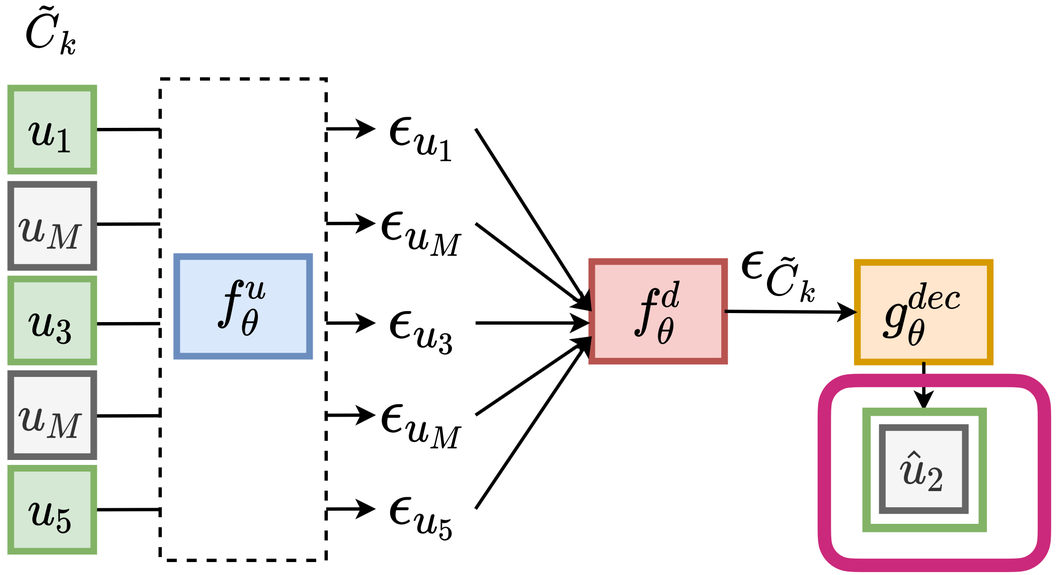
Dialog Level Pretraining

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

Masked Sequence Generation (MSG)

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

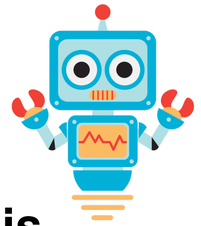




Sequence Labelling Tasks

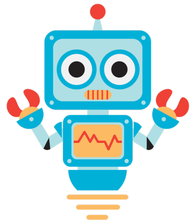
I'm worried about something.

What's that?



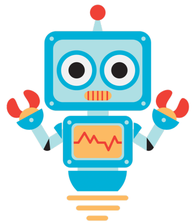
Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.

That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.



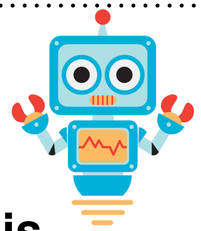
Ok, I'll try that.

Is there anything else bothering you?



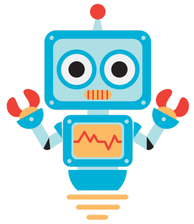
Sequence Labelling Tasks

I'm worried about something. y_1



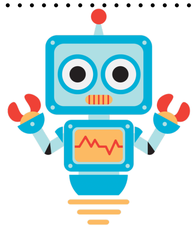
What's that? y_2

Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic. y_3



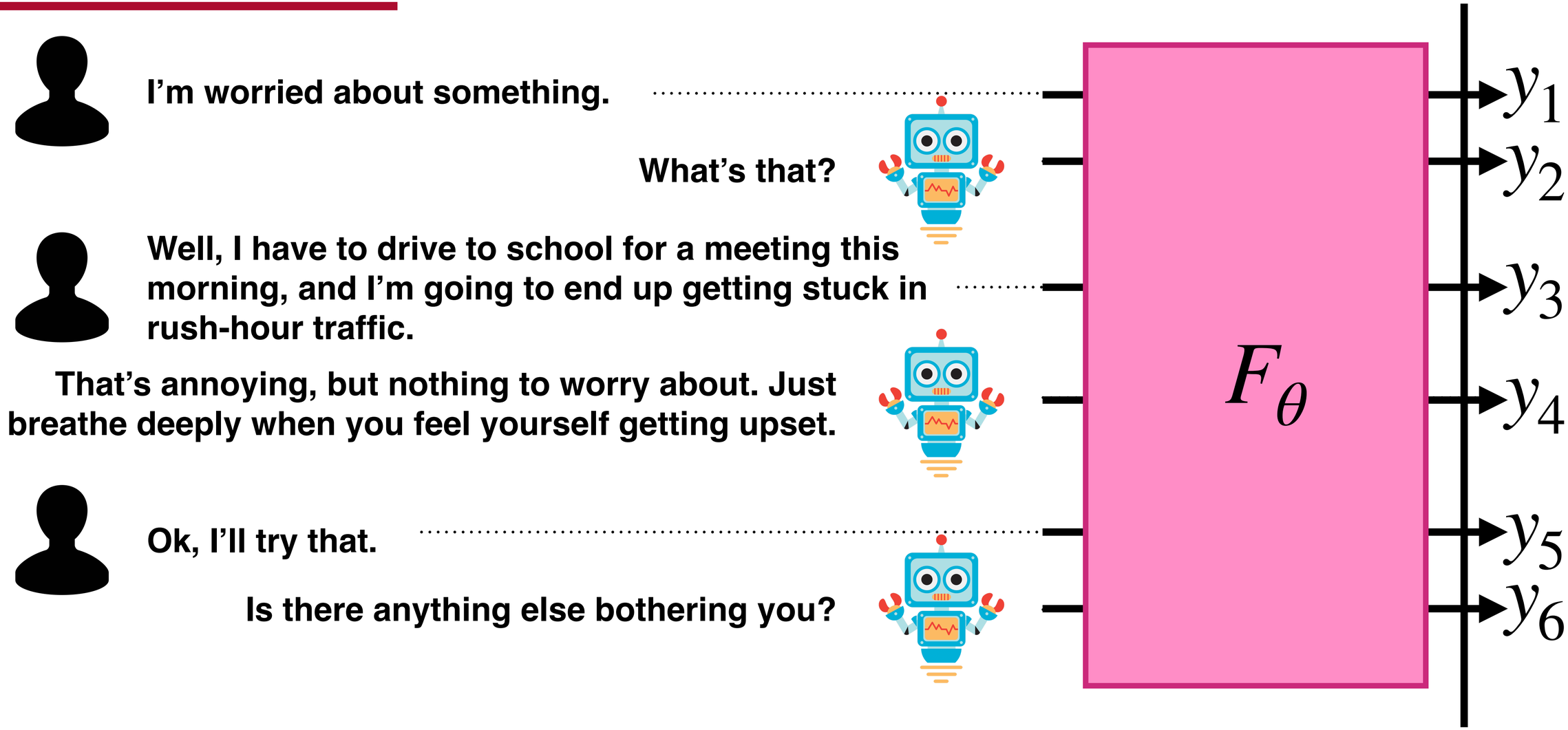
That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset. y_4

Ok, I'll try that. y_5



Is there anything else bothering you? y_6

Sequence Labelling Tasks



Evaluation on SILICONE

Evaluation on SILICONE

Target Task: Sequence Labelling

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE)

Sizes

Schemas

Evaluation on SILICONE

Target Task: Sequence Labelling

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE)

Sizes

Schemas

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

Evaluation on SILICONE

Target Task: Sequence Labelling

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE)

Sizes

Schemas

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

Evaluation on SILICONE

Target Task: Sequence Labelling

SILICONE (Sequence labelllng evaLuatIon benChmark fOr spoken laNguagE)

Sizes

Schemas

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

Results on SILICONE

Results on SILICONE

Pretraining Corpora	OpenSubtitles	Lison and Tiedemann, 2016	Spoken dialogs	> 2.3 billion

Results on SILICONE

Pretraining Corpora OpenSubtitles Lison and Tiedemann, 2016 Spoken dialogs > 2.3 billion

	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**

Results on SILICONE

Pretraining Corpora OpenSubtitles Lison and Tiedemann, 2016 Spoken dialogs > 2.3 billion

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

The proposed hierarchical pretraining helps.

Results on SILICONE

Pretraining Corpora OpenSubtitles Lison and Tiedemann, 2016 Spoken dialogs > 2.3 billion

	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**

The proposed hierarchical pretraining helps.

Reduced model size thanks to hierarchy.

Limitations & Futur Work

Limitations & Futur Work

Limitations

The proposed model only handle monolingual English dialogues

The proposed model is mono-modal

Limitations & Futur Work

Limitations

The proposed model only handle monolingual English dialogues

The proposed model is mono-modal

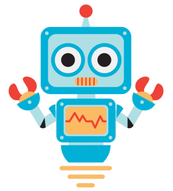
Extensions

Most of people in the world are bilingual [Grosjean and Li, 2013]

When speakers share more than one language, they inevitably will engage in code-switching

Multimodal extension

Contributions



For this talk.

- E.Chapuis*, **P.Colombo***,M. Labeau,and C.Clavel. Code-switched inspired losses for generic spoken dialog representations. **EMNLP 2021**.
- **P. Colombo**, E. Chapuis, M. Labeau, and C. Clavel. Improving multimodal fusion via mutual dependency maximisation. **(oral) EMNLP 2021**.
- E.Chapuis*, **P.Colombo***, M. Manica, M.Labeau, and C.Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. **Findings of EMNLP 2020.**
- **P.Colombo***, E.Chapuis*, M.Manica, E.Vignon, G.Varni, and C.Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. **(oral) AAAI 2020.**
- A. Garcia*, **P.Colombo***, S. Essid, F. D'Alché But, C. Clavel . From the token to the review: A hierarchical multimodal approach to opinion mining. **EMNLP 2019.**
- T. Dinkar*, **P. Colombo***, M. Labeau, and C. Clavel. The importance of fillers for text representations of speech transcripts. **EMNLP 2020**

On representation learning

- **P. Colombo**, P. Piantanida, and C. Clavel. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations **(oral) ACL 2021**
- H. Jalalzai*, **P. Colombo** *, C. Clavel, E. Gaussier, G. Varni, E. Vignon, and A. Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. **NeurIPS 2020**
- **P. Colombo**, N. Noiry, et al. Learning Disentangled Textual Representations via Statistical Measures of Similarity. **ACL 2022**.
- Georg Pichler*, **P. Colombo***, Malik Boudiaf*, Günther Koliander, Pablo Piantanida. KNIFE: Kernelized-Neural Differential Entropy Estimation.
- **P. Colombo**, N. Noiry et al. SOLDIER: Learning to Disentangle Textual Representations via Supervised Contrastive Learning

This presentation is the result of a team effort

Chouchang Yack, Giovanna Varni, Chloé Clavel, Emile Chapuis, Matthieu Labeau, Guillaume Staerman, Pablo Piantanida, Tanvi Dinkar, Hamid Jalalzai, Matteo Manica, Eric Gaussier, Emmanuel Vignon, Anne Sabourin, Alexandre Garcia, Slim Essid, Florence D'Alché-Buc, Wojciech Witon, Ashutosh Modi, James Kennedy, Mubbasir Kapadia, Georg Pichler, Malik Boudiaf, Günther Koliander, Nathan Noiry, Pavlo Mozharovskyi, Stephan Clémençon

Thanks for listening

Title: **Learning generic dialog embeddings for sequence labelling tasks**

Corresponding Authors:

Pierre Colombo

Emile Chapuis

Code Switching

Intra-sentential

What have you done, Prison Mike ?

I stole et j’ai volé.

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

C’est quite le rap sheet, Prison Mike.

And they never caught me either

Well, t’es in prison…

Code Switching

Inter-sentential

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.

That is quite the rap sheet, Prison Mike.

Et on ne m’a jamais chopé non plus!

Well, you are in prison…

Formalisation

Conversation C

$u_1^{L_1}$ 

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.

 $u_2^{L_2}$

$u_3^{L_2}$ 

That is quite the rap sheet, Prison Mike.

Et on ne m’a jamais chopé non plus!

 $u_4^{L_1}$

$u_5^{L_2}$ 

Well, you are in prison…

$$C_i = (u_1^{L_1}, u_2^{L_2}, \ldots, u_{|C_i|}^{L_{C_i}})$$

Summary & Conclusions

Summary & Conclusions

Dialogue act classification

Emotions/sentiments classification

Summary & Conclusions

Sequence labelling tasks

Dialogue act classification

Emotions/sentiments classification

Summary & Conclusions

Sequence labelling tasks

Dialogue act classification

Emotions/sentiments classification

English monolingual data

Summary & Conclusions

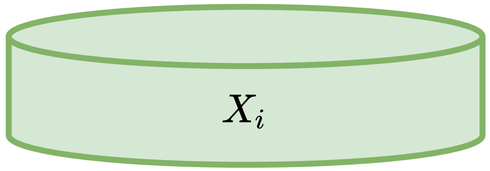
Sequence labelling tasks

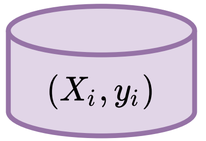
Dialogue act classification

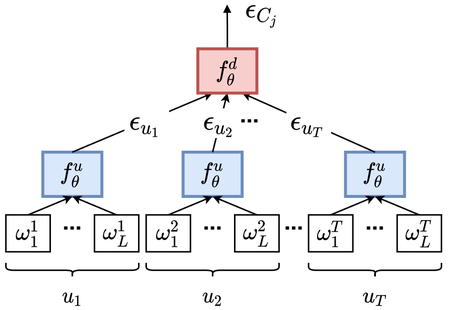
Emotions/sentiments classification

English monolingual data

Small annotated corpus







Summary & Conclusions

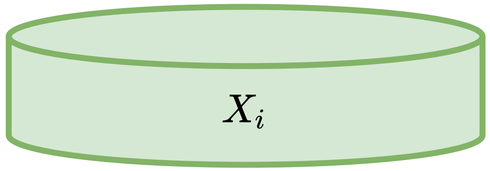
Sequence labelling tasks

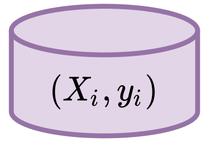
Dialogue act classification

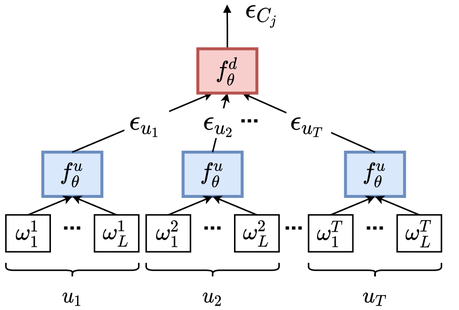
Emotions/sentiments classification

English monolingual data

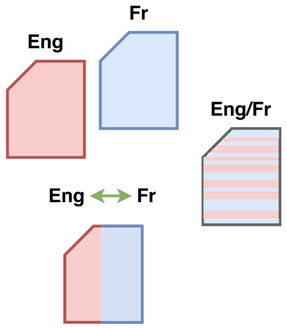
Small annotated corpus

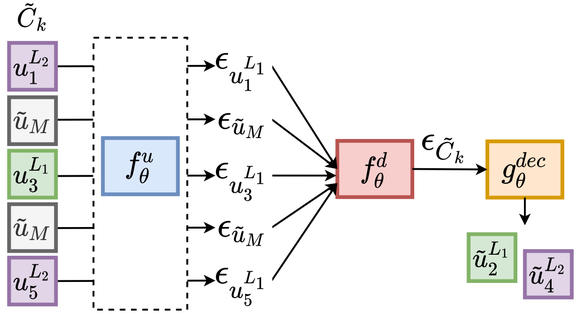






Multilingual data





Perspectives

Perspectives

Many Future works can be drawn from the works I presented

Perspectives

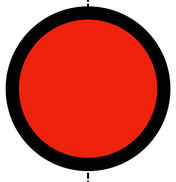
Many Future works can be drawn from the works I presented

