

# Learning to Represent and Generate Text using Information Measures

Pierre Colombo

Thesis Advisors

Chloe Clavel, Telecom Paris, Palaiseau, France

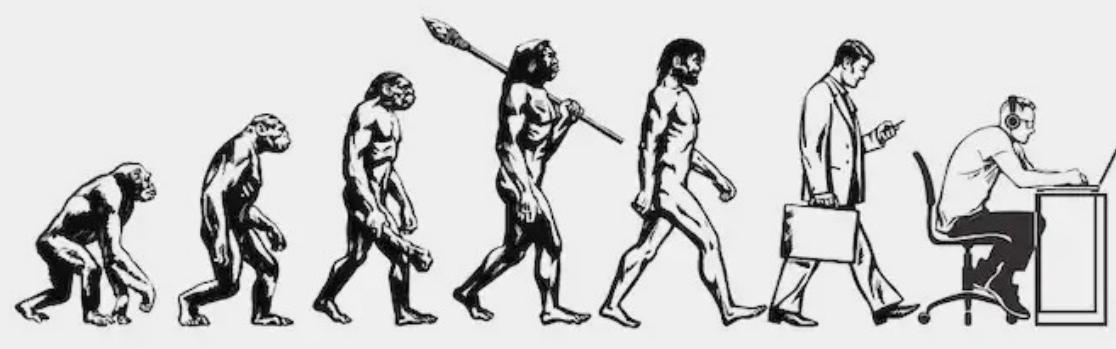
Giovanna Varni, Telecom Paris, Palaiseau, France

Emmanuel Vignon, IBM GBS, Bois Colombes, France

Joffrey Martinez, IBM GBS, Bois Colombes, France

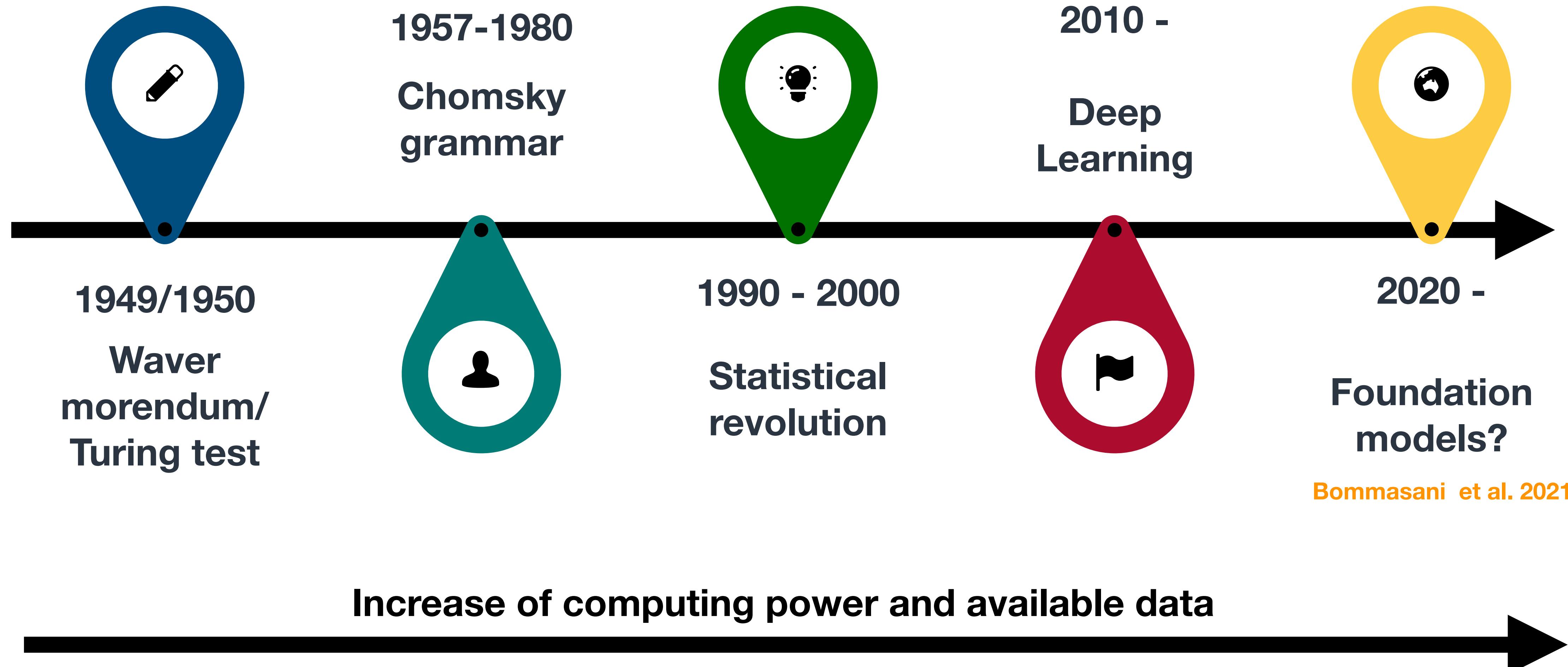
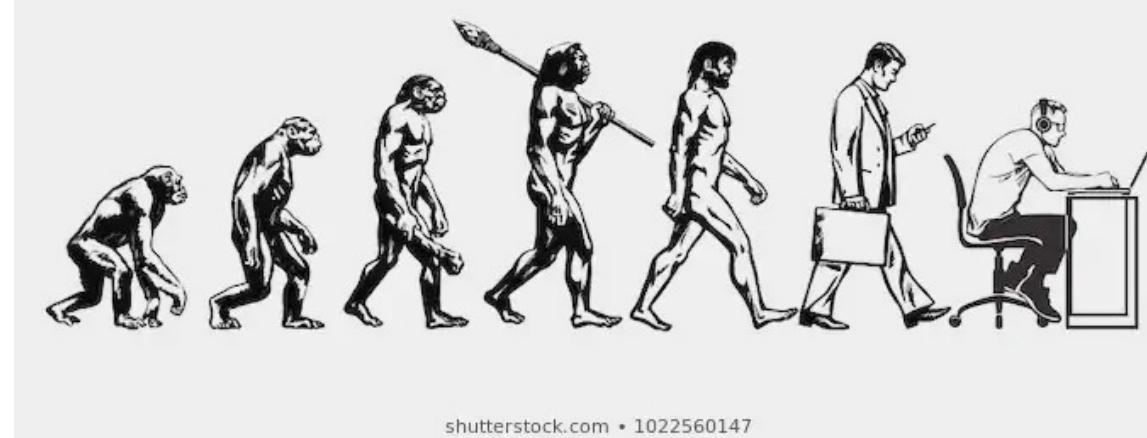
# A brief history of NLP

---

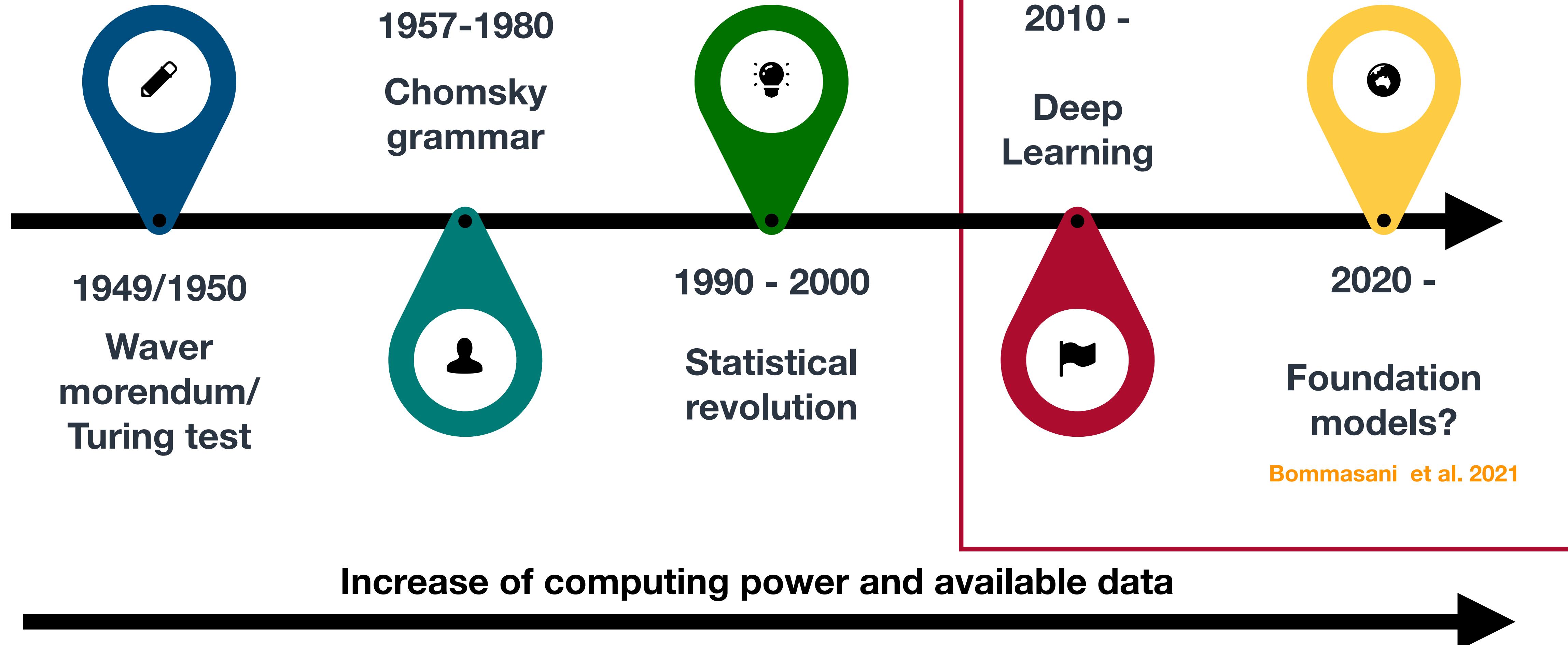
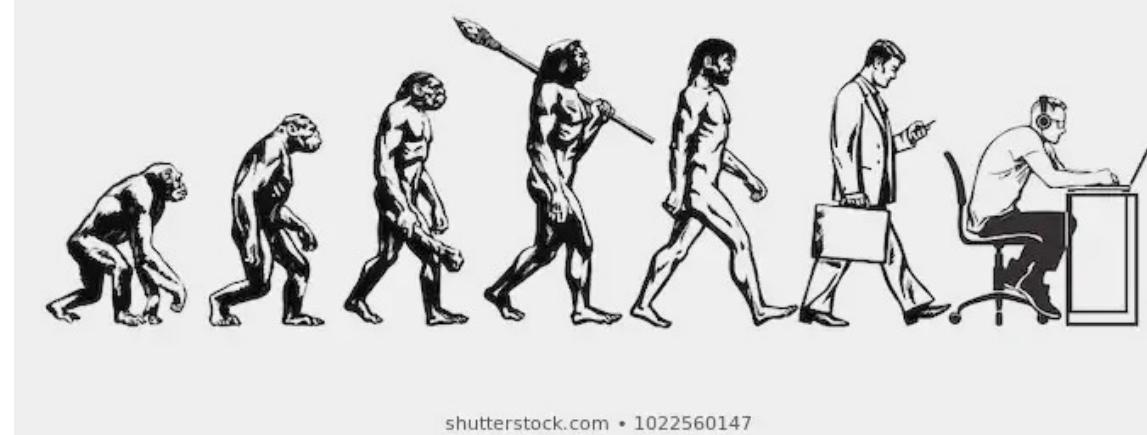