Extentions

Explore the interplay between dialogue act and emotion

Including speaker information in our models

Enrich our multilignual models developed with new languages



Perspectives

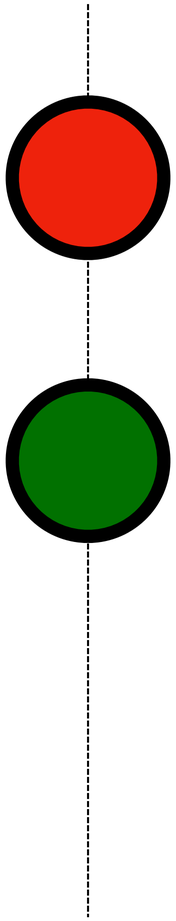
Many Future works can be drawn from the works I presented

Extentions

Explore the interplay between dialogue act and emotion

Including speaker information in our models

Enrich our multilignual models developed with new languages



Generation

Generation of emotional driven responses

Perspectives

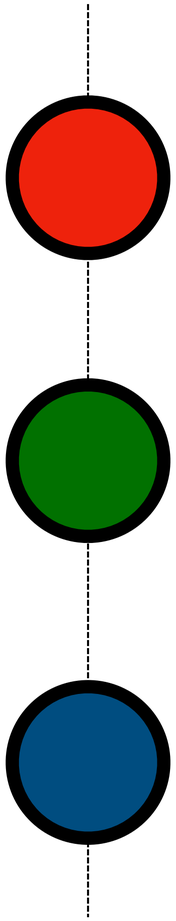
Many Future works can be drawn from the works I presented

Extentions

Explore the interplay between dialogue act and emotion

Including speaker information in our models

Enrich our multilignual models developed with new languages



Generation

Generation of emotional driven responses

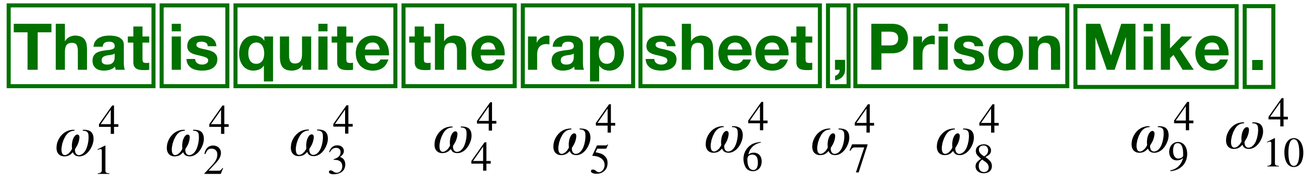
Evaluation

Dialogue Evaluation

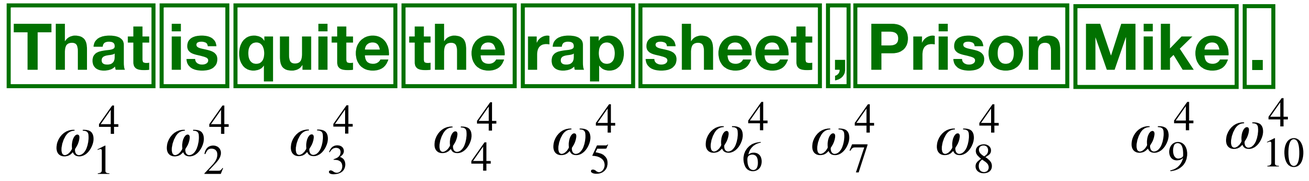
Studying Domain Adaptation

Pretrained Language Models

Pretrained Language Models

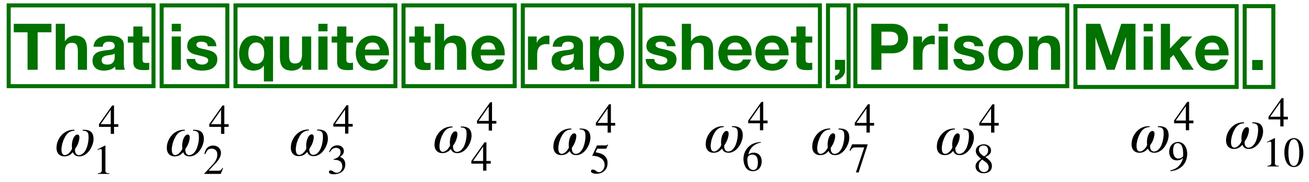


Pretrained Language Models

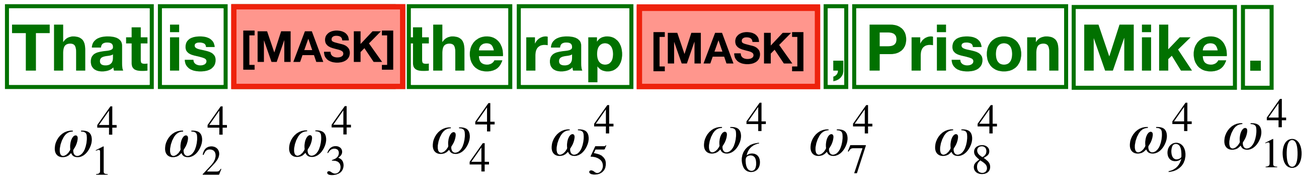


Language modelling:

Pretrained Language Models



Language modelling:



Pretrained Language Models

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

Language modelling:

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

quite sheet

Pretrained Language Models

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

Language modelling:

Classification

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

quite sheet

Pretrained Language Models

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

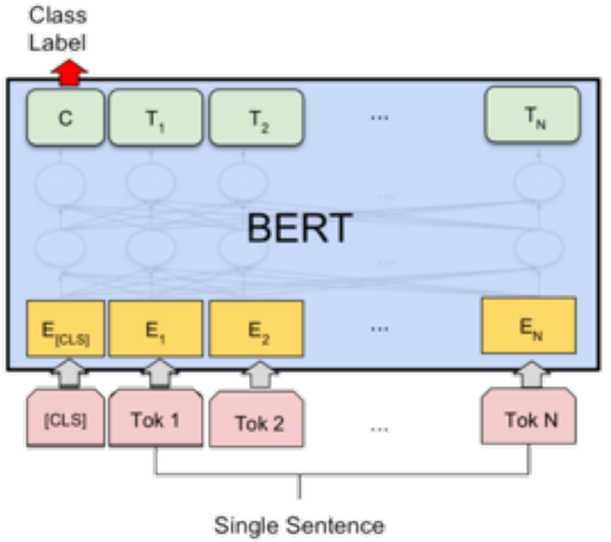
Language modelling:

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

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Pretrained Language Models

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

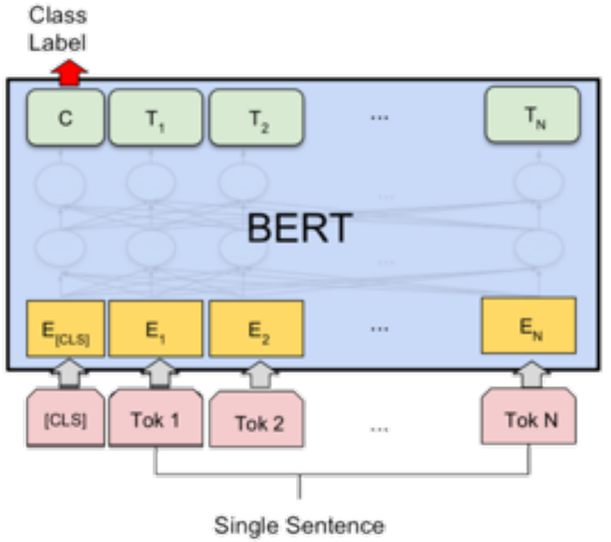
Language modelling:

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

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Predict → Label

Pretrained Language Models

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

Language modelling:

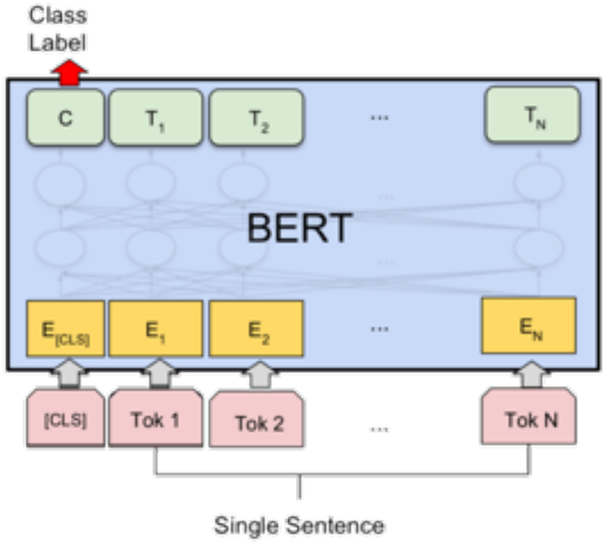
That is [MASK] the rap [MASK] , Prison Mike .

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

Finetuning

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

→ Predict: Label

Pretrained Language Models

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

Language modelling:

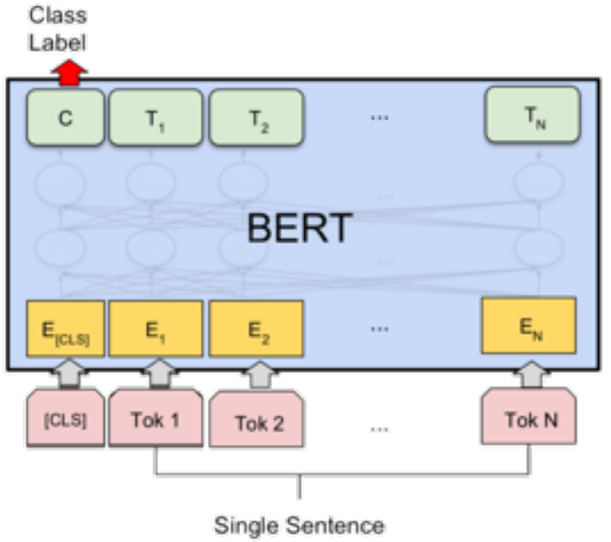
That is [MASK] the rap [MASK] , Prison Mike .

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

Finetuning

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Predict → Label

Pretrained Language Models

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

Language modelling:

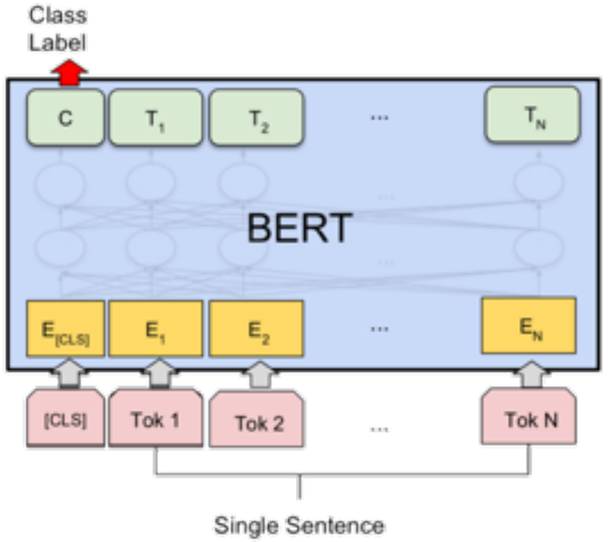
That is [MASK] the rap [MASK] , Prison Mike .

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

Finetuning

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Predict → Label

Pretrained Language Models

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

Language modelling:

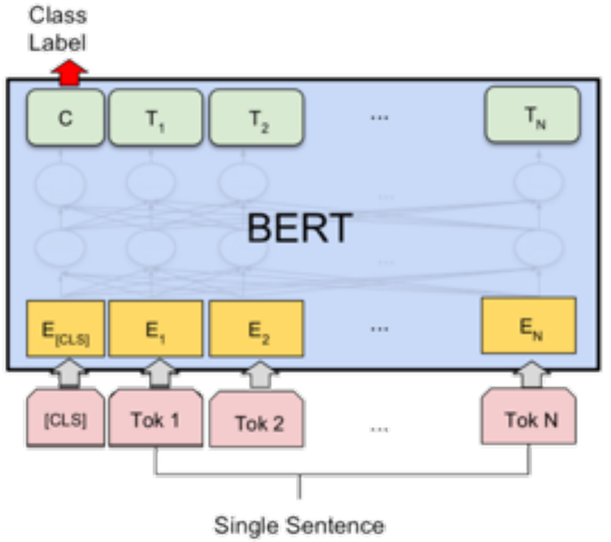
That is [MASK] the rap [MASK] , Prison Mike .

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

Predict → quite sheet

Finetuning

Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Predict → Label

Pre-training Corpus

Pre-training Corpus

Wikipedia

Pre-training Corpus

Wikipedia

BookCorpus

Pre-training Corpus

Wikipedia

SwitchBoard

BookCorpus

Pre-training Corpus

Wikipedia

Cornell Movie Dialog Corpus

SwitchBoard

BookCorpus

Pre-training Corpus

Wikipedia

Cornell Movie Dialog Corpus

SwitchBoard

BookCorpus

Wizard of Oz

Pre-training Corpus

Wikipedia

Cornell Movie Dialog Corpus

SwitchBoard

BookCorpus

Wizard of Oz

Reddit Corpus

Pre-training Corpus

Wikipedia

Cornell Movie Dialog Corpus

SwitchBoard

BookCorpus

Wizard of Oz

Reddit Corpus

Twitter

Pre-training Corpus

Wikipedia

Cornell Movie
Dialog Corpus

SwitchBoard

Opensubtitles

Wizard of Oz

BookCorpus

Reddit Corpus

Twitter

Pre-training Corpus

Wikipedia

Cornell Movie Dialog Corpus

SwitchBoard

Opensubtitles

Wizard of Oz

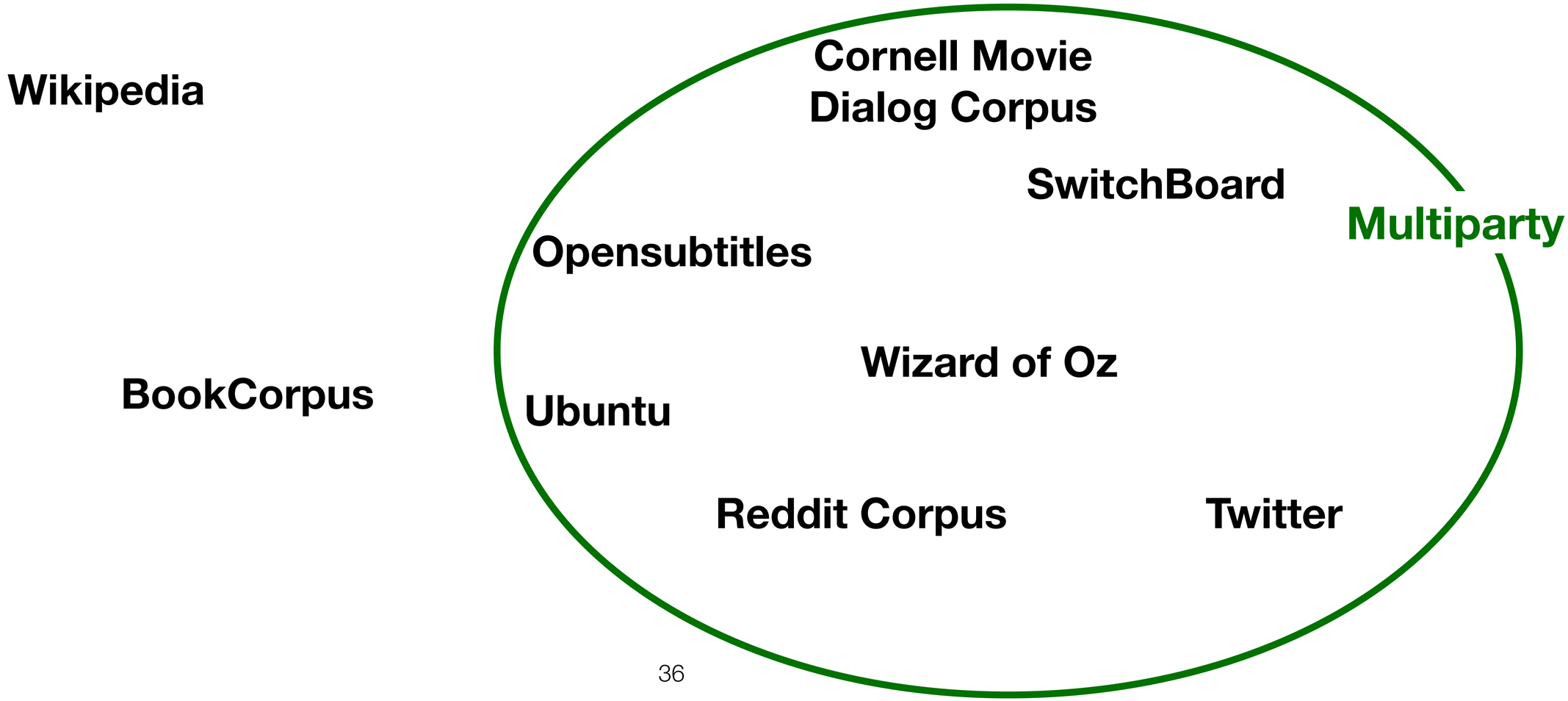
BookCorpus

Ubuntu

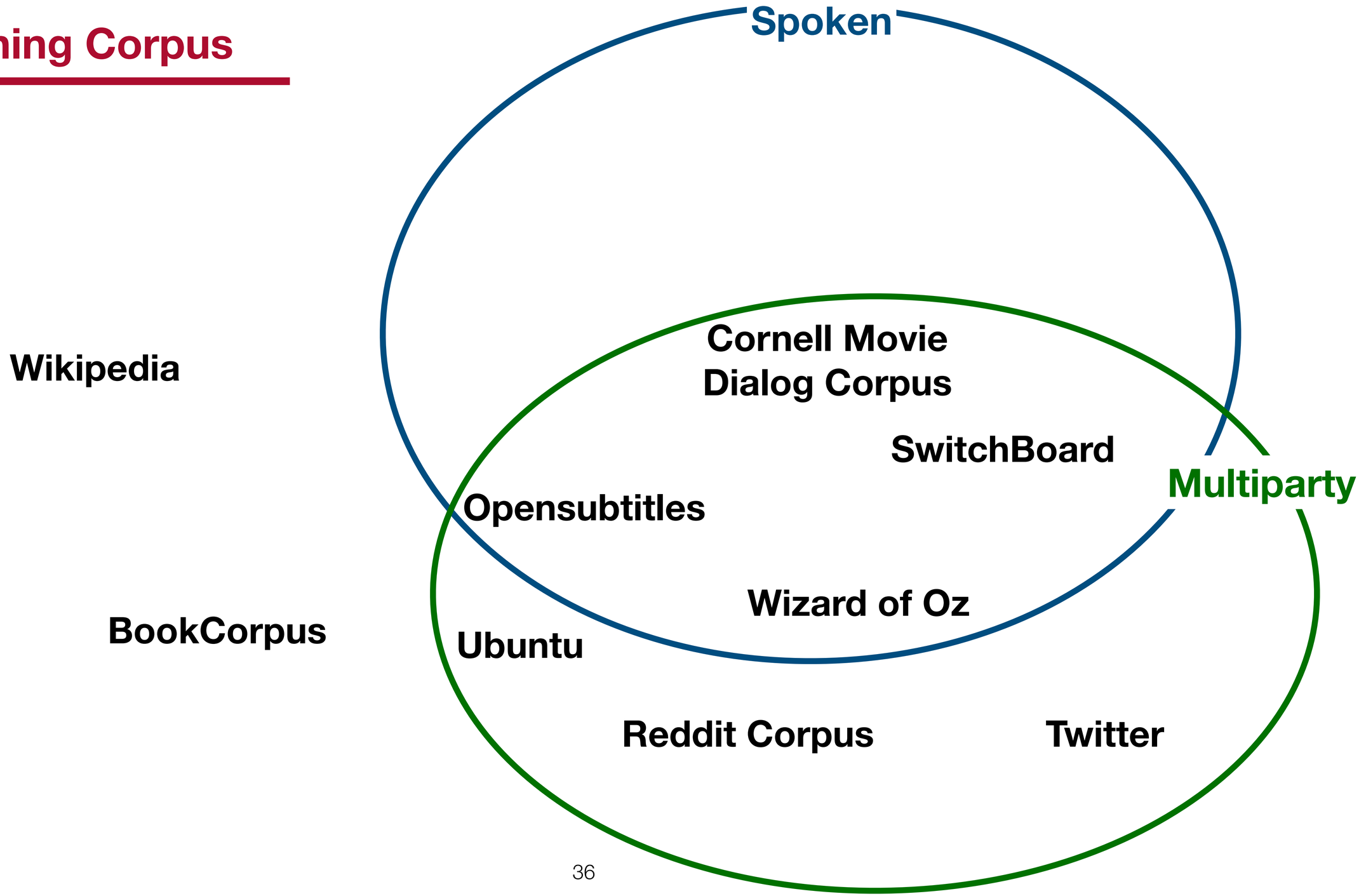
Reddit Corpus

Twitter

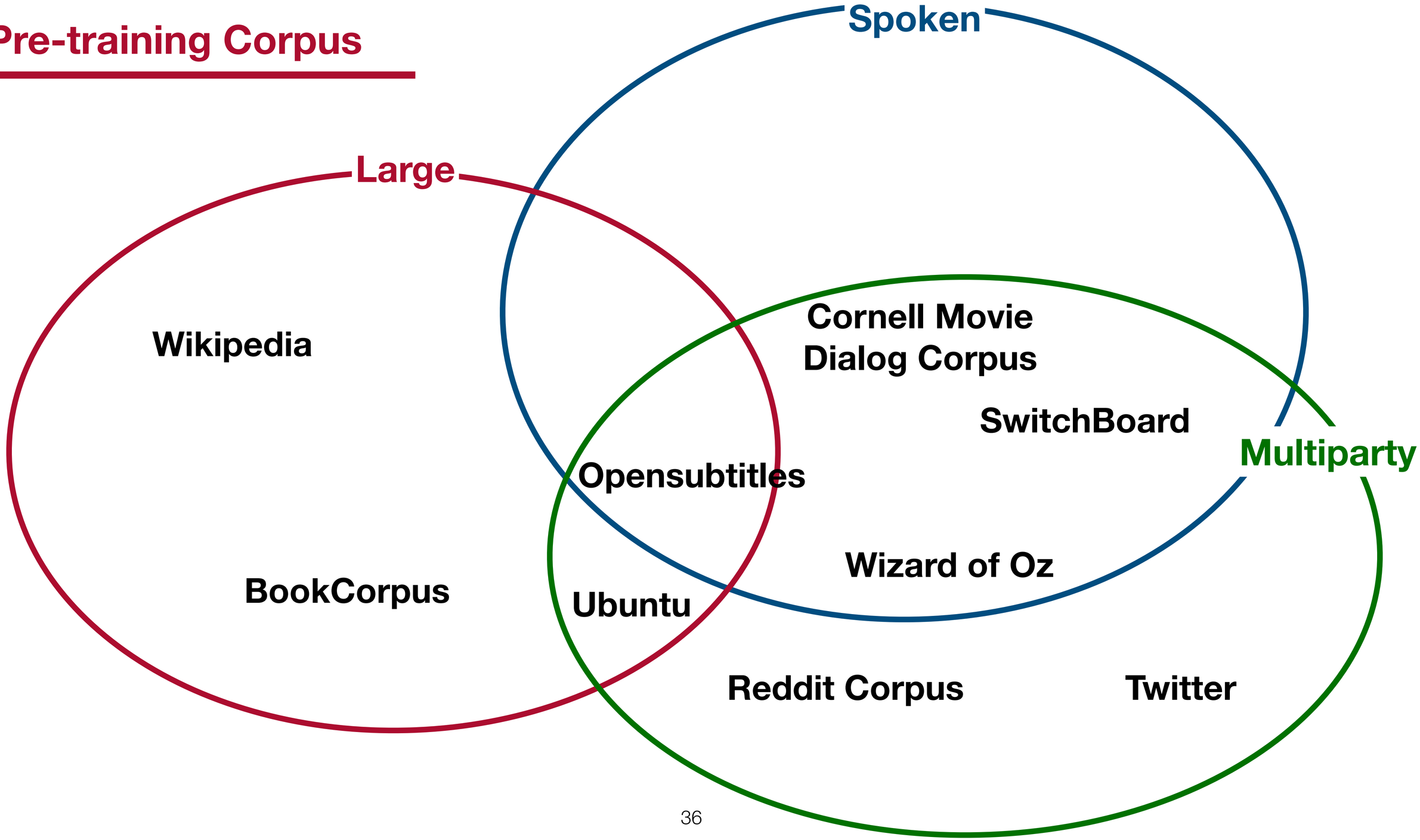
Pre-training Corpus



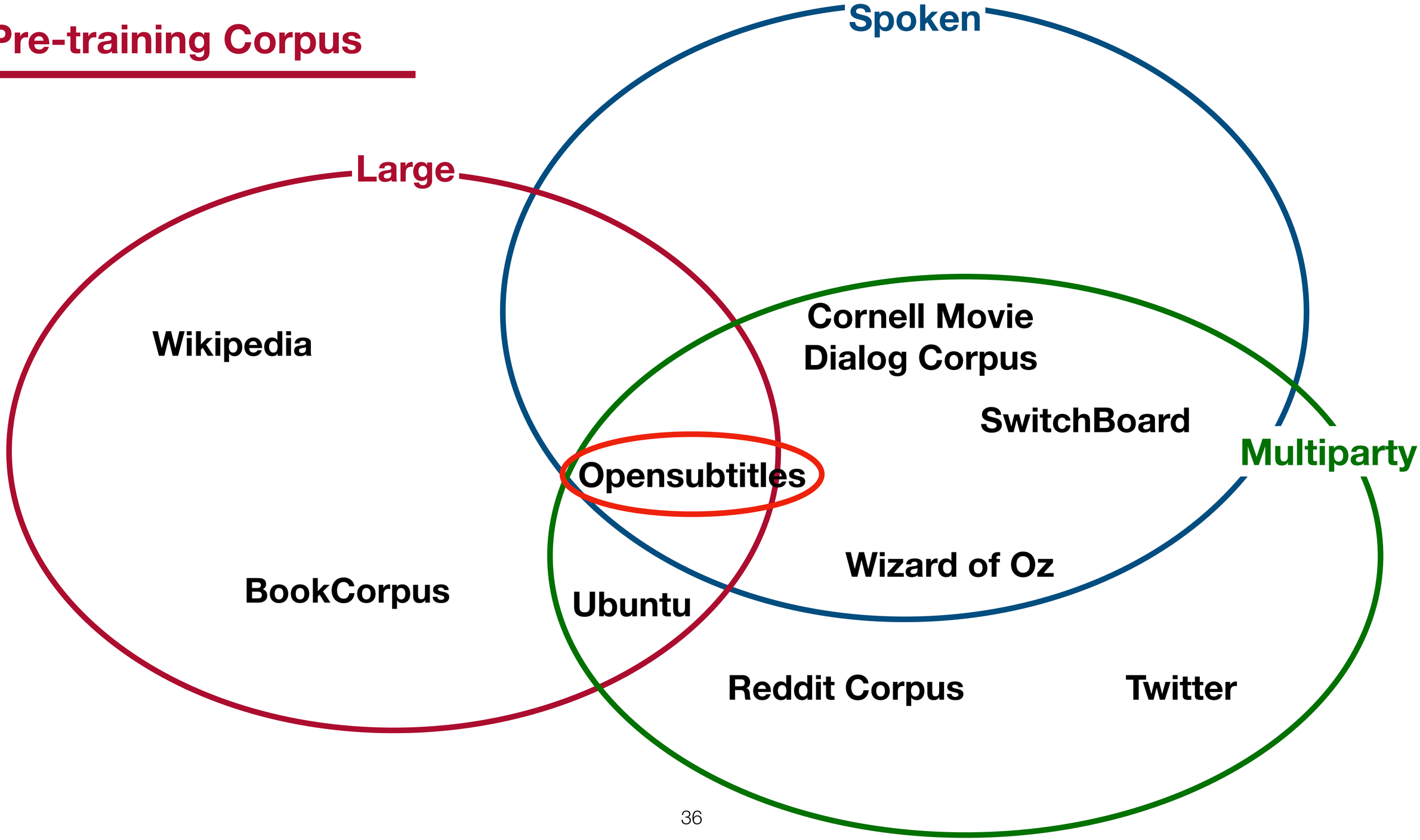
Pre-training Corpus



Pre-training Corpus



Pre-training Corpus



Formalisation

$u_1^{L_1}$ 

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.