shutterstock.com • 1022560147

# A brief history of NLP



# A brief history of NLP



# Information Measures

---

# Information Measures

---

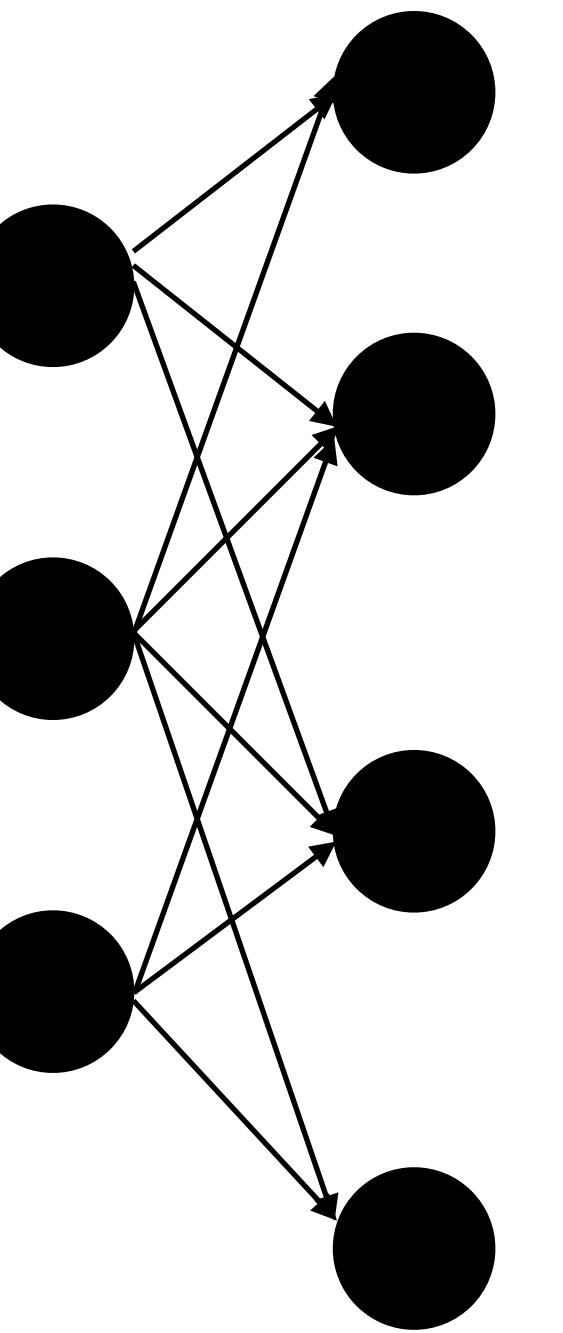
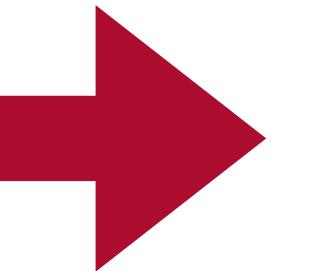
**Hello, Chicago.**  
**If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [....].**  
**Yes we can!**

**Input Text**

# Information Measures

---

Hello, Chicago.  
If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [....].  
Yes we can!



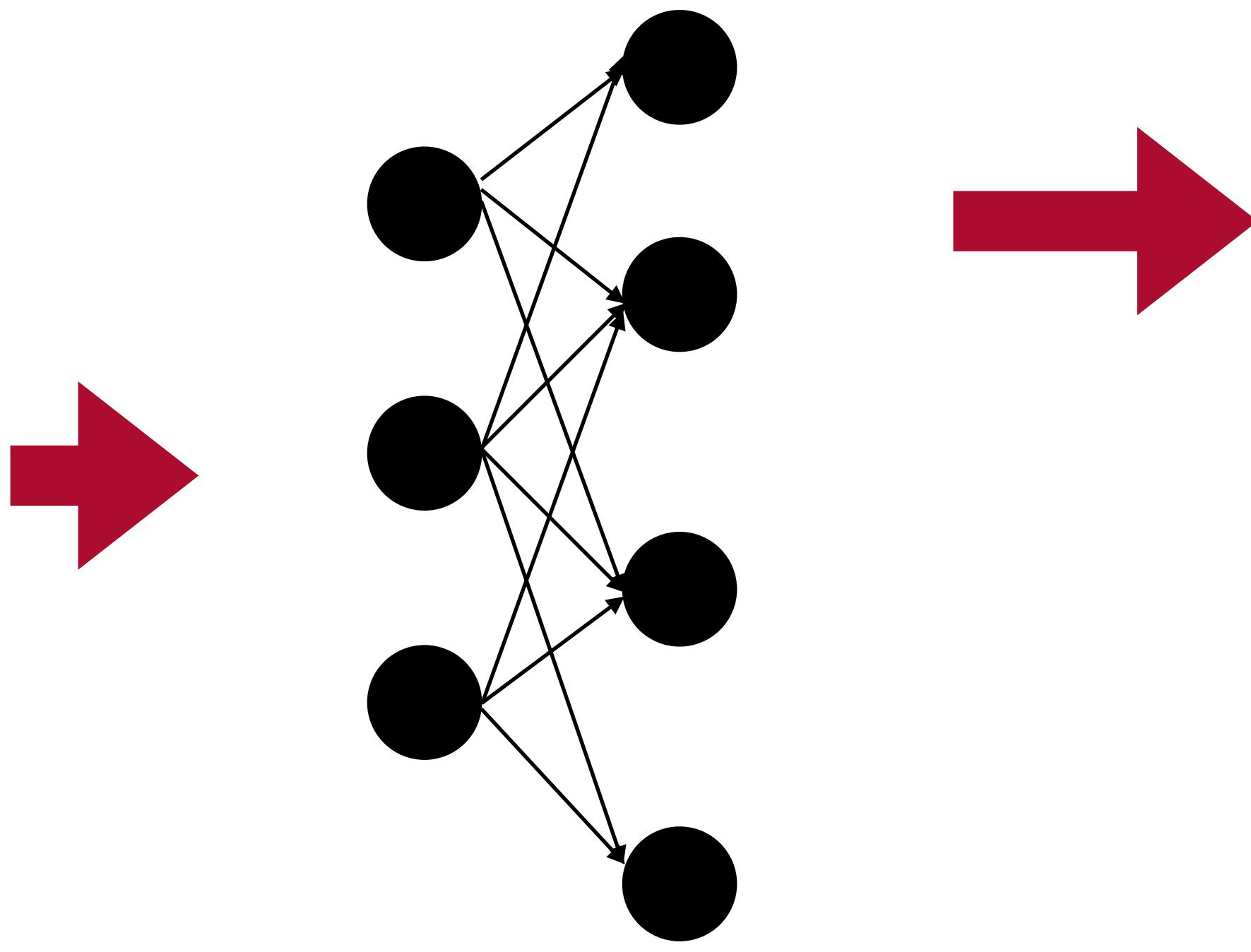
Input Text

Neural Network

# Information Measures

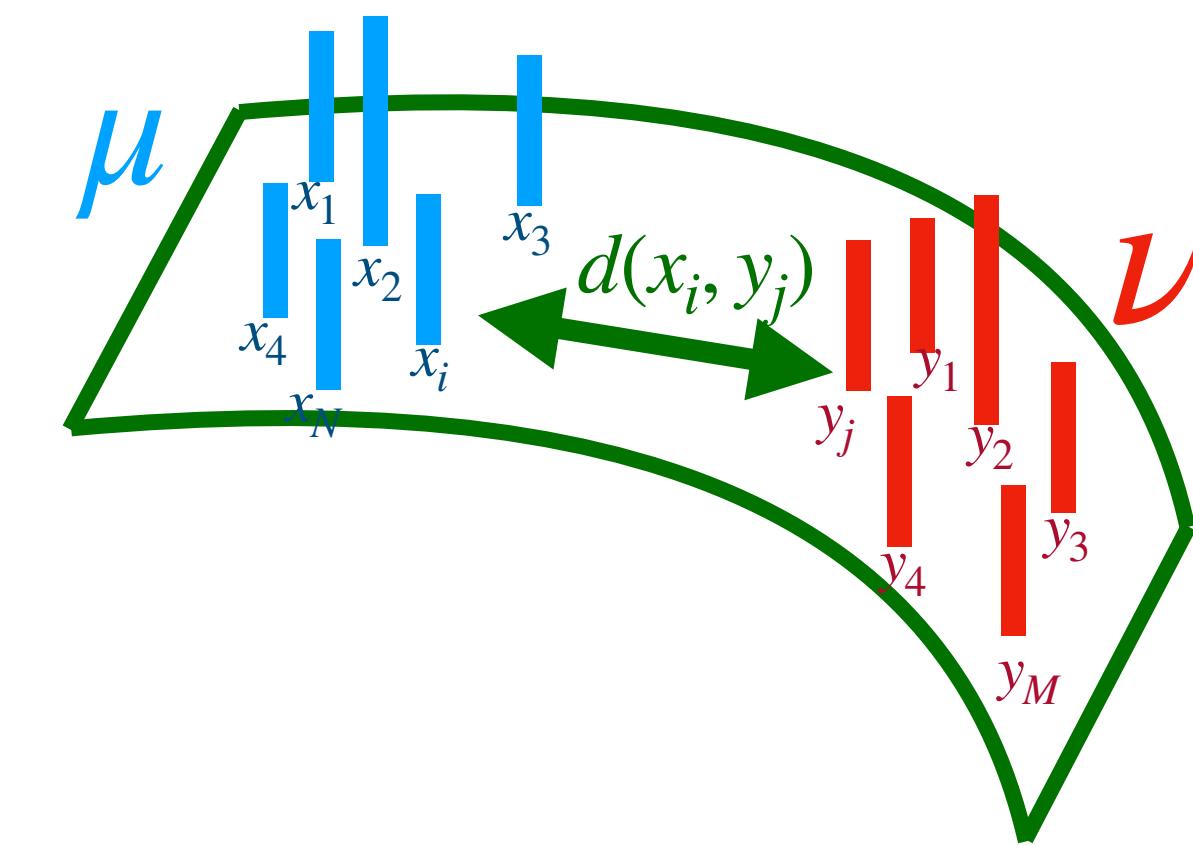
Hello, Chicago.  
If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [....].  
Yes we can!

Input Text



Neural Network

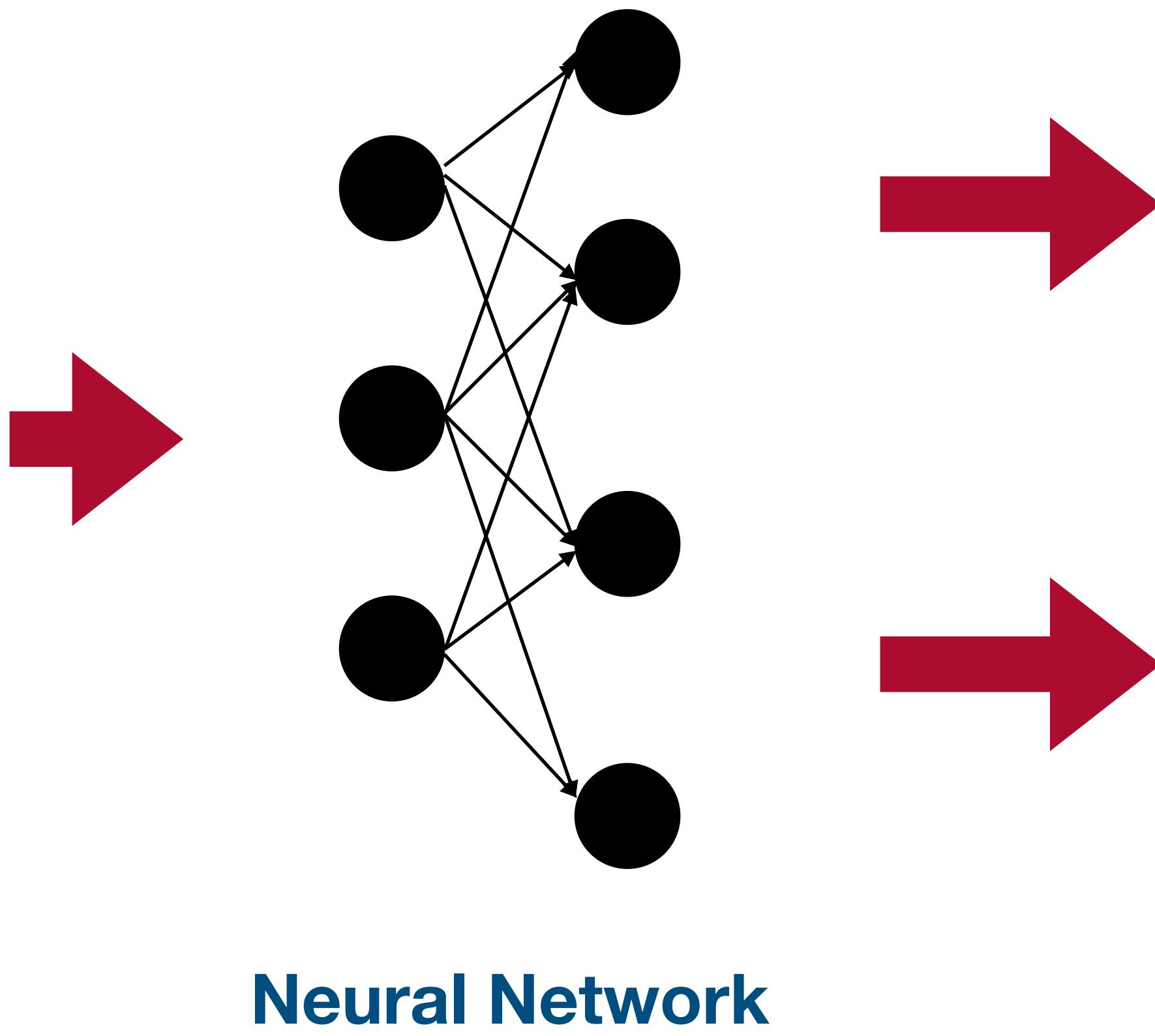
# High dimensional data



# Information Measures

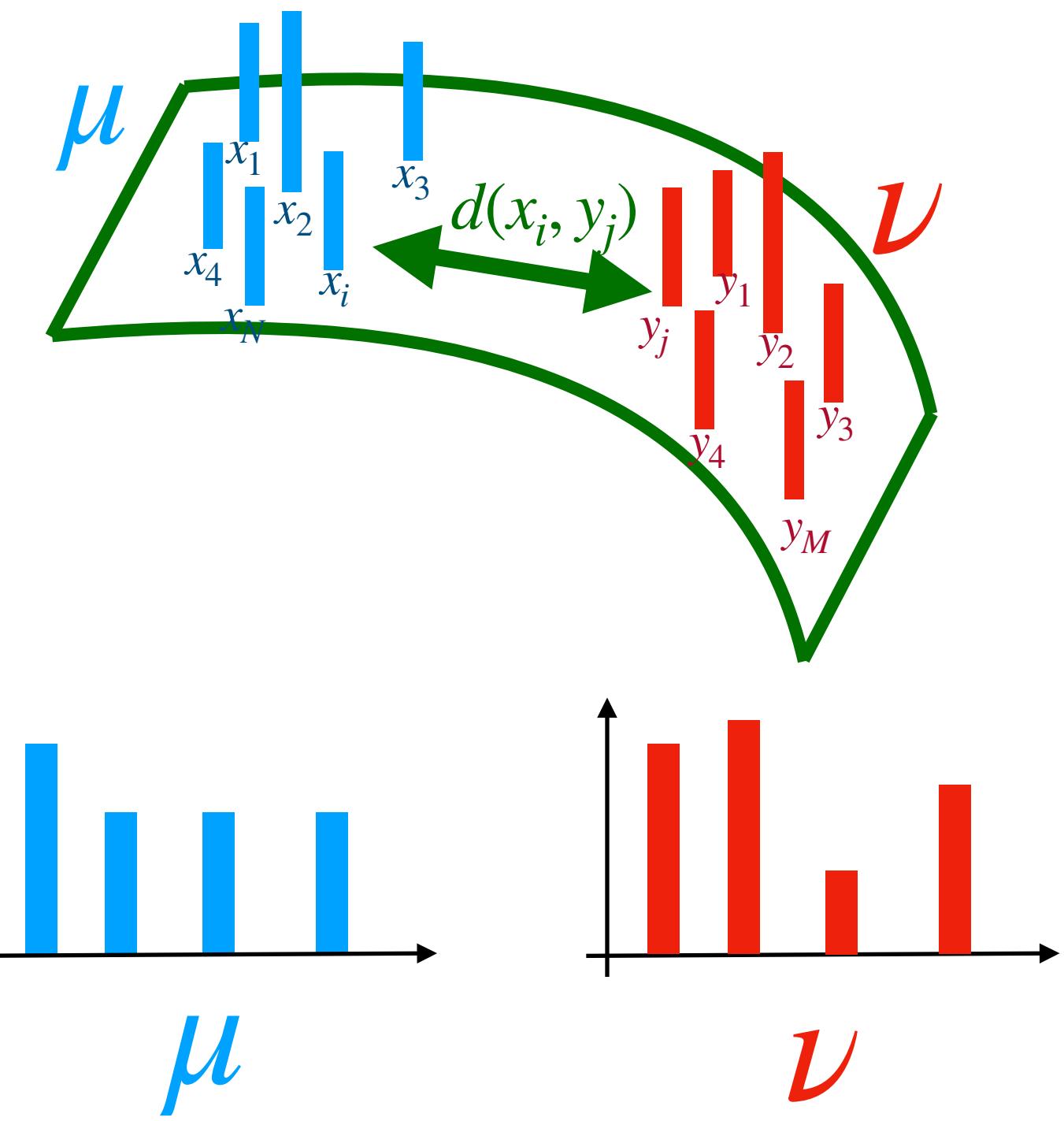
Hello, Chicago.  
If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [...].  
Yes we can!

Input Text



Neural Network

# High dimensional data

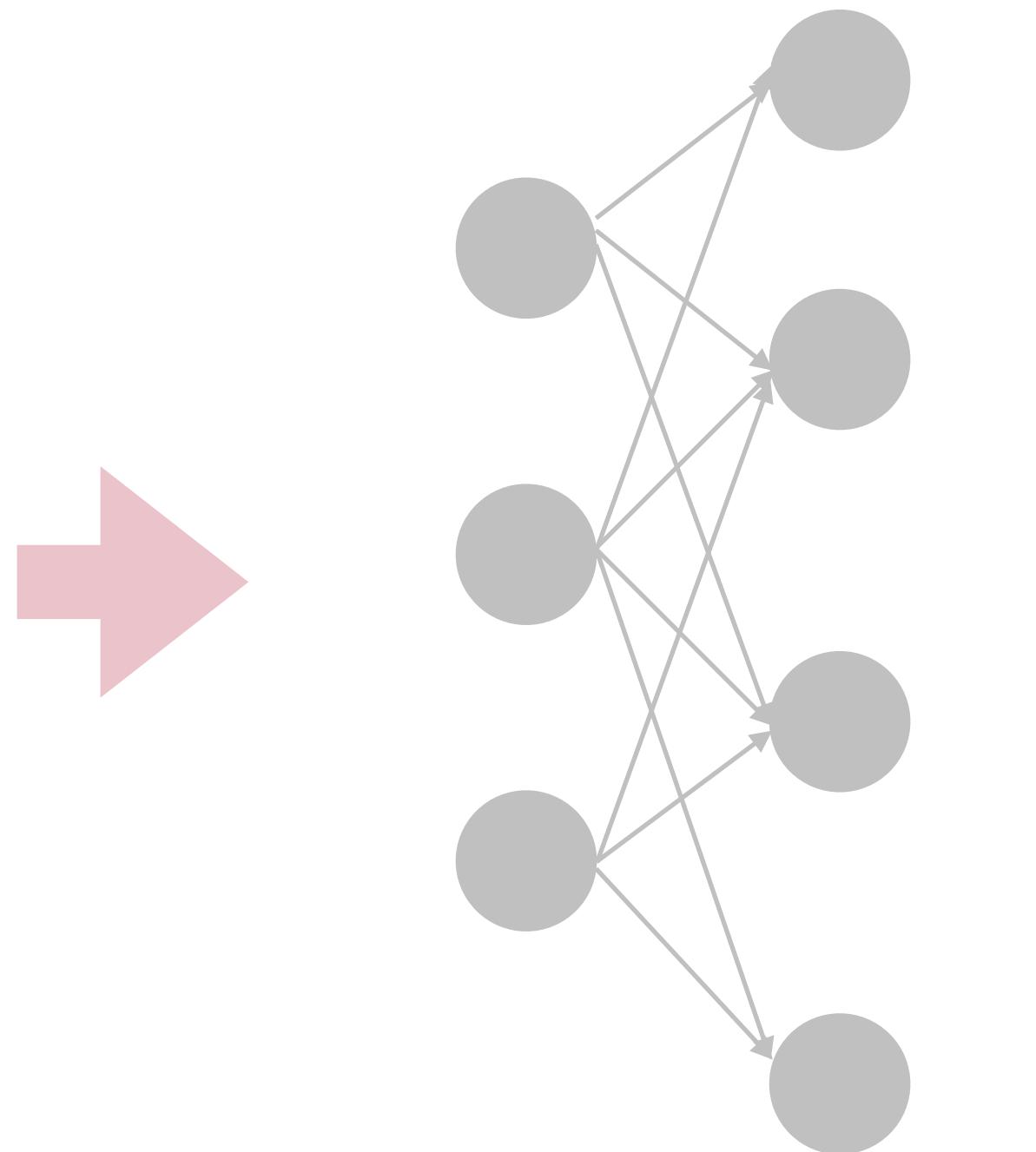


Soft Probabilities

# Information Measures

Hello, Chicago.  
If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [...].  
Yes we can!