 $u_2^{L_2}$

$u_3^{L_2}$ 

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!

 $u_4^{L_1}$

$u_5^{L_2}$ 

Well, you are in prison…

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

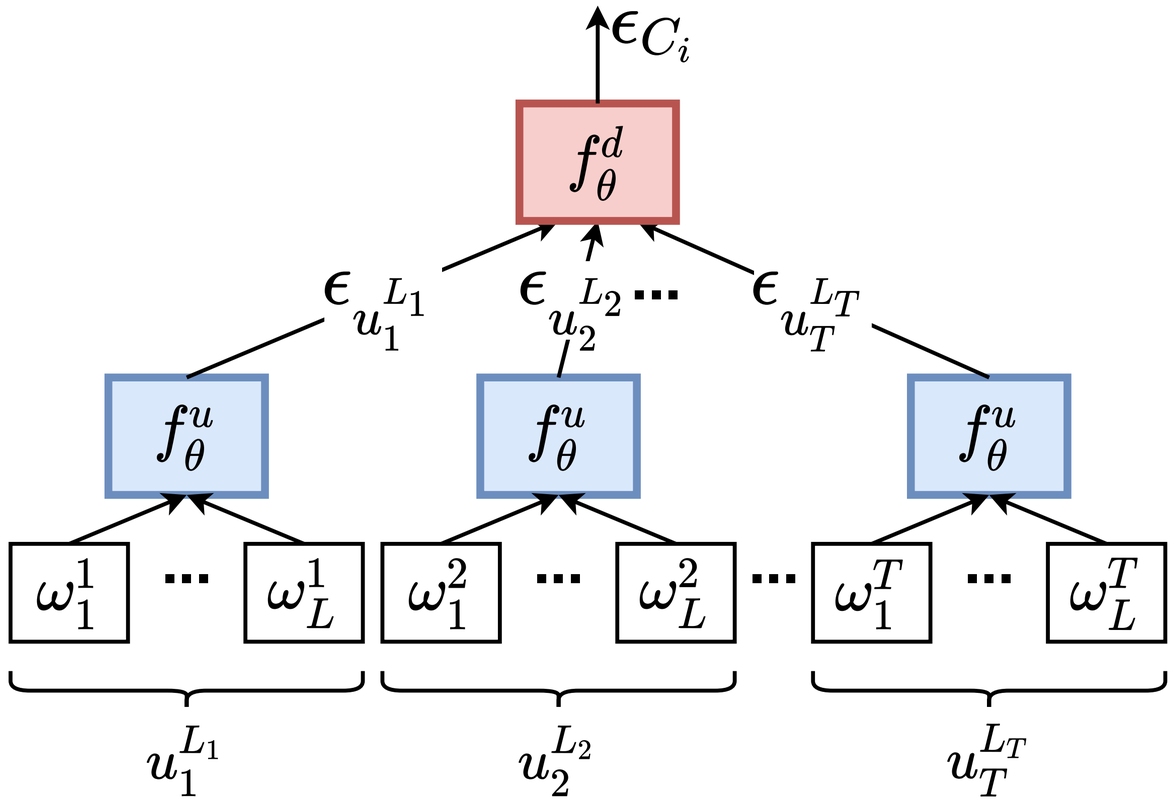
Models

Goal: **Learn a multi-purpose dialog embedding**

Hierarchical Transformer Encoder: f_θ^u **and** f_θ^d

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

$$\mathscr{E}_{C_i} = f_\theta^d(\mathscr{E}_{u_1^{L_1}}, \ldots, \mathscr{E}_{u_T^{L_T}}),$$



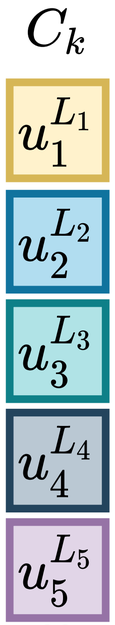
Losses

Dialog Level Pretraining

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

Generic Framework

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



Losses

Dialog Level Pretraining

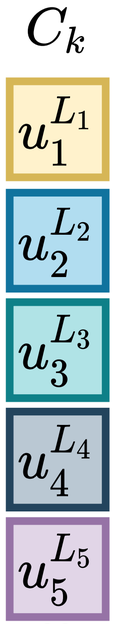
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Generic Framework

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

Set of masked tokens

Set of masked indices



Losses

Dialog Level Pretraining

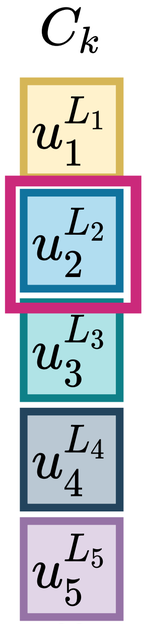
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Generic Framework

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

Set of masked tokens → $\{u_2^{L_2}, u_4^{L_4}\}$

Set of masked indices → $\{2,4\}$



Losses

Dialog Level Pretraining

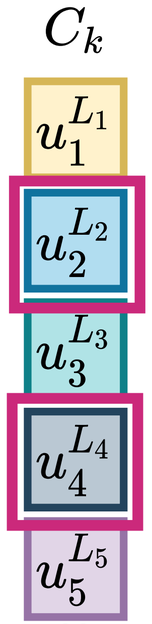
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Generic Framework

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

Set of masked tokens → $\{u_2^{L_2}, u_4^{L_4}\}$

Set of masked indices → $\{2,4\}$



Losses

Dialog Level Pretraining

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

Generic Framework

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

Set of masked tokens $\{u_2^{L_2}, u_4^{L_4}\}$

Set of masked indices $\{2,4\}$

$\tilde{C}_k$

$u_1^{L_1}$

$\tilde{u}_M$

$u_3^{L_3}$

$\tilde{u}_M$

$u_5^{L_5}$

Losses

Dialog Level Pretraining

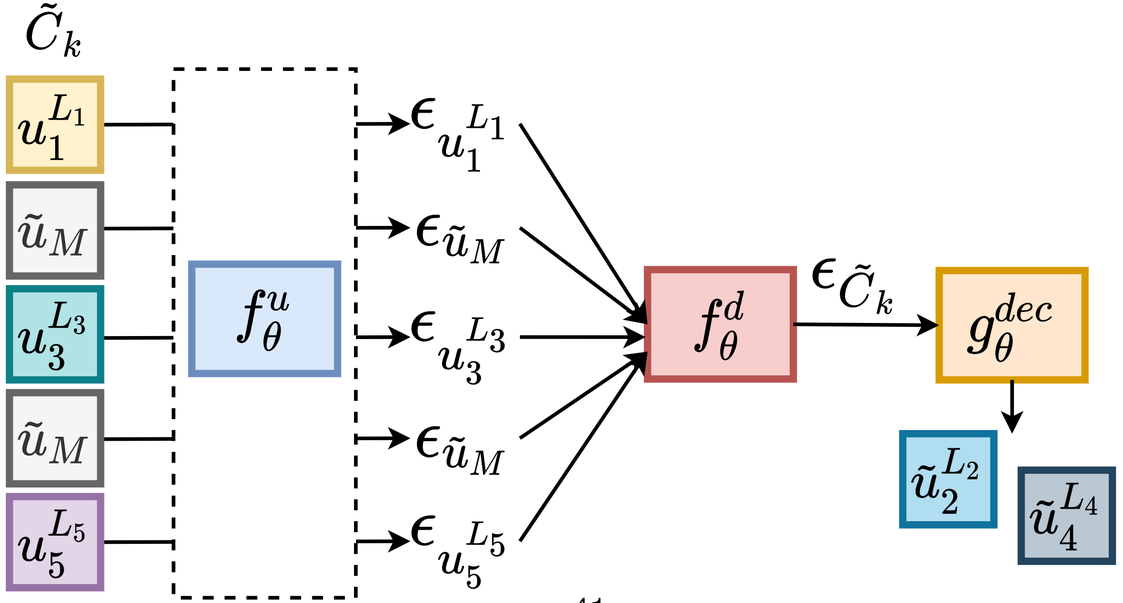
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Generic Framework

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

Set of masked tokens: $\{u_2^{L_2}, u_4^{L_4}\}$

Set of masked indices: $\{2,4\}$



Losses

Dialog Level Pretraining

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

Masked Utterance Generation (MUG)

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

C_k

$u_1^{L_1}$

$u_2^{L_1}$

$u_3^{L_1}$

$u_4^{L_1}$

$u_5^{L_1}$

Losses

Dialog Level Pretraining

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

Masked Utterance Generation (MUG)

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

Context Monolingual

C_k

$u_1^{L_1}$

$u_2^{L_1}$

$u_3^{L_1}$

$u_4^{L_1}$

$u_5^{L_1}$

Losses

Dialog Level Pretraining

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

Masked Utterance Generation (MUG)

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

Context Monolingual

Losses

Dialog Level Pretraining

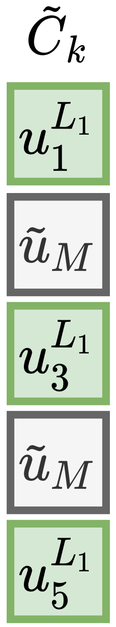
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Masked Utterance Generation (MUG)

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

Set of masked tokens: $\{u_2^{L_1}, u_4^{L_1}\}$

Set of masked indices: $\{2,4\}$



Losses

Dialog Level Pretraining

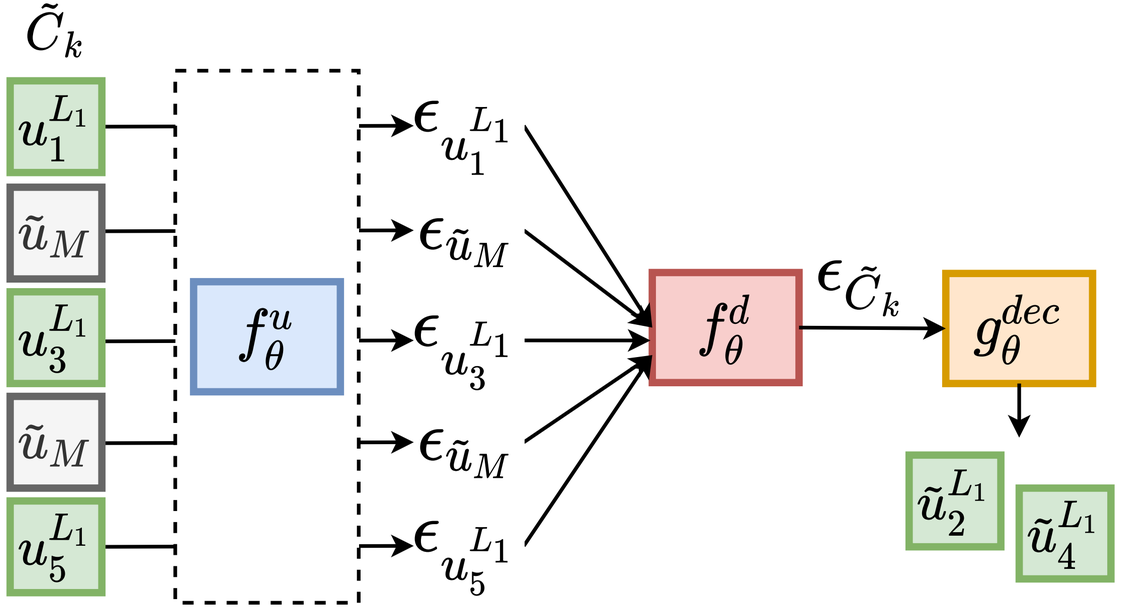
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Masked Utterance Generation (MUG)

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

Set of masked tokens: $\{u_2^{L_1}, u_4^{L_1}\}$

Set of masked indices: $\{2,4\}$



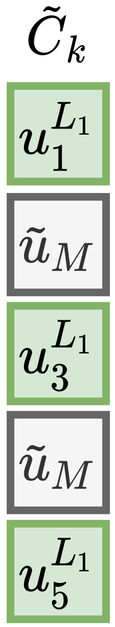
Losses

Dialog Level Pretraining

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

Translation Masked Utterance Generation (TMUG)

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



Losses

Dialog Level Pretraining

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

Translation Masked Utterance Generation (TMUG)

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

Context Monolingual

$\tilde{C}_k$

$u_1^{L_1}$

$\tilde{u}_M$

$u_3^{L_1}$

$\tilde{u}_M$

$u_5^{L_1}$

Losses

Dialog Level Pretraining

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

Translation Masked Utterance Generation (TMUG)

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

Set of masked indices ⟶ $\{2,4\}$

Context Monolingual

$\tilde{C}_k$

$u_1^{L_1}$

$\tilde{u}_M$

$u_3^{L_1}$

$\tilde{u}_M$

$u_5^{L_1}$

Losses

Dialog Level Pretraining

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

Translation Masked Utterance Generation (TMUG)