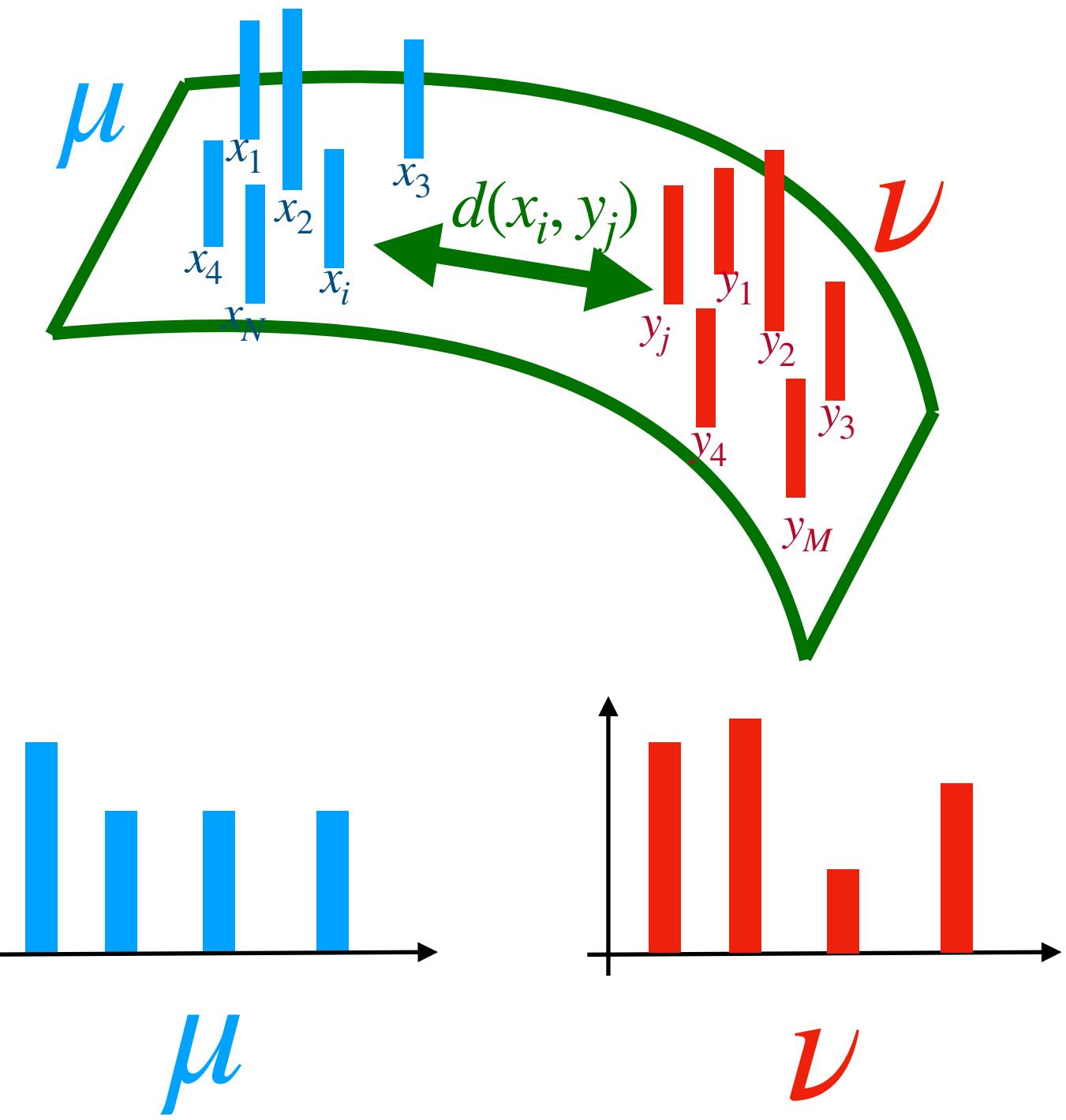
Input Text



Neural Network



# High dimensional data



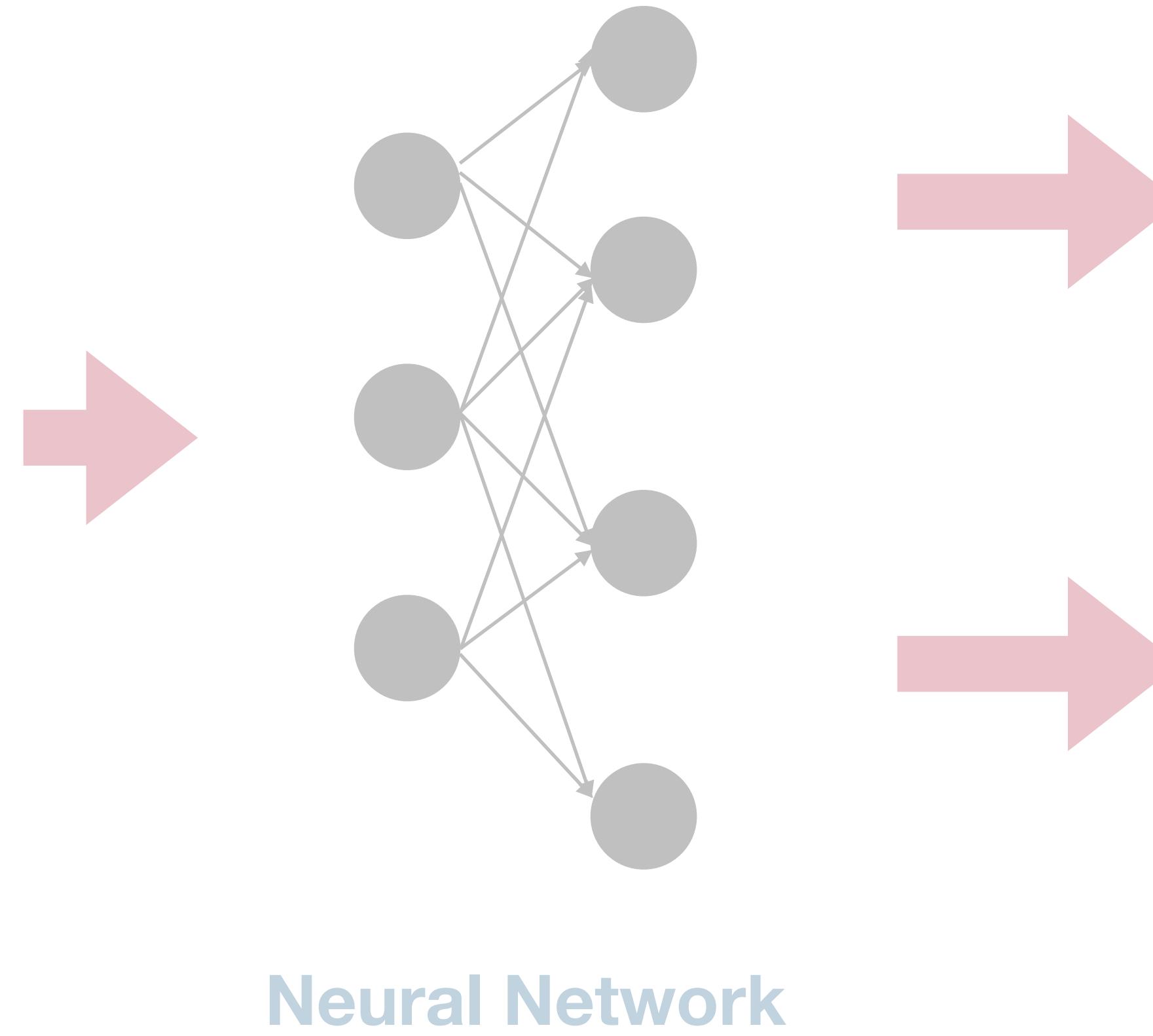
Soft Probabilities

When working with Neural Networks, we need to compare probability distributions

# Information Measures

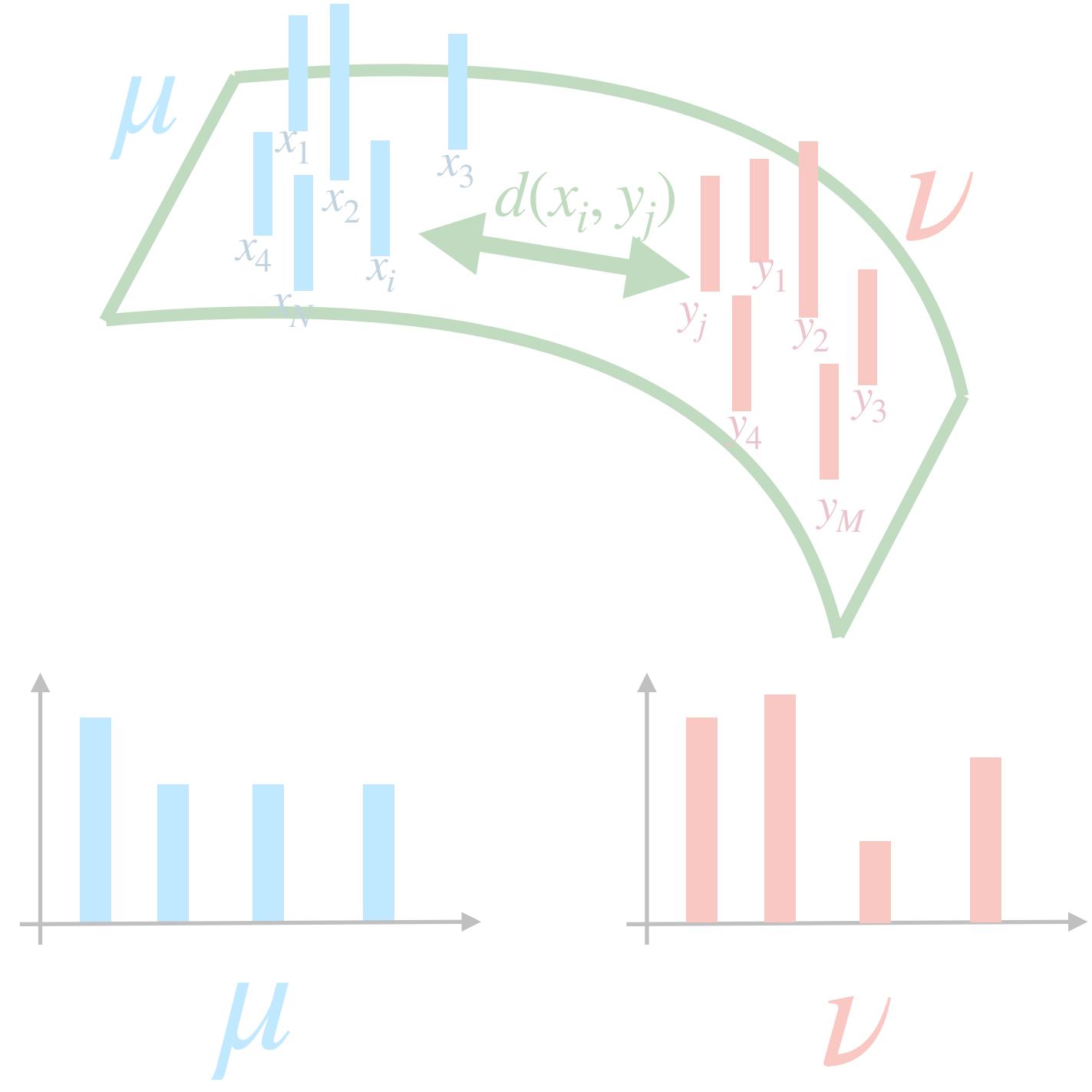
Hello, Chicago.  
If there is anyone out  
there who still doubts  
that America is a place  
where all things are  
possible, who still  
wonders if the dream of  
our founders is alive in  
our time, [....].  
Yes we can!

Input Text



Neural Network

High dimensional data



Soft Probabilities

When working with Neural Networks, we need to compare probability distributions

The measures of information are a tool measure this similarity/dissimilarity!



**Applications  
of  
Information**

# Information Measures

# Disentanglement

Cheng. et al EMNLP & ICML 2020

# Contrastive learning

van den Oord et al Preprint 2019

# Finetuning

Karimi Mahabad et al. ICLR 2021

# Pretraining

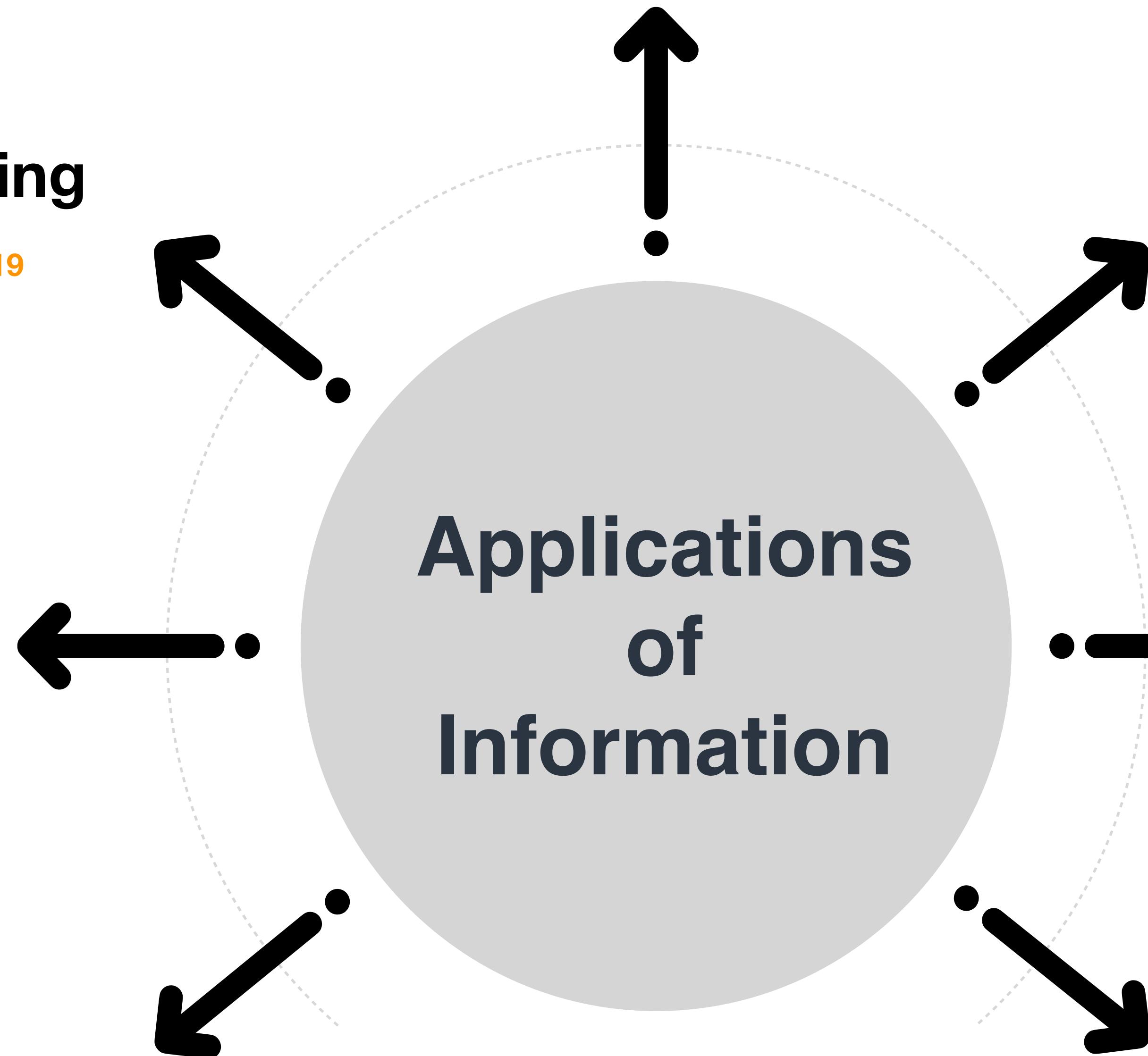
Clark et al ICLR 2020

# Diverse NLG

Li et al NAACL 2016

# Evaluation of NLG

Thompson et al. EMNLP 2020



# Conversational Agents

---

# Conversational Agents

## Definition

A conversational agent is any dialogue system that not only conducts natural language processing but also responds automatically using human language.



# Conversational Agents

## Definition

A conversational agent is any dialogue system that not only conducts natural language processing but also responds automatically using human language.

## Natural Language Understanding (NLU)

The conversation agent has to understand the user.

Spoken Conversations

Multimodal Conversation (audio, video, language)



# Conversational Agents

## Definition

A conversational agent is any dialogue system that not only conducts natural language processing but also responds automatically using human language.

## Natural Language Understanding (NLU)

The conversation agent has to understand the user.

Spoken Conversations

Multimodal Conversation (audio, video, language)

## Natural Language Generation (NLG)

The conversation agent has to produce a response.

Generate controlled text

Control text quality



# NLU Research Questions

---



# NLU Research Questions



Pretrained representations are trained on **written texts**:

**WORD2VEC, ... , BERT**

Mikolov et al. 2014 Delvin et al. 2019

**MI maximization principle**

Linkser 1988

# NLU Research Questions



Pretrained representations are trained on **written texts**:

**WORD2VEC, ... , BERT**

Mikolov et al. 2014 Delvin et al. 2019

**MI maximization principle**

Linkser 1988

**Conversational Agents are processing **spoken conversations**:**

**Conversation composed of spoken texts**

**Conversation are hierarchical**

Caller	Utterance
A	um, did you do through a public school system or private ?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference ?
B	Oh, yeah,
B	a big difference.
A	Like in what sense ?

# NLU Research Questions



Pretrained representations are trained on **written texts**:

**WORD2VEC, ... , BERT**

Mikolov et al. 2014 Delvin et al. 2019

**MI maximization principle**

Linkser 1988

**Conversational Agents are processing **spoken conversations**:**

**Conversation composed of spoken texts**

**Conversation are hierarchical**

Caller	Utterance
A	um, did you do through a public school system or private ?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference ?
B	Oh, yeah,
B	a big difference.
A	Like in what sense ?

**Conversational Agents can be exposed to multimodal conversations (e.g video, text, audio)**

**Pretrained representations are mono-modal**

# NLU Research Questions



Pretrained representations are trained on **written texts**:

WORD2VEC, ... , BERT

Mikolov et al. 2014 Delvin et al. 2019

MI maximization principle

Linkser 1988

Conversational Agents are processing **spoken conversations**:

Conversation composed of spoken texts

Conversation are hierarchical

Caller	Utterance
A	um, did you do through a public school system or private?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference?
B	Oh, yeah,
B	a big difference.
A	Like in what sense?

Conversational Agents can be exposed to **multimodal conversations** (e.g video, text, audio)

Pretrained representations are mono-modal

**How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?**

# NLG Research Questions

---



# NLG Research Questions

Conversational Agent are required to **generate text**.

**Well formed** (i.e grammatically correct, coherent)

Gatt et al 2018

**Informative**



# NLG Research Questions



Conversational Agent are required to generate text.

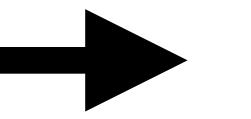
Well formed (i.e grammatically correct, coherent)

Gatt et al 2018

Informative

There is a need to control text generation both style and content.

I really hate these stupid cats



I love this wonderful cat!

Negative

Positive

# NLG Research Questions



Conversational Agents are required to generate text.

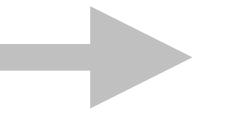
Well formed (i.e grammatically correct, coherent)

Gatt et al 2018

Informative

There is a need to control text generation both style and content.

I really hate these stupid cats



I love this wonderful cat!

Negative

Positive

Controlled NLG

# NLG Research Questions



Conversational Agents are required to generate text.

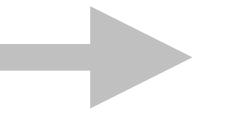
Well formed (i.e grammatically correct, coherent)

Gatt et al 2018

Informative

There is a need to control text generation both style and content.

I really hate these stupid cats



I love this wonderful cat!

Negative

Positive

Controlled NLG

Text is highly variable and there are multiple ways to express the same idea

It is sunny outside!



The weather looks great today!

# NLG Research Questions



Conversational Agent are required to generate text.

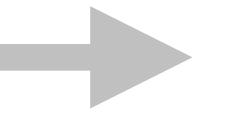
Well formed (i.e grammatically correct, coherent)

Gatt et al 2018

Informative

There is a need to control text generation both style and content.

I really hate these stupid cats



I love this wonderful cat!

Negative

Positive

Controlled NLG

Text is highly variable and there are multiple ways to express the same idea

It is sunny outside!



The weather looks great today!

NLG evaluation

# NLG Research Questions



Conversational Agent are required to generate text.

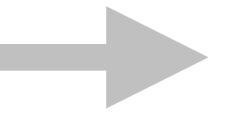
Well formed (i.e grammatically correct, coherent)

Gatt et al 2018

Informative

There is a need to control text generation both style and content.

I really hate these stupid cats



I love this wonderful cat!

Negative

Positive

Controlled NLG

Text is highly variable and there are multiple ways to express the same idea

It is sunny outside!



The weather looks great today!

NLG evaluation

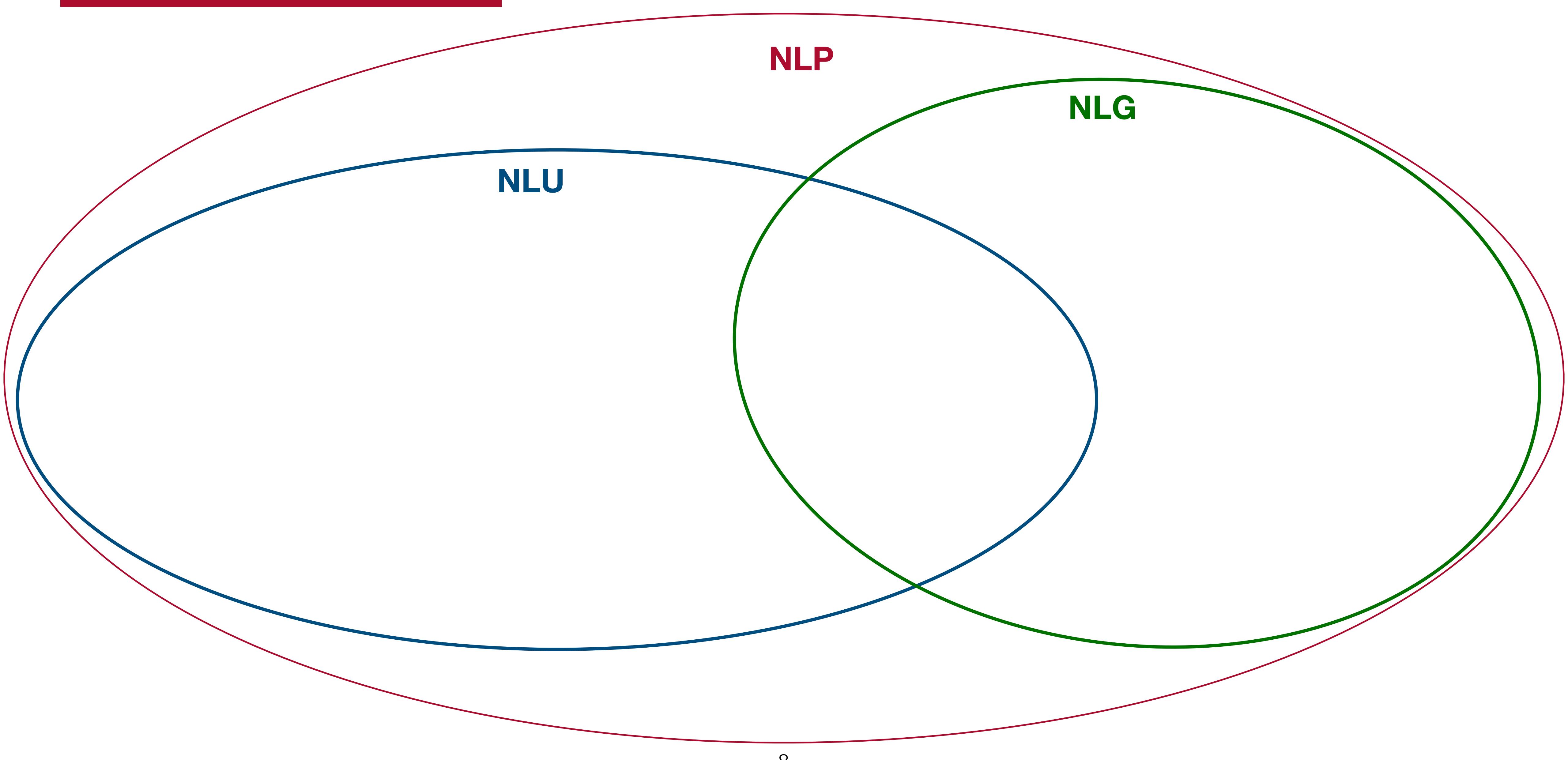
How to Use the Geometrical Properties of the Measures of Information to Generate and Evaluate Generated Text?