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

Set of masked tokens

$\{u_2^{L_2}, u_4^{L_2}\}$

Set of masked indices → $\{2,4\}$

Context Monolingual

$\tilde{C}_k$

$u_1^{L_1}$

$\tilde{u}_M$

$u_3^{L_1}$

$\tilde{u}_M$

$u_5^{L_1}$

Losses

Dialog Level Pretraining

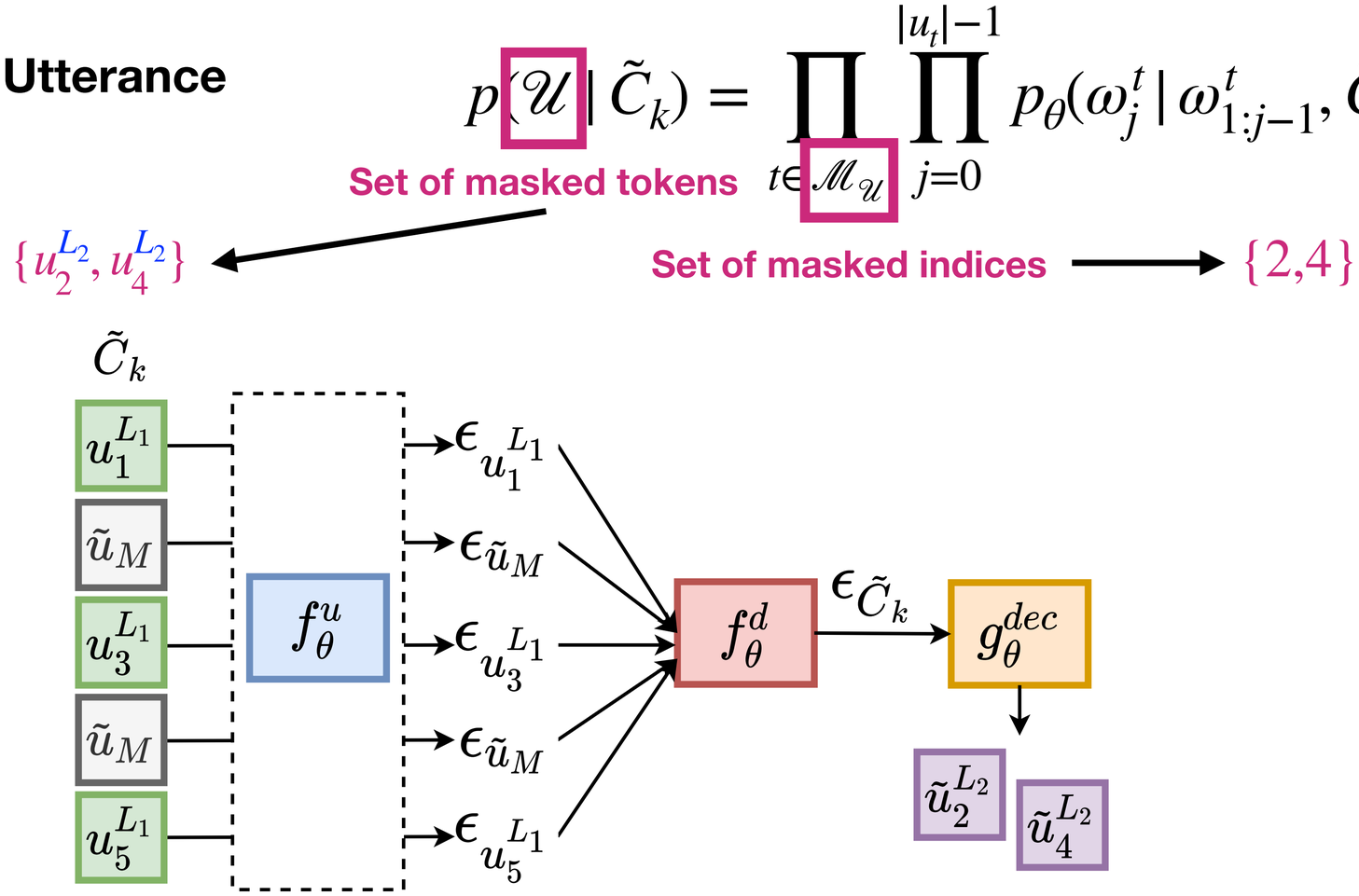
$$\mathscr{L}(\theta) = \lambda_u \mathscr{L}^u(\theta) + \lambda_d \mathscr{L}^d(\theta) .$$

Translation Masked Utterance Generation (TMUG)

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

Set of masked tokens

Set of masked indices



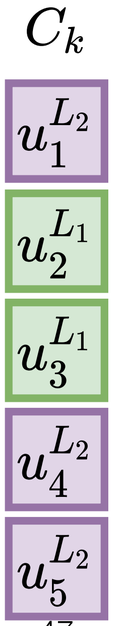
Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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



Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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

Context Multilingual

C_k

$u_1^{L_2}$

$u_2^{L_1}$

$u_3^{L_1}$

$u_4^{L_2}$

$u_5^{L_2}$

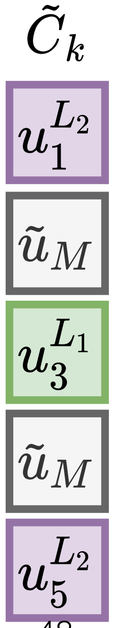
Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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



Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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

Set of masked indices $\longrightarrow \{2,4\}$

$\tilde{C}_k$

$u_1^{L_2}$

$\tilde{u}_M$

$u_3^{L_1}$

$\tilde{u}_M$

$u_5^{L_2}$

Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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

Set of masked tokens

Set of masked indices

$\{u_2^{L_1}, u_4^{L_2}\}$

$\{2,4\}$

$\tilde{C}_k$

$u_1^{L_2}$

$\tilde{u}_M$

$u_3^{L_1}$

$\tilde{u}_M$

$u_5^{L_2}$

Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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

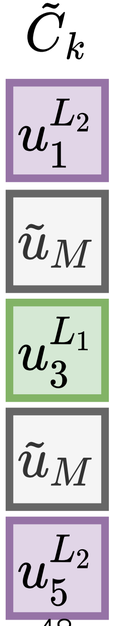
Set of masked tokens

Set of masked indices

$\{u_2^{L_1}, u_4^{L_2}\}$

$\{2,4\}$

Context Multilingual



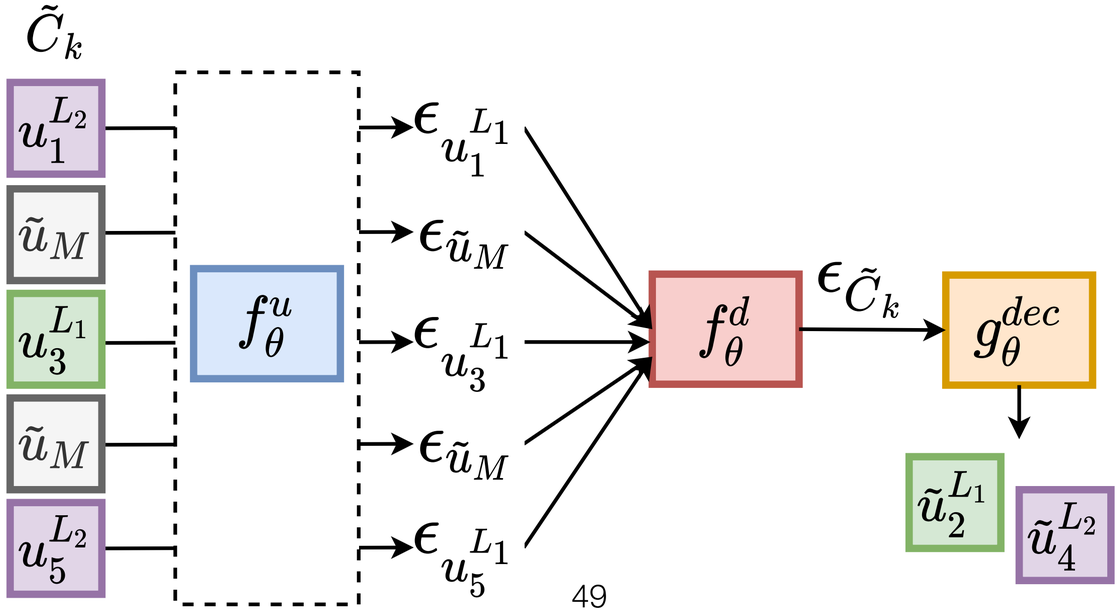
Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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



Losses

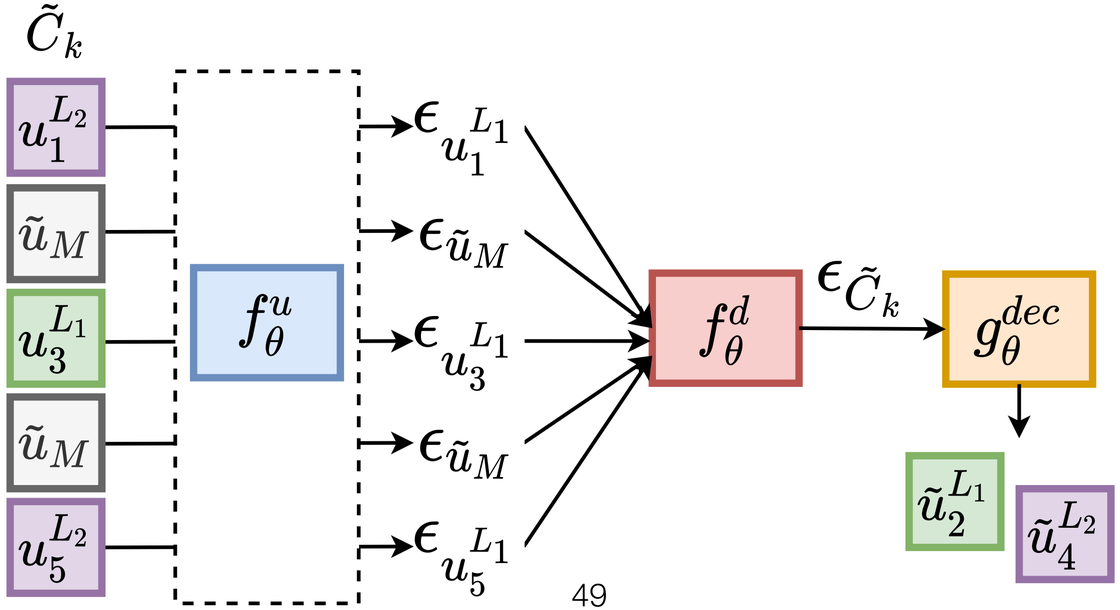
Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

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

Set of masked indices ⟶ $\{2,4\}$



Losses

Dialog Level Pretraining

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

Multilingual Masked Utterance Generation (MMUG)

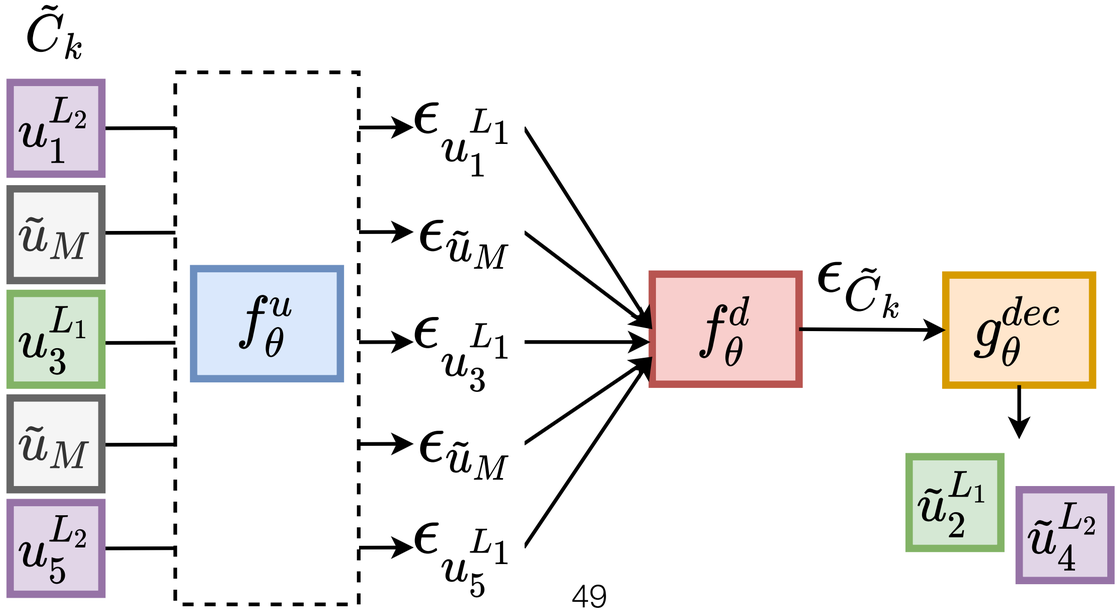
$$p(\mathscr{U} \mid \tilde{C}_k) = \prod_{t \in \mathscr{M}_{\mathscr{U}}} \prod_{j=0}^{|u_t|-1} p_\theta(\omega_j^t \mid \omega_{1:j-1}^t, \tilde{C}_k) .$$

Set of masked tokens

Set of masked indices

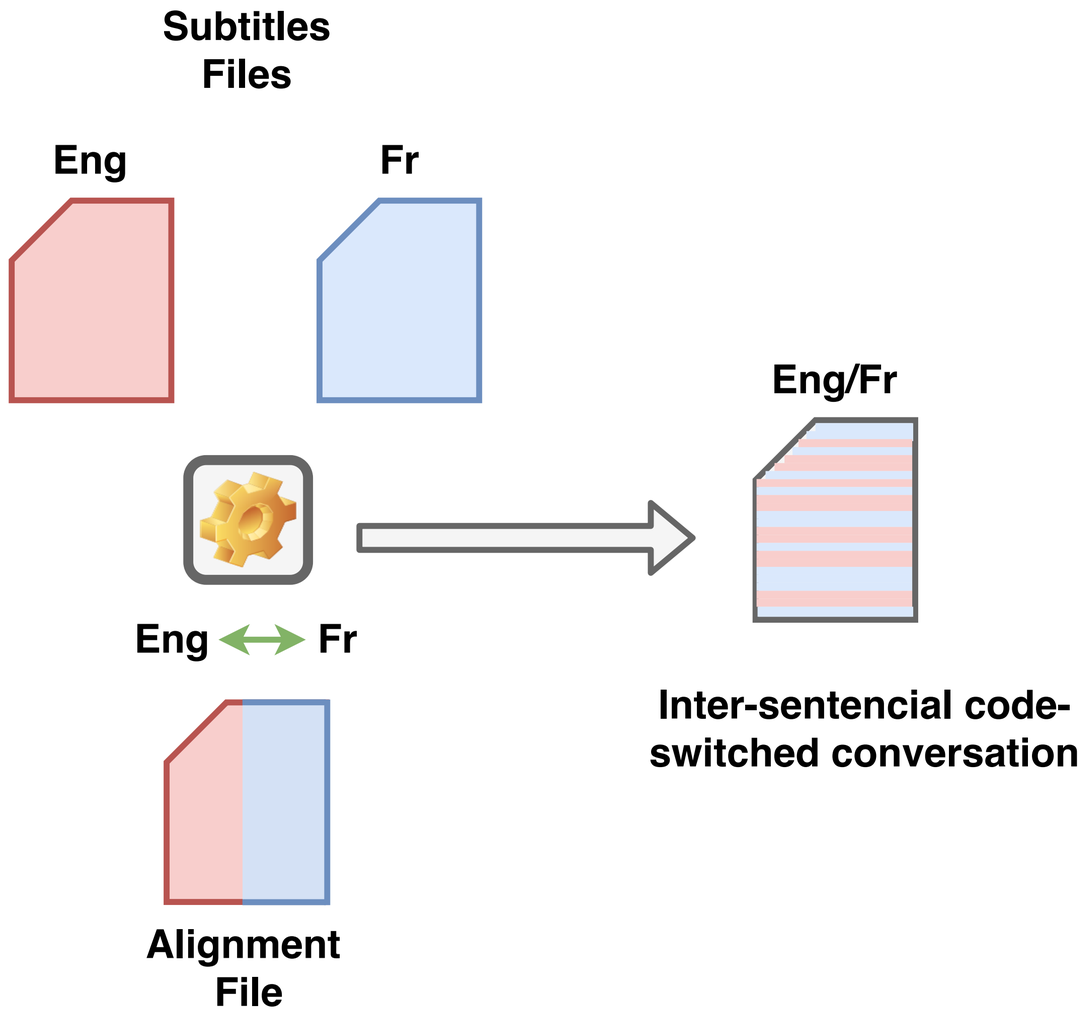
$\{u_2^{L_1}, u_4^{L_2}\}$

$\{2,4\}$



Pre-training Corpus

OpenSubtitles



Evaluation

Three main tasks:

- **Dialog Act Classification (DA)**
- **Next Utterance Recognition (NUR)**
- **Inconsistency Identification (II)**

Two types of evaluation:

- **Monolingual dialog context**
- **Multilingual dialog context**

Evaluation

Three main tasks:

Dialog Act Classification (DA)

Next Utterance Recognition (NUR)

Inconsistency Identification (II)

Two types of evaluation:

Monolingual dialog context

Multilingual dialog context

Multilingual dialog act benchmark (MIAM)

- English: MapTask
- French: LORIA
- German: Verbmobil (VM2)
- Italian: Ilisten
- Spanish: Dihana

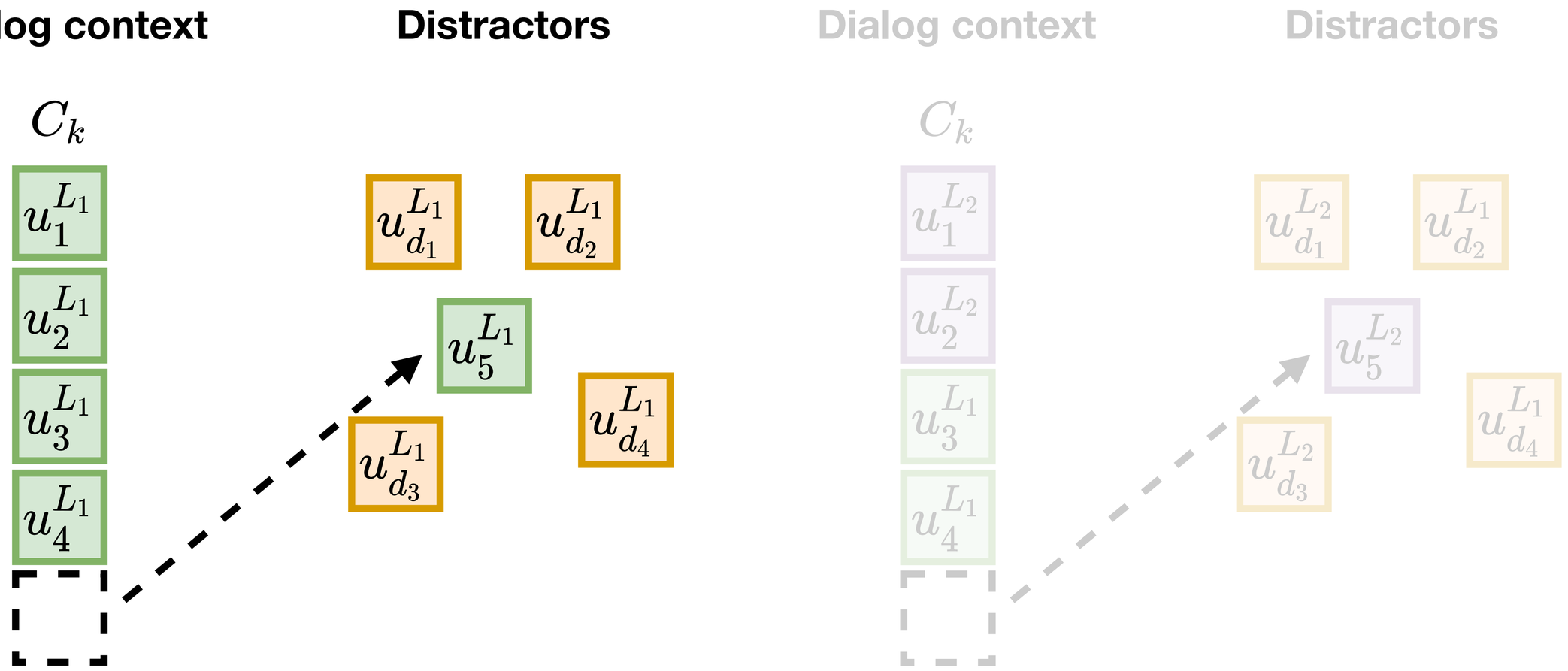
Results on MIAM

	Toke.	VM2	Map Task	Dihana	Loria	Ilisten	Total
BERT	lang	54.7	66.4	86.0	50.2	74.9	66.4
BERT - 4*layers*	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	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 - 4*layers*	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**

addition of cross-lingual generation during pre-training helps.

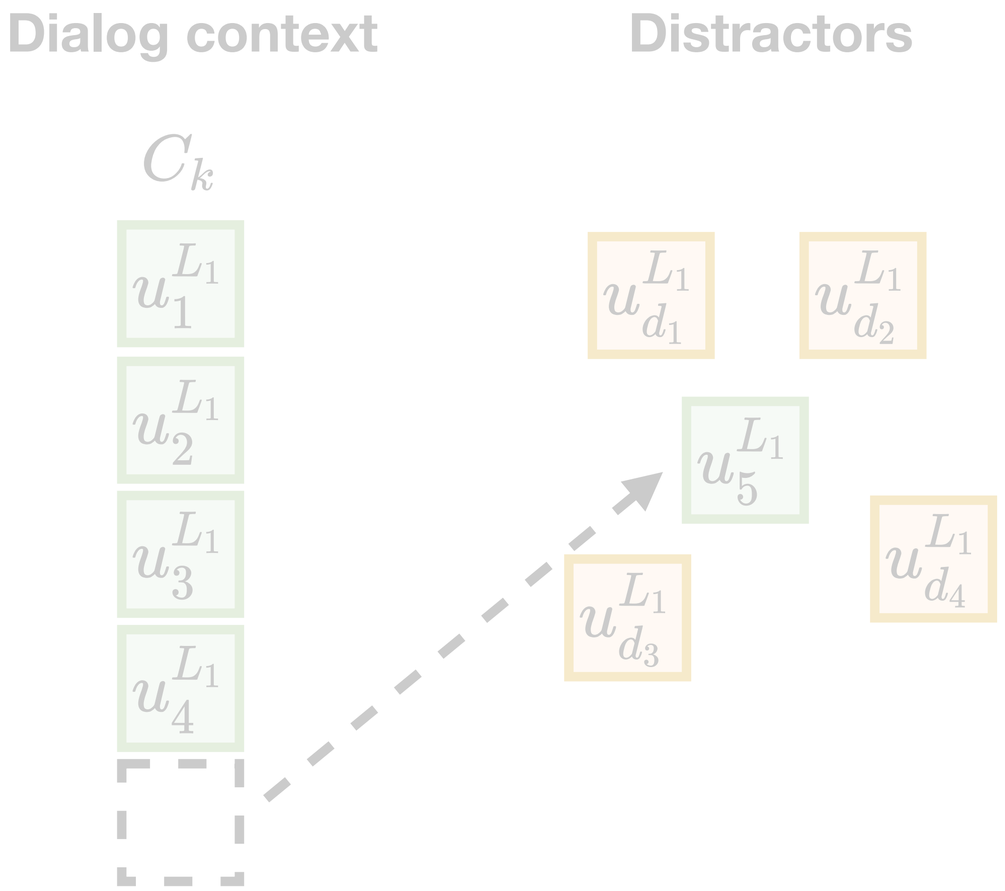
Evaluation

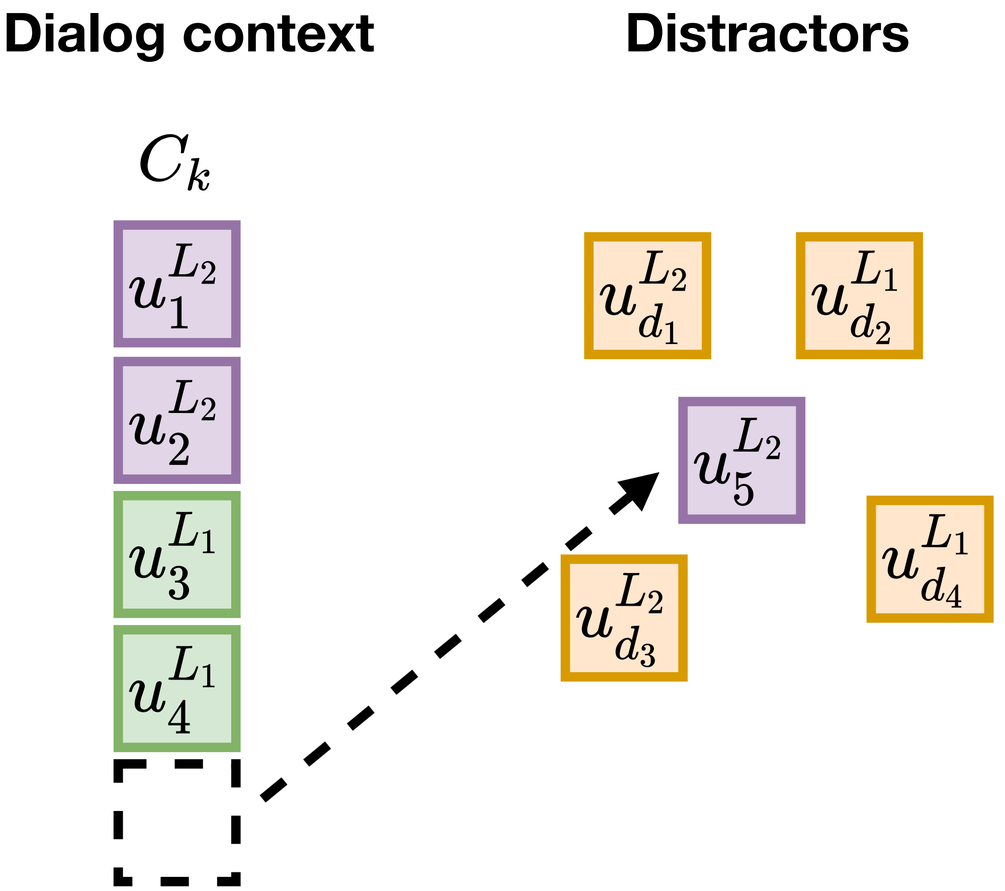
Next Utterance Recognition (NUR)



Evaluation

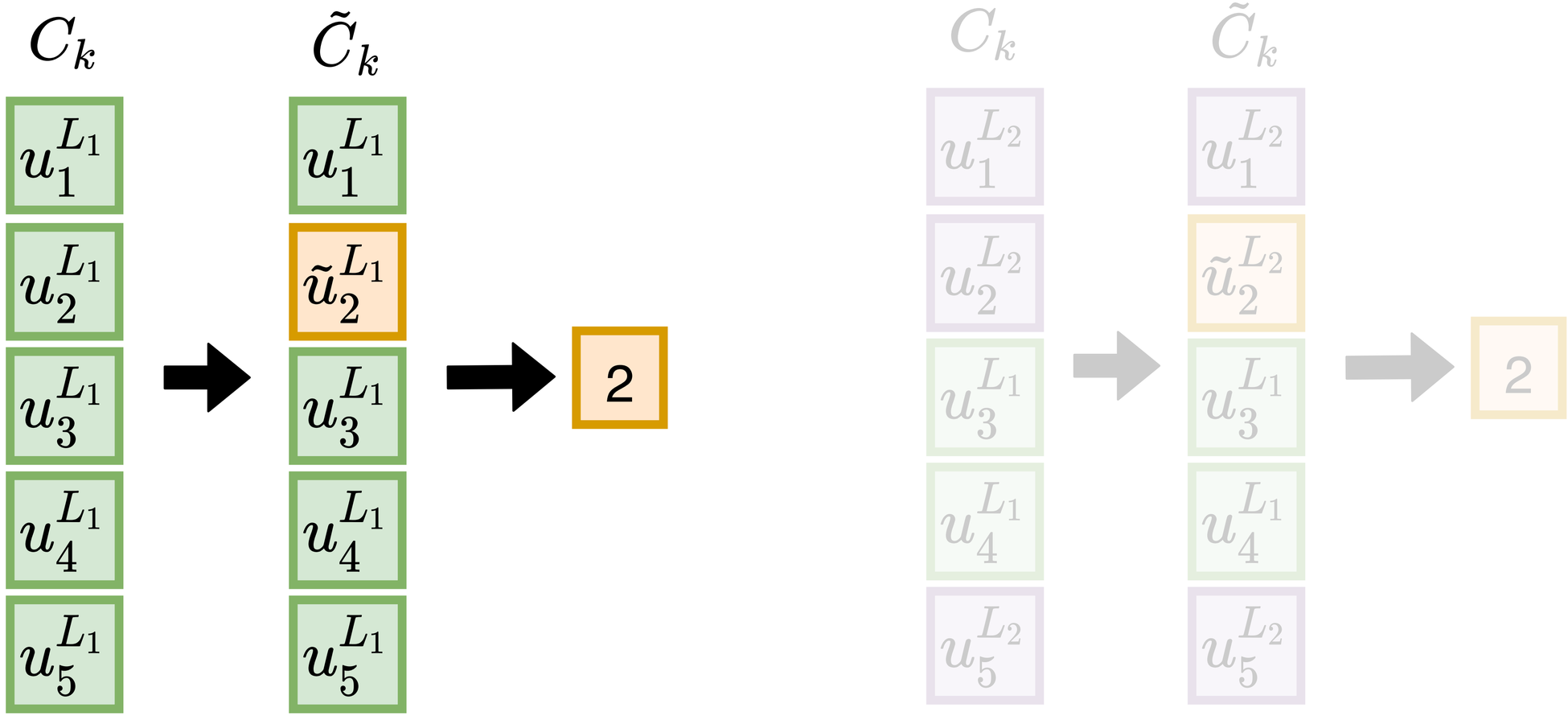
Next Utterance Recognition (NUR)