# Contributions

---

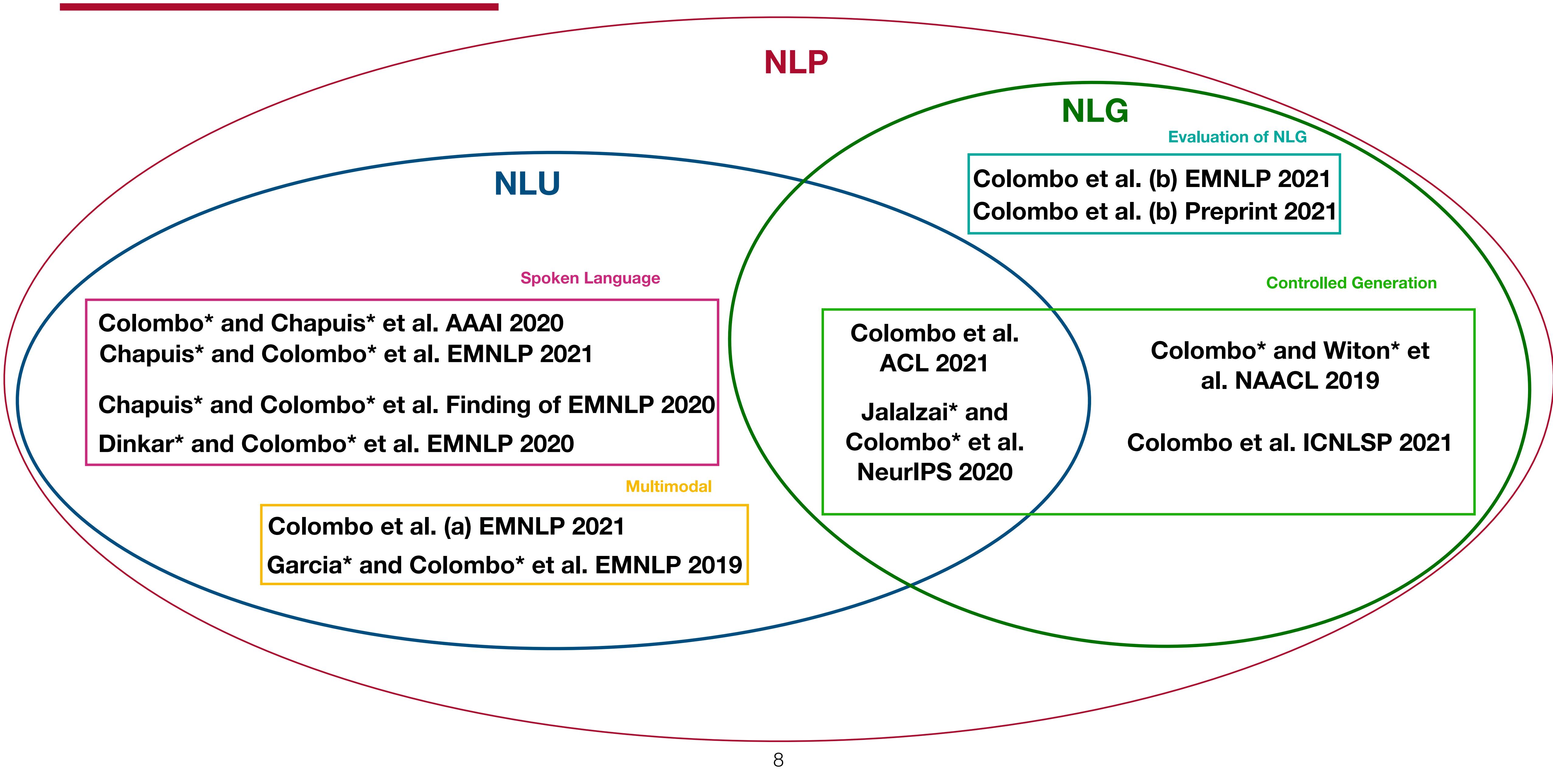
# Contributions

---

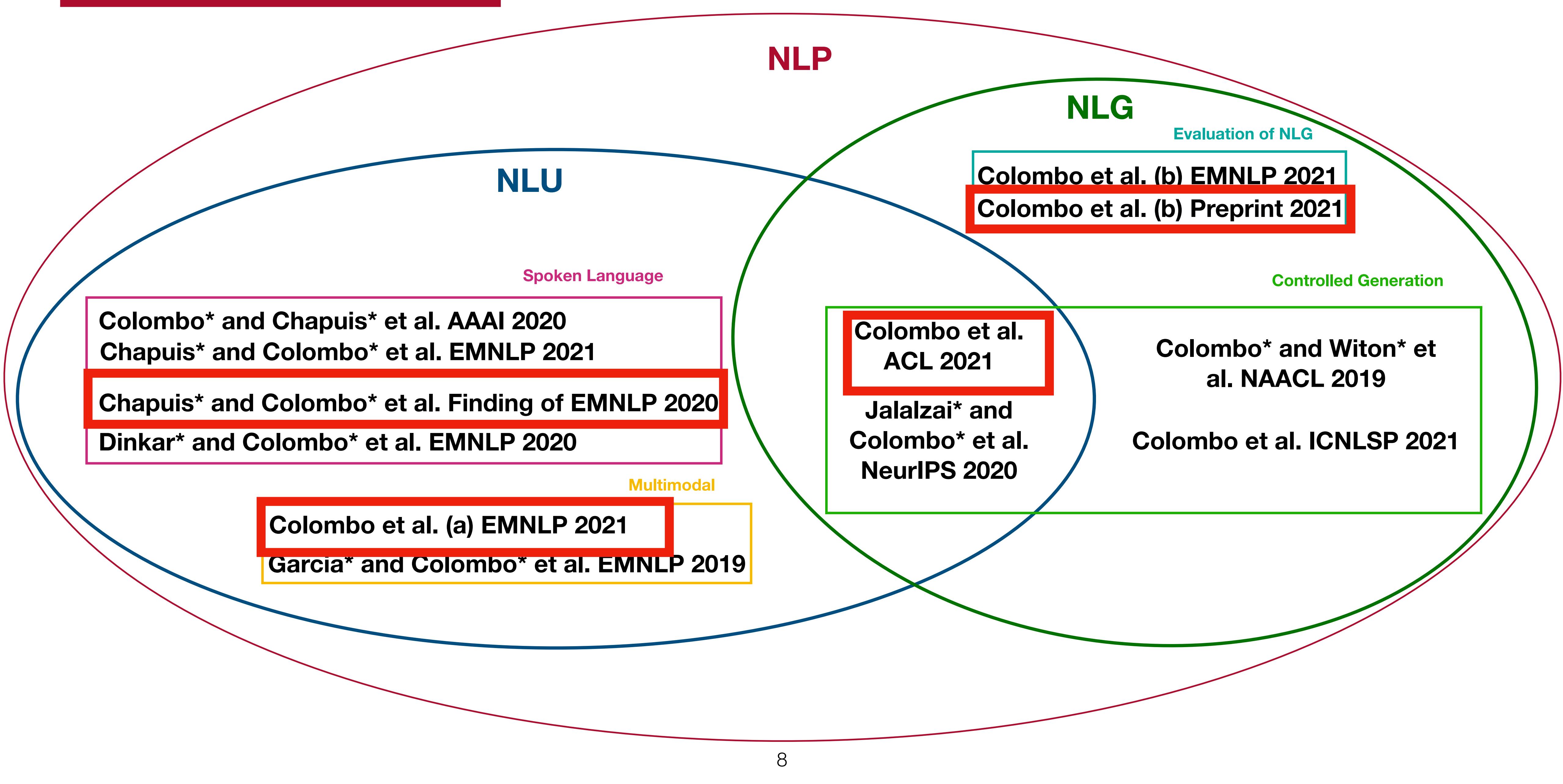


# Contributions

---



# Contributions



# Outline

---

# Outline

---

## 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

# Outline

---

## 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

### 1.1 Integrating Conversational Dimension in Pretrained Representations

Emile Chapuis\*, Pierre Colombo\*, Matteo Manica, Matthieu Labeau and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. Finding of EMNLP 2020

### 1.2 Including Multimodal Dimension in Representations of Spoken Transcripts

Pierre Colombo, Emile Chapuis, Matthieu Labeau and Chloé Clavel. Improving Multimodal fusion via Mutual Dependency Maximisation. (oral) EMNLP 2021

# Outline

---

## 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

### 1.1 Integrating Conversational Dimension in Pretrained Representations

Emile Chapuis\*, Pierre Colombo\*, Matteo Manica, Matthieu Labeau and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. Finding of EMNLP 2020

### 1.2 Including Multimodal Dimension in Representations of Spoken Transcripts

Pierre Colombo, Emile Chapuis, Matthieu Labeau and Chloé Clavel. Improving Multimodal fusion via Mutual Dependency Maximisation. (oral) EMNLP 2021

## 2. How to Use the Information Measures to Generate and Evaluate Generated Text?

# Outline

---

## 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

### 1.1 Integrating Conversational Dimension in Pretrained Representations

Emile Chapuis\*, Pierre Colombo\*, Matteo Manica, Matthieu Labeau and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. Finding of EMNLP 2020

### 1.2 Including Multimodal Dimension in Representations of Spoken Transcripts

Pierre Colombo, Emile Chapuis, Matthieu Labeau and Chloé Clavel. Improving Multimodal fusion via Mutual Dependency Maximisation. (oral) EMNLP 2021

## 2. How to Use the Information Measures to Generate and Evaluate Generated Text?

### 2.1 Learning to Disentangle Textual Representations and Attributes via MI

Pierre Colombo, Pablo Piantanida and Chloé Clavel. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations. (oral) ACL 2021

### 2.2 Automatic Text Generation Evaluation

Pierre Colombo, Chloé Clavel and Pablo Piantanida InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. Preprint 2021

# Outline

---

## 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

### 1.1 Integrating Conversational Dimension in Pretrained Representations

Emile Chapuis\*, Pierre Colombo\*, Matteo Manica, Matthieu Labeau and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. Finding of EMNLP 2020

### 1.2 Including Multimodal Dimension in Representations of Spoken Transcripts

Pierre Colombo, Emile Chapuis, Matthieu Labeau and Chloé Clavel. Improving Multimodal fusion via Mutual Dependency Maximisation. (oral) EMNLP 2021

## 2. How to Use the Information Measures to Generate and Evaluate Generated Text?

### 2.1 Learning to Disentangle Textual Representations and Attributes via MI

Pierre Colombo, Pablo Piantanida and Chloé Clavel. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations. (oral) ACL 2021

### 2.2 Automatic Text Generation Evaluation

Pierre Colombo, Chloé Clavel and Pablo Piantanida InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. Preprint 2021

## 3. Conclusions

# 1. How to Adapt the MI Maximization Principle to Learn Transcripts Representations with Conversational and Multimodal Dimensions?

# Mutual Information Maximization Principle

---

# Mutual Information Maximization Principle

---

**Mutual Information**

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

# Mutual Information Maximization Principle

---

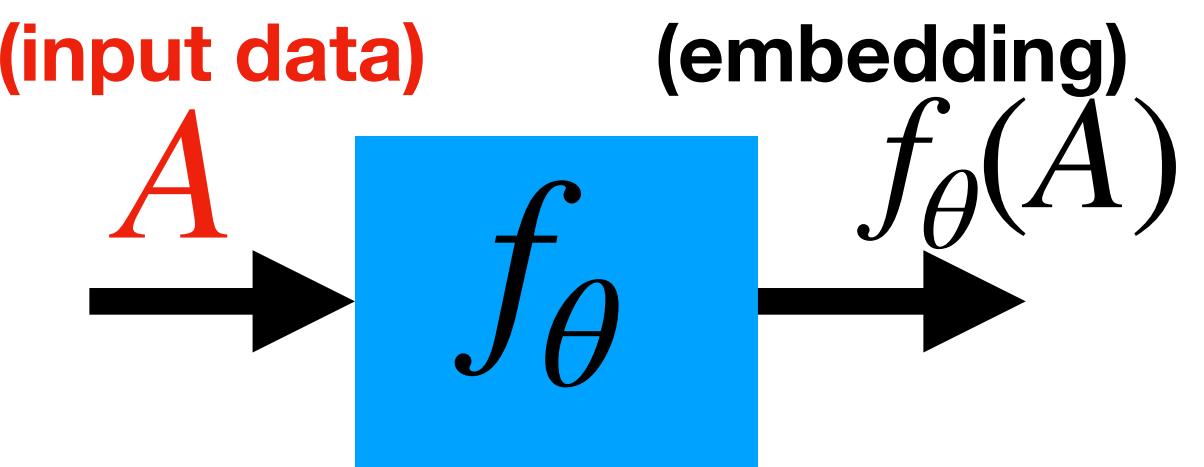
## Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v A while observing r.v B?

# Mutual Information Maximization Principle



Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v A while observing r.v B?

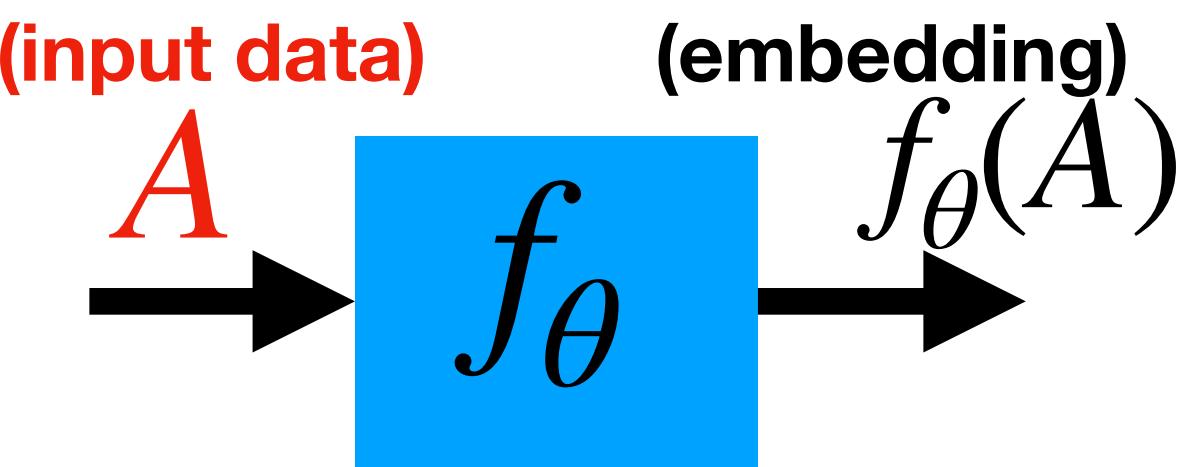
Unsupervised setting

Hjelm et al. 2019

Encoder  $f_\theta$

Input  $A$

# Mutual Information Maximization Principle



Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v  $A$  while observing r.v  $B$ ?

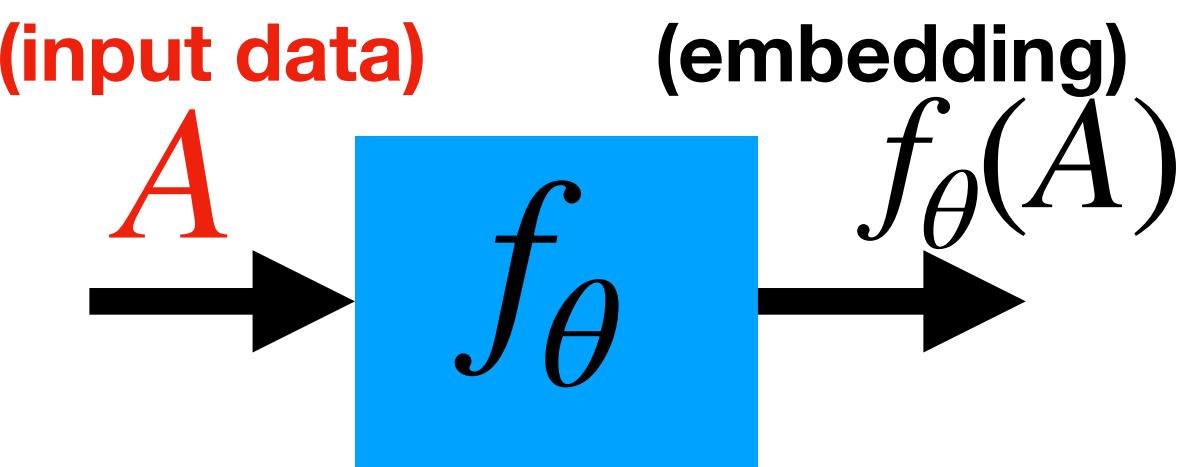
Unsupervised setting

Hjelm et al. 2019

Encoder  $f_\theta$   
Input  $A$

**Goal:** Learn a good feature representation  $f_\theta(A)$

# Mutual Information Maximization Principle



Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v A while observing r.v B?

Unsupervised setting

Hjelm et al. 2019

Encoder  $f_\theta$

Input  $A$

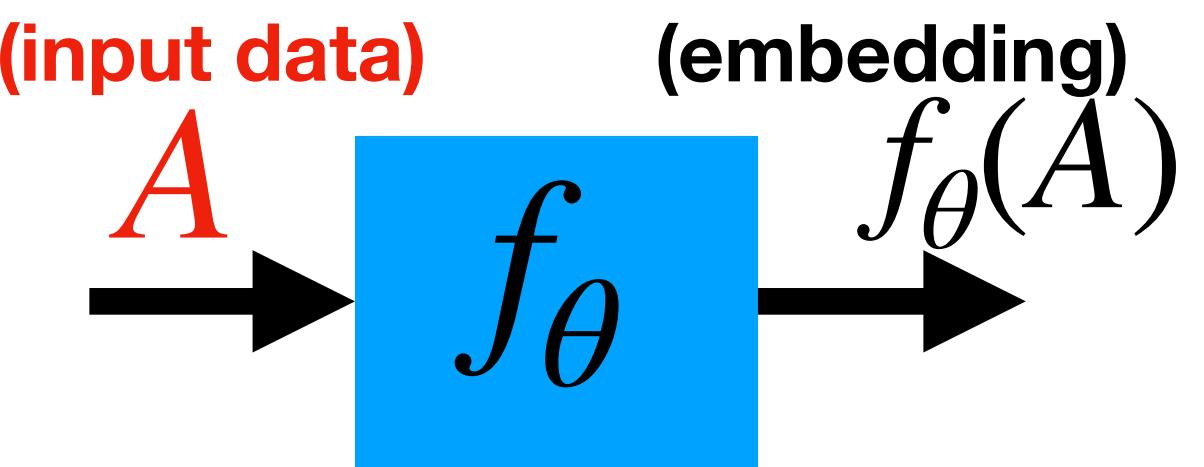
**Goal:** Learn a good feature representation  $f_\theta(A)$

InfoMax Principle

Linsker et al 1988

**Find**  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

# Mutual Information Maximization Principle



Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v A while observing r.v B?

Unsupervised setting

Hjelm et al. 2019

Encoder  $f_\theta$   
Input  $A$

**Goal:** Learn a good feature representation  $f_\theta(A)$

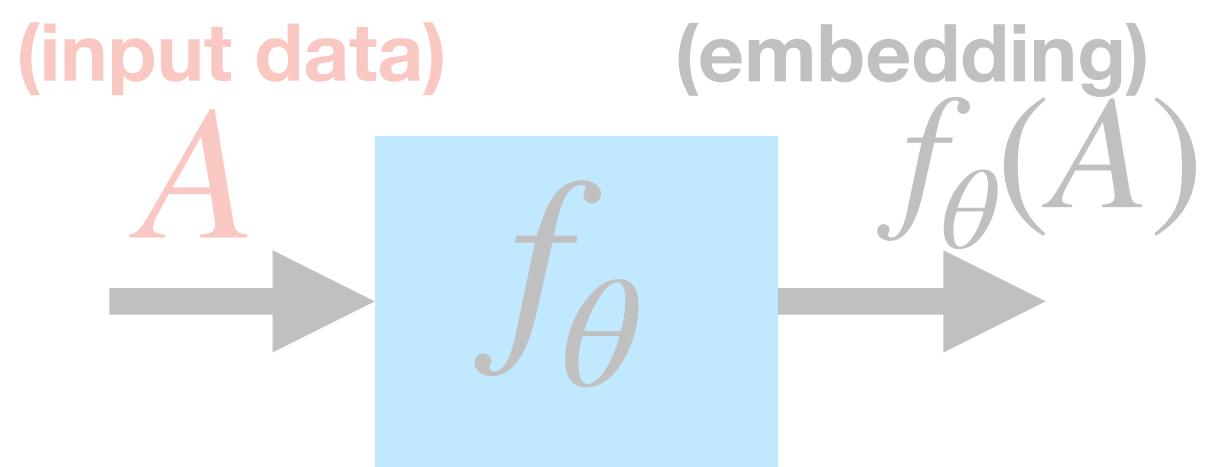
InfoMax Principle

Linsker et al 1988

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

A good feature map should have high mutual information with the input data.

# Mutual Information Maximization Principle



Mutual Information

$$I(A; B) \triangleq H(A) - H(A | B) \triangleq KL(p(a, b) || p(a) \times p(b))$$

Cover and Thomas 2007

How much can we learn from the r.v  $A$  while observing r.v  $B$ ?

Unsupervised setting

Hjelm et al. 2019

Encoder  $f_\theta$   
Input  $A$

Goal: Learn a good feature representation  $f_\theta(A)$

InfoMax Principle

Linsker et al 1988

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

A good feature map should have high mutual information with the input data.

**Estimating the MI is hard in high dimension!**

Pichler et al 2020

# InfoMax and Pretraining

---

Kong et al 2020

**InfoNCE is an estimator of Mutual Information.**

InfoNCE is an estimator of Mutual Information.

<b>InfoNCE</b>	$I(A; B) \geq \mathbf{E}_{p(A,B)} \left[ f_{\theta}(a, b) - \mathbf{E}_{q(\tilde{B})} \left( \log \sum_{\tilde{b} \in \tilde{B}} \exp f_{\theta}(a, \tilde{b}) \right) \right] - \log  \tilde{B} $	<b>Gutmann &amp; Hyvarinen, 2012</b> <b>Logeswaran &amp; Lee, 2018</b> <b>van den Oord et al., 2019</b>
<b>Cross entropy loss</b>	$\mathbf{E}_{p(A,B)} \left[ f_{\theta}(a, b) - \log \sum_{\tilde{b} \in \mathcal{V}} \exp f_{\theta}(a, \tilde{b}) \right] - \log  \mathcal{V} $	$\tilde{B} = \mathcal{B} \quad q(\tilde{B})$ $f_{\theta}(a, b) = g_{\psi}(b)^T g_{\phi}(a)$

InfoNCE is an estimator of Mutual Information.

InfoNCE	$I(A; B) \geq \mathbb{E}_{p(A,B)} \left[ f_\theta(a, b) - \mathbb{E}_{q(\tilde{B})} \left( \log \sum_{\tilde{b} \in \tilde{B}} \exp f_\theta(a, \tilde{b}) \right) \right] - \log  \tilde{B} $	<span style="color: orange;">Gutmann &amp; Hyvärinen, 2012</span> <span style="color: orange;">Logeswaran &amp; Lee, 2018</span> <span style="color: orange;">van den Oord et al., 2019</span>
Cross entropy loss	$\mathbb{E}_{p(A,B)} \left[ f_\theta(a, b) - \log \sum_{\tilde{b} \in \mathcal{V}} \exp f_\theta(a, \tilde{b}) \right] - \log  \mathcal{V} $	$\tilde{B} = \mathcal{B} \quad q(\tilde{B})$ $f_\theta(a, b) = g_\psi(b)^T g_\phi(a)$

Objective	$a$	$b$	$p(a, b)$	$g_\omega$	$g_\psi$
Skip-gram	word	word	word and its context	lookup	lookup
MLM	context	masked word	masked tokens probability	Transformer	lookup
NSP	sentence	sentence	(non-)consecutive sentences	Transformer	lookup
XLNet	context	masked word	factorization permutation	TXL++	lookup
DIM	context	masked $n$ -grams	sentence and its $n$ -grams	Transformer	not used

InfoNCE is an estimator of Mutual Information.

			Gutmann & Hyvarinen, 2012
InfoNCE	$I(A; B) \geq \mathbb{E}_{p(A,B)} \left[ f_\theta(a,b) - \mathbb{E}_{q(\tilde{B})} \left( \log \sum_{\tilde{b} \in \tilde{B}} \exp f_\theta(a, \tilde{b}) \right) \right] - \log  \tilde{B} $		Logeswaran & Lee, 2018
			van den Oord et al., 2019
Cross entropy loss	$\mathbb{E}_{p(A,B)} \left[ f_\theta(a,b) - \log \sum_{\tilde{b} \in \mathcal{V}} \exp f_\theta(a, \tilde{b}) \right] - \log  \mathcal{V} $	$\tilde{B} = \mathcal{B} \quad q(\tilde{B})$	
			$f_\theta(a,b) = g_\psi(b)^T g_\phi(a)$

No hierarchy!

Objective	$a$	$b$	$p(a, b)$	$g_\omega$	$g_\psi$
Skip-gram	word	word	word and its context	lookup	lookup
MLM	context	masked word	masked tokens probability	Transformer	lookup
NSP	sentence	sentence	(non-)consecutive sentences	Transformer	lookup
XLNet	context	masked word	factorization permutation	TXL++	lookup
DIM	context	masked $n$ -grams	sentence and its $n$ -grams	Transformer	not used

# 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!

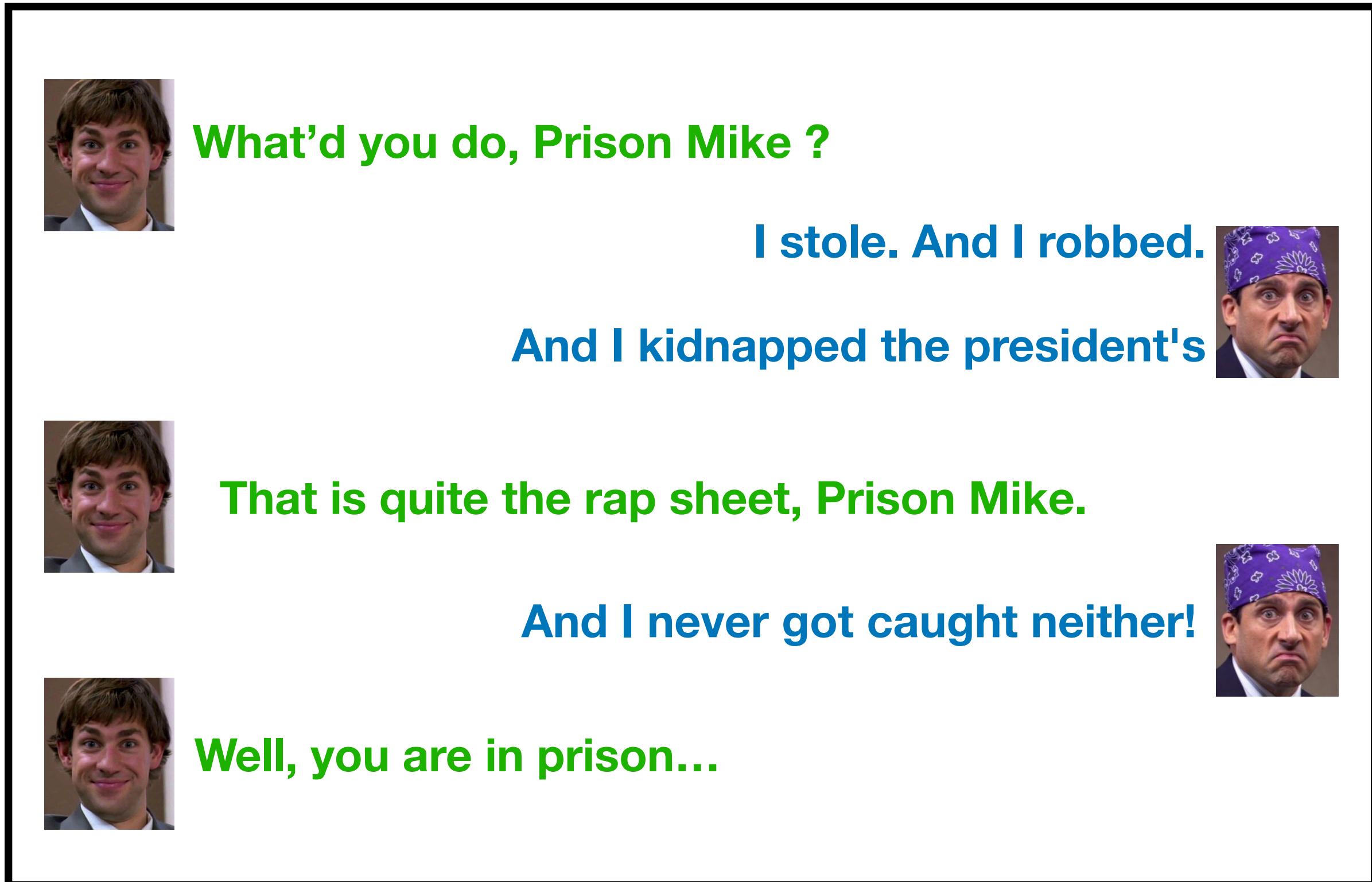


Well, you are in prison...