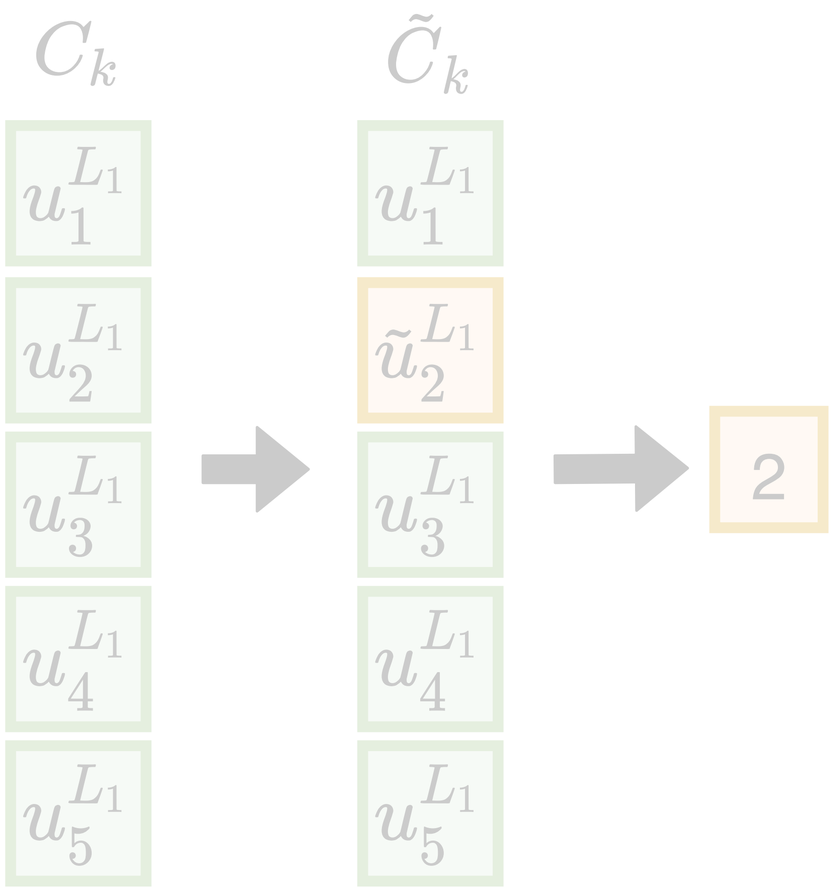
Evaluation

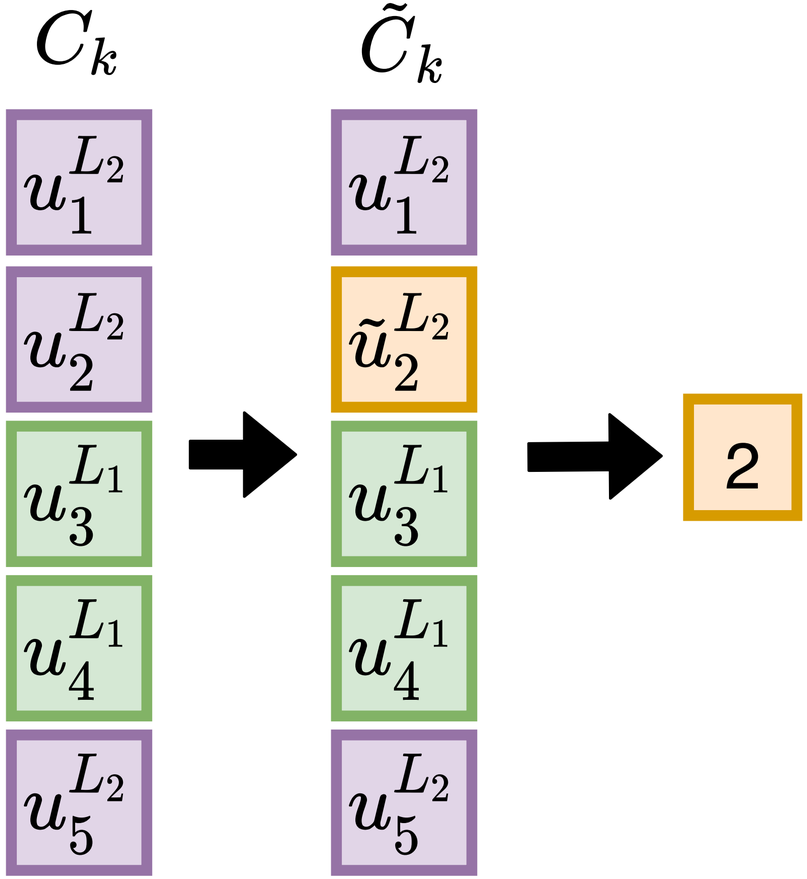
Inconsistency Identification (II)



Evaluation

Inconsistency Identification (II)





Results

Next Utterance Recognition (NUR)

	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**

Results

Inconsistency Identification (II)

	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**

Results

Multilingual Next Utterance Recognition (mNUR)

	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**

Results

Multilingual Inconsistency Identification (mII)

	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**

From monomodal to multimodal transcript embeddings

From monomodal to multimodal transcript embeddings

Previous embeddings were limited to text.

From monomodal to multimodal transcript embeddings

Previous embeddings were limited to text.

However Human Communication is multi-modal.

From monomodal to multimodal transcript embeddings

Previous embeddings were limited to text.

However Human Communication is multi-modal.

Verbal

« What you say »

- Lexicon:
  - Words
- Syntax:
  - POS
- Pragmatics:
  - DA
  - Emotion

Language X_l

From monomodal to multimodal transcript embeddings

Previous embeddings were limited to text.

However Human Communication is multi-modal.

Verbal

« What you say »

- Lexicon:
  - Words
- Syntax:
  - POS
- Pragmatics:
  - DA
  - Emotion

Language X_l

« How you say it »

Vocal

- Prosody
  - Intonation
  - Voice quality
- Vocal expressions:
  - Laughter
  - Moans

Audio X_a

Visual

- Gestures:
  - Head & Eye
- Body language
  - Body posture
- Eye contact
- Facial expressions
  - Action units

Video X_v

Multimodal Task Description

Multimodal Task Description

Goal: Include Multimodal Dimension in Representations of Spoken Transcripts

Multimodal Task Description

Goal: Include Multimodal Dimension in Representations of Spoken Transcripts

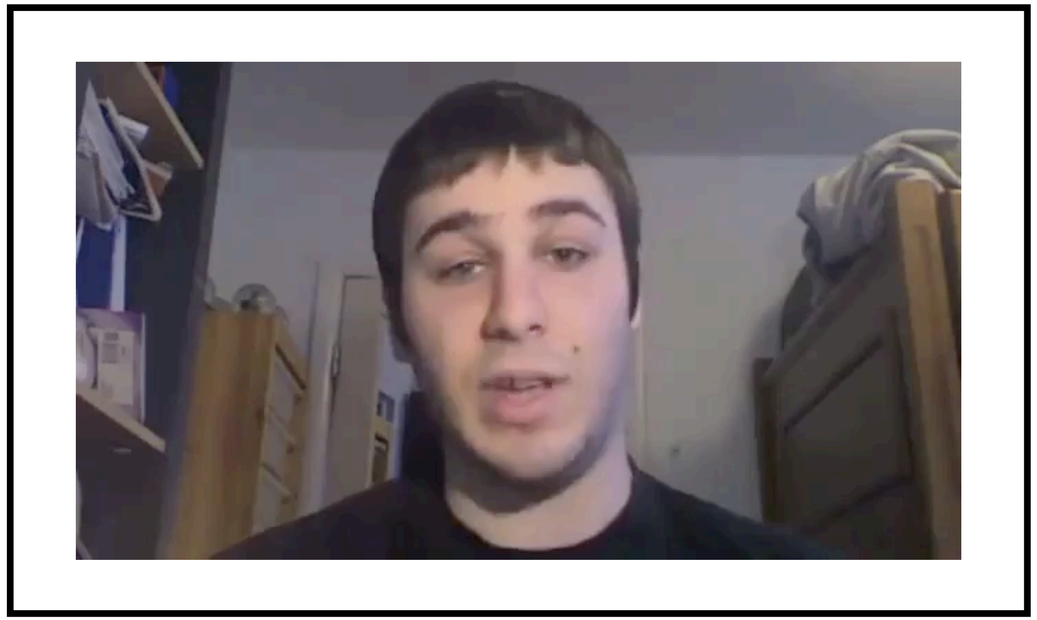
Input Video

The action is fucking awesome!

Multimodal Task Description

Goal: Include Multimodal Dimension in Representations of Spoken Transcripts

Input Video



The action is fucking awesome!

Emotion Predictor

Emotion

Positive +3

.....

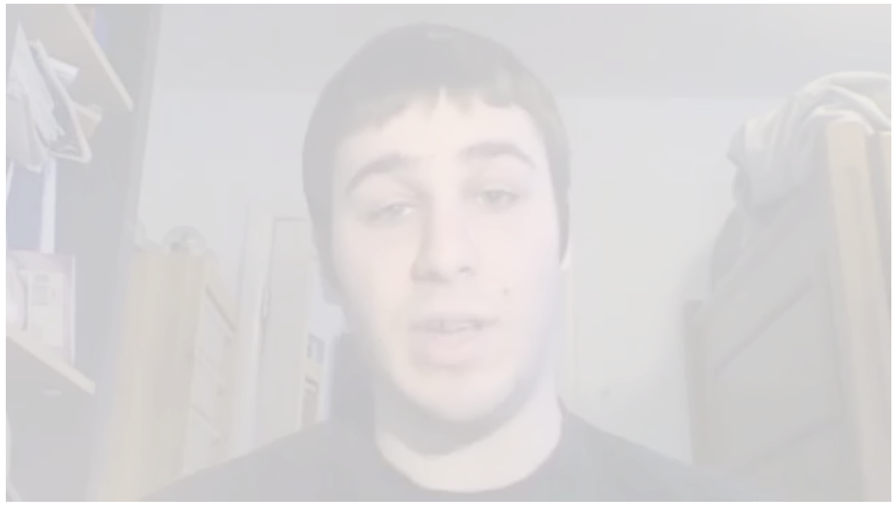
Negative -3

Multimodal Task Description

Goal: Include Multimodal Dimension in Representations of Spoken Transcripts

Task: Learn a multi-modal emotion predictor

Input Video



The action is fucking awesome!

Emotion Predictor

Emotion

Positive +3

.....

Negative -3

Core Challenges in Multimodal Learning

Core Challenges in Multimodal Learning

5 challenges of multimodal learning

Core Challenges in Multimodal Learning

5 challenges of multimodal learning

Representation	Alignement	Fusion	Translation	Co-learning
Represent multimodal data (leverage complementarity, redundancy)	**Identify** relations between elements of different modalities	**Join** information from modalities	**Translate** one modality to another	**Transfer** knowledge between modalities

Core Challenges in Multimodal Learning

5 challenges of multimodal learning

Representation

Represent multimodal data (leverage complementarity, redundancy)

Alignement

Identify relations between elements of different modalities

Fusion

Join information from modalities

Translation

Translate one modality to another

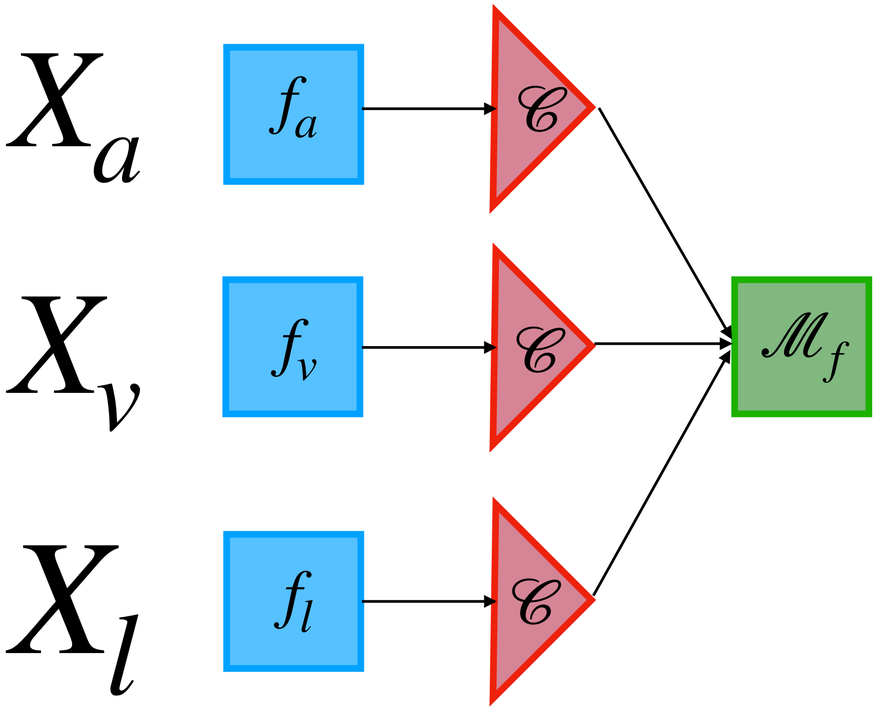
Co-learning

Transfer knowledge between modalities

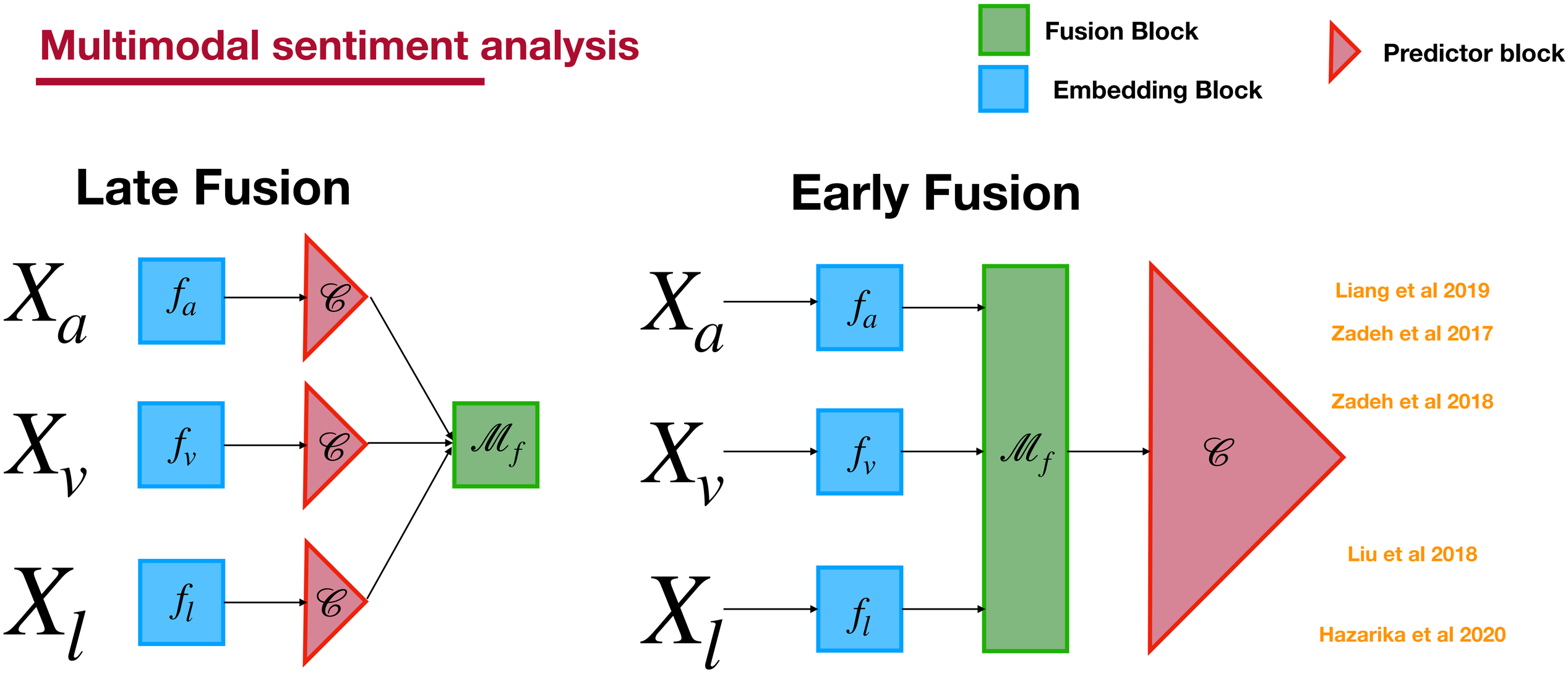
Multimodal sentiment analysis

Multimodal sentiment analysis

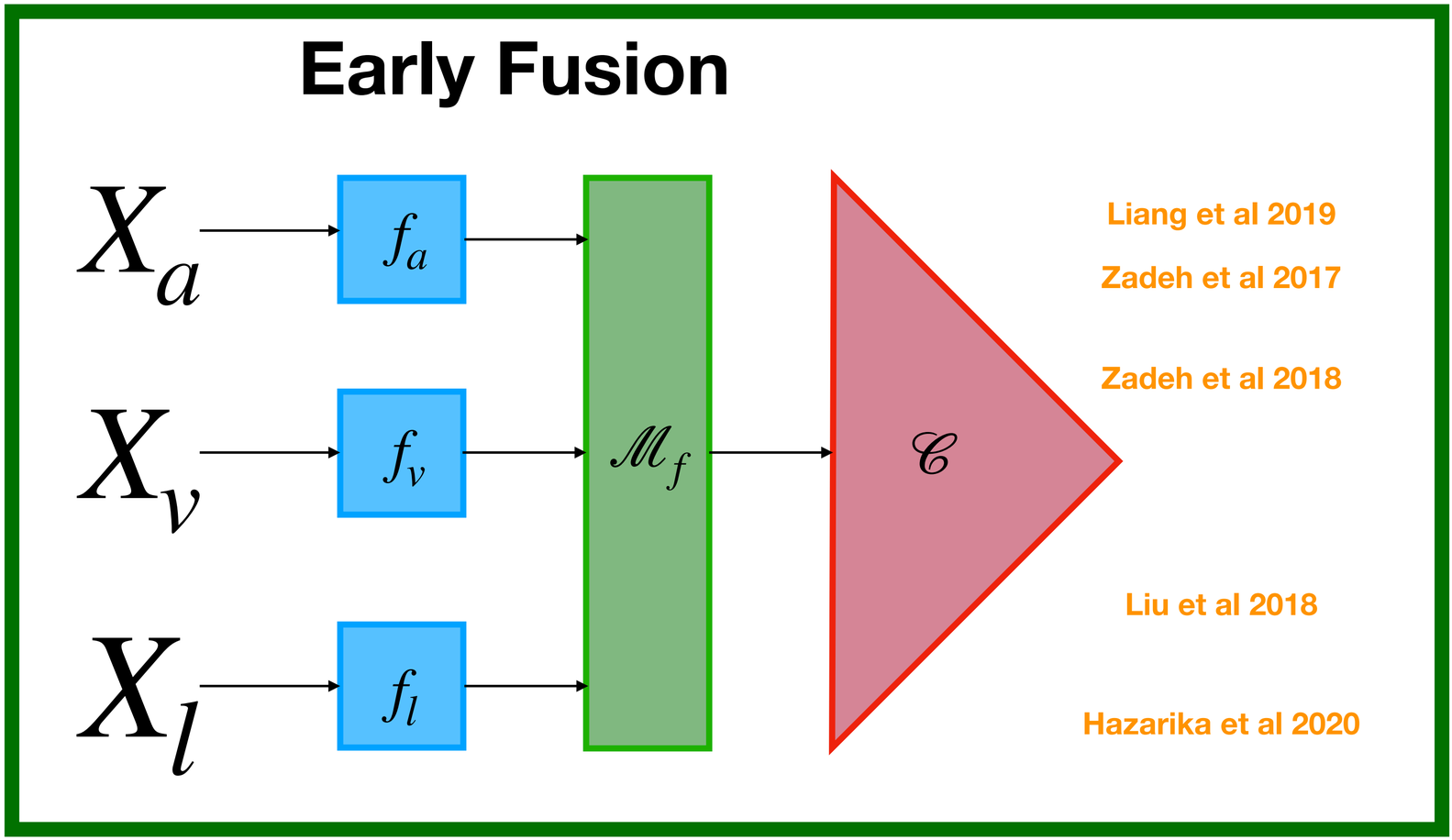
Late Fusion



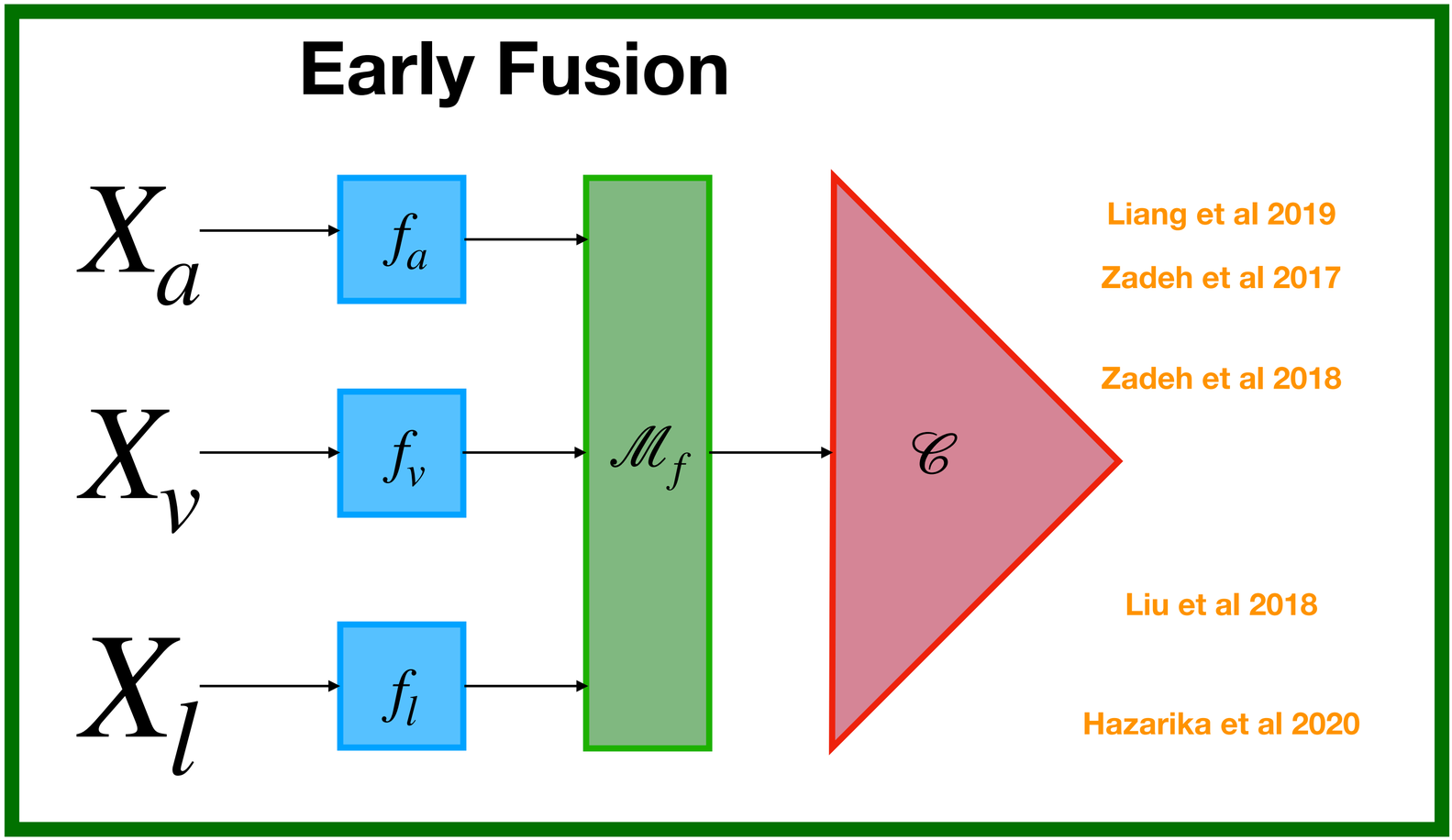
Multimodal sentiment analysis



Multimodal sentiment analysis



Multimodal sentiment analysis



Fusion in previous work mainly rely on complex neural networks

Few previous works on improving fusion using loss function!

Conclusion

- New codes-witched inspired losses help to learn representations for both monolingual and multilingual dialogs.
- first that explicitly includes code switching during pretraining to learn multilingual spoken dialog representations.

Future works

- Work on OPS to obtain fine-grained alignments (e.g at the span and word levels).
- Extend MIAM with Emotion/Sentiment corpora.
- Few-shots scenarios.

Formalisation

Conversation C

$u_1^{L_1}$ 

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.

 $u_2^{L_2}$

$u_3^{L_2}$ 

That is quite the rap sheet, Prison Mike.

Et on ne m’a jamais chopé non plus!

 $u_4^{L_1}$

$u_5^{L_2}$ 

Well, you are in prison…

$$C_i = (u_1^{L_1}, u_2^{L_2}, \ldots, u_{|C_i|}^{L_{C_i}})$$

Importance of Pre-training for SILICONE

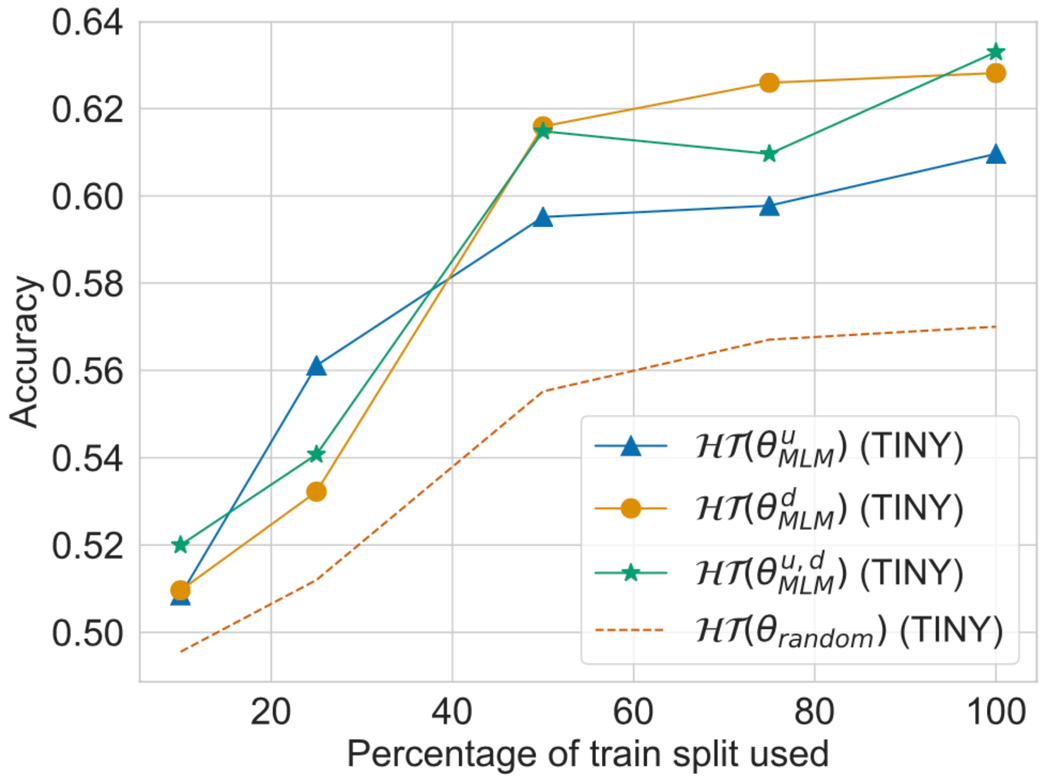


Figure 5.7 – A comparison of different parameters initialisation on $\texttt{MELD}_\texttt{s}$. Training is performed using a different percentage of complete training set. Validation and test set are fixed over all experimentation. Each score is the averaged accuracy over 10 random runs.

Formalisation

$u_1^{L_1}$ 

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.

 $u_2^{L_2}$

$u_3^{L_2}$ 

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!

 $u_4^{L_1}$

$u_5^{L_2}$ 

Well, you are in prison…

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

Evaluation

Sentence Representation Evaluation

- **Glue** [Wang et al., 2018]
- **SuperGlue** [Wang et al., 2019]
- **SentEval** [Conneau et al., 2018]

Dialogue Representation Evaluation

- **DialoGLUE** [Mehri et al., 2020]
- **Dialogue Dodecathlon** [Shuster et al., 2019]

Formalisation

$u_1^{L_1}$ 

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

I stole and I robbed.

$u_2^{L_2}$

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

$u_3^{L_2}$ 

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!

$u_4^{L_1}$

$u_5^{L_2}$ 

Well, you are in prison…

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

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…

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

Importance of Pre-training for SILICONE

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