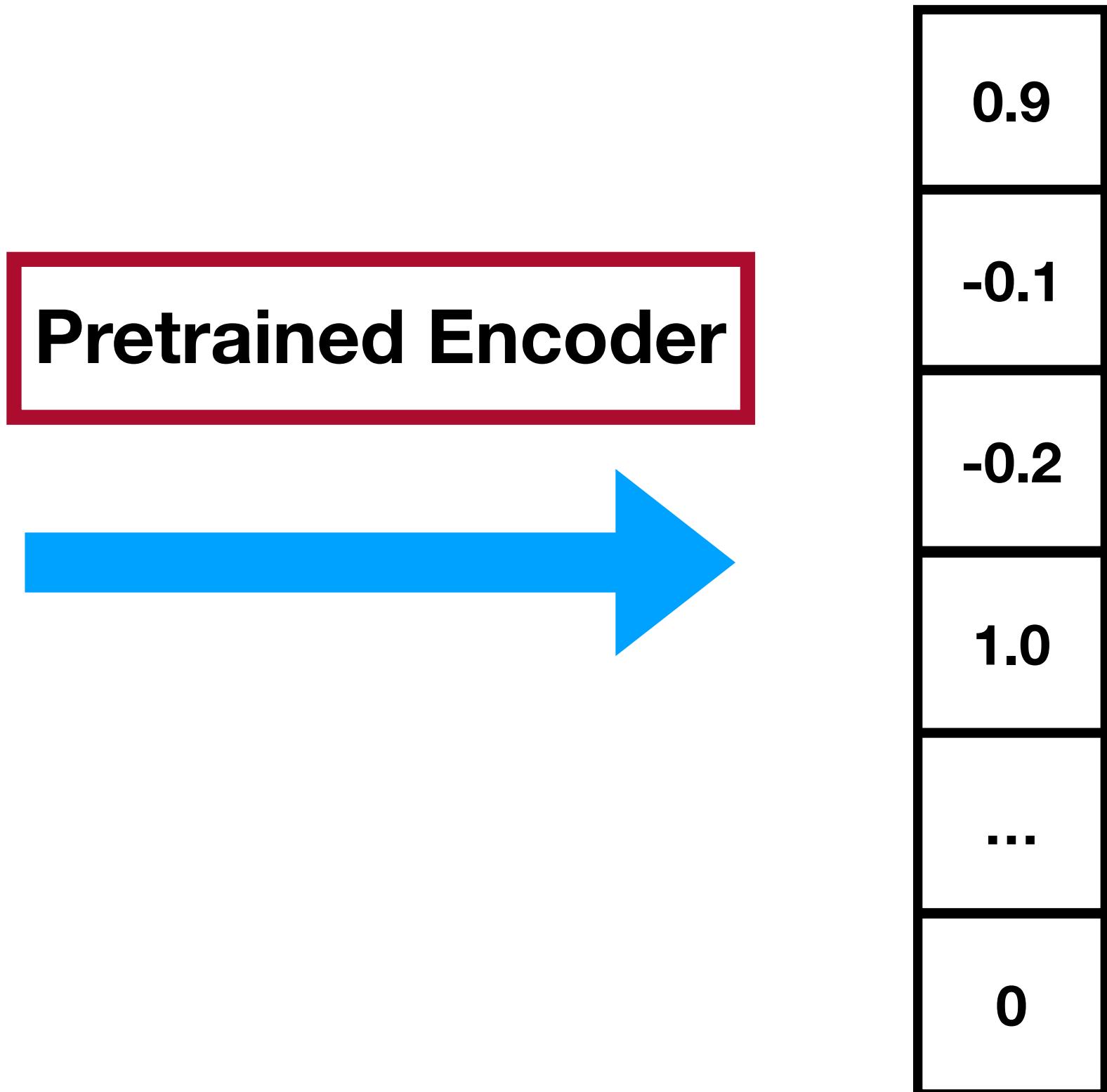
# Goals and Approach



## Input Conversation



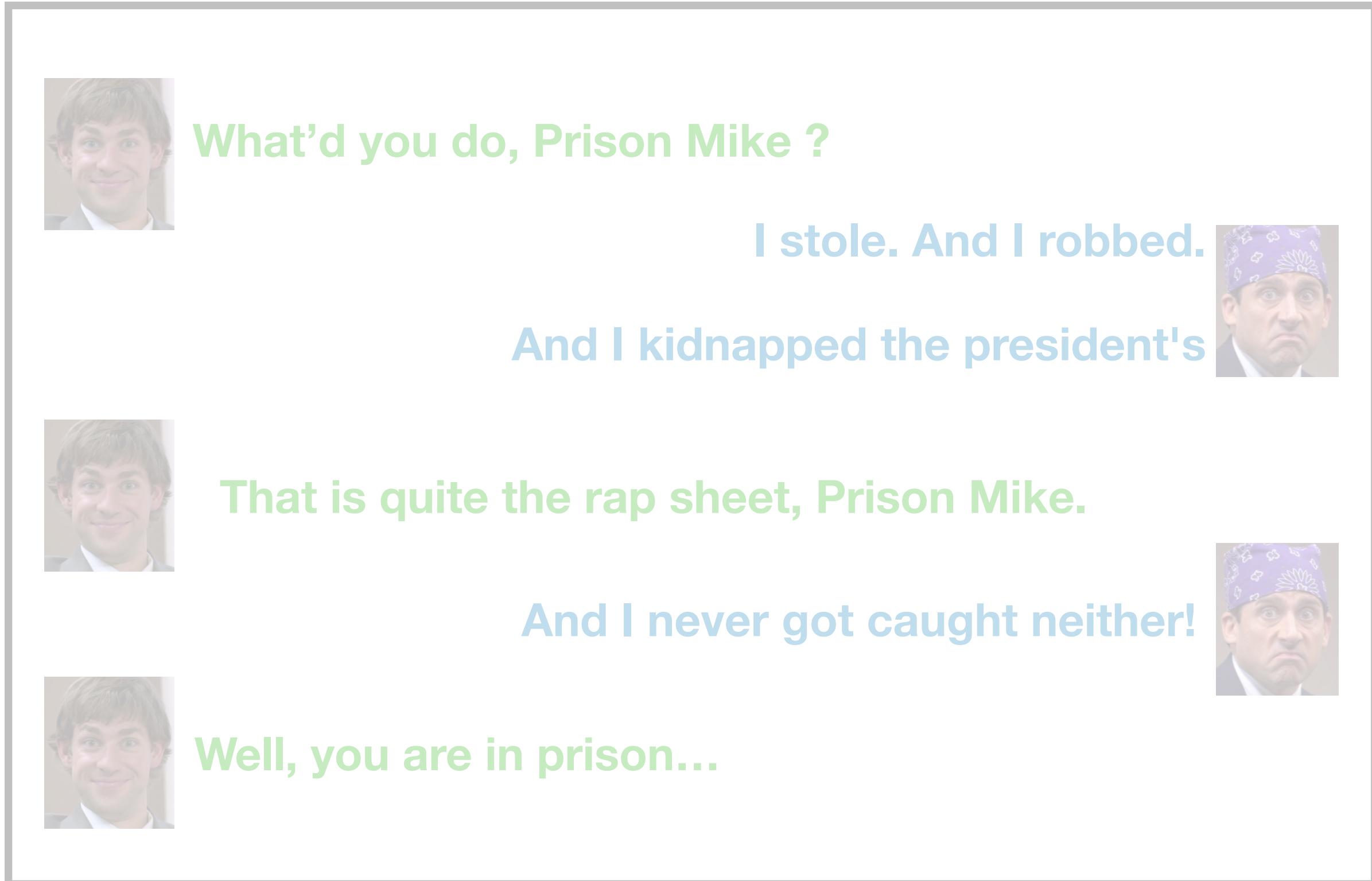
## Multi-purpose embedding



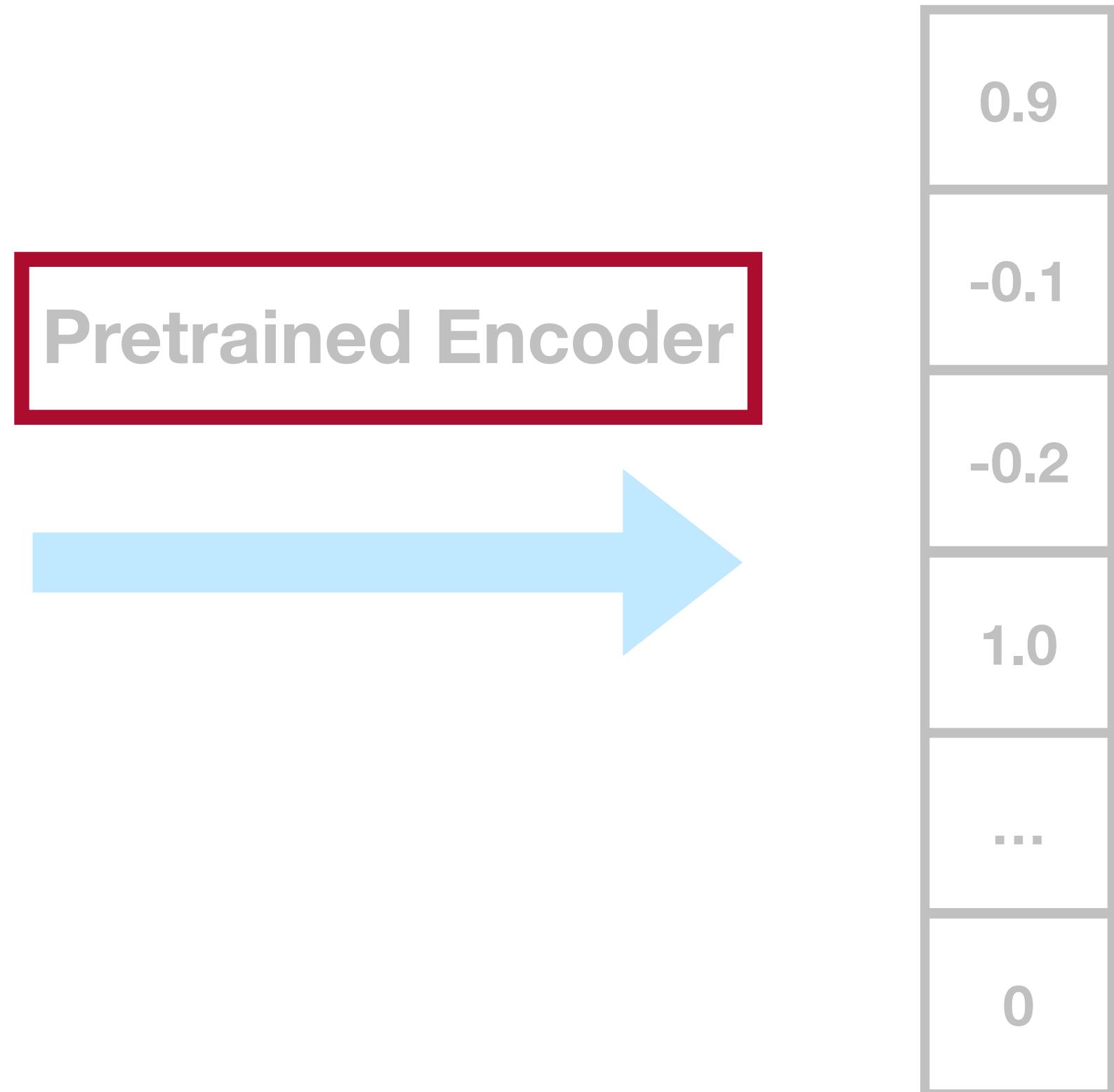
# Goals and Approach



## Input Conversation



## Multi-purpose embedding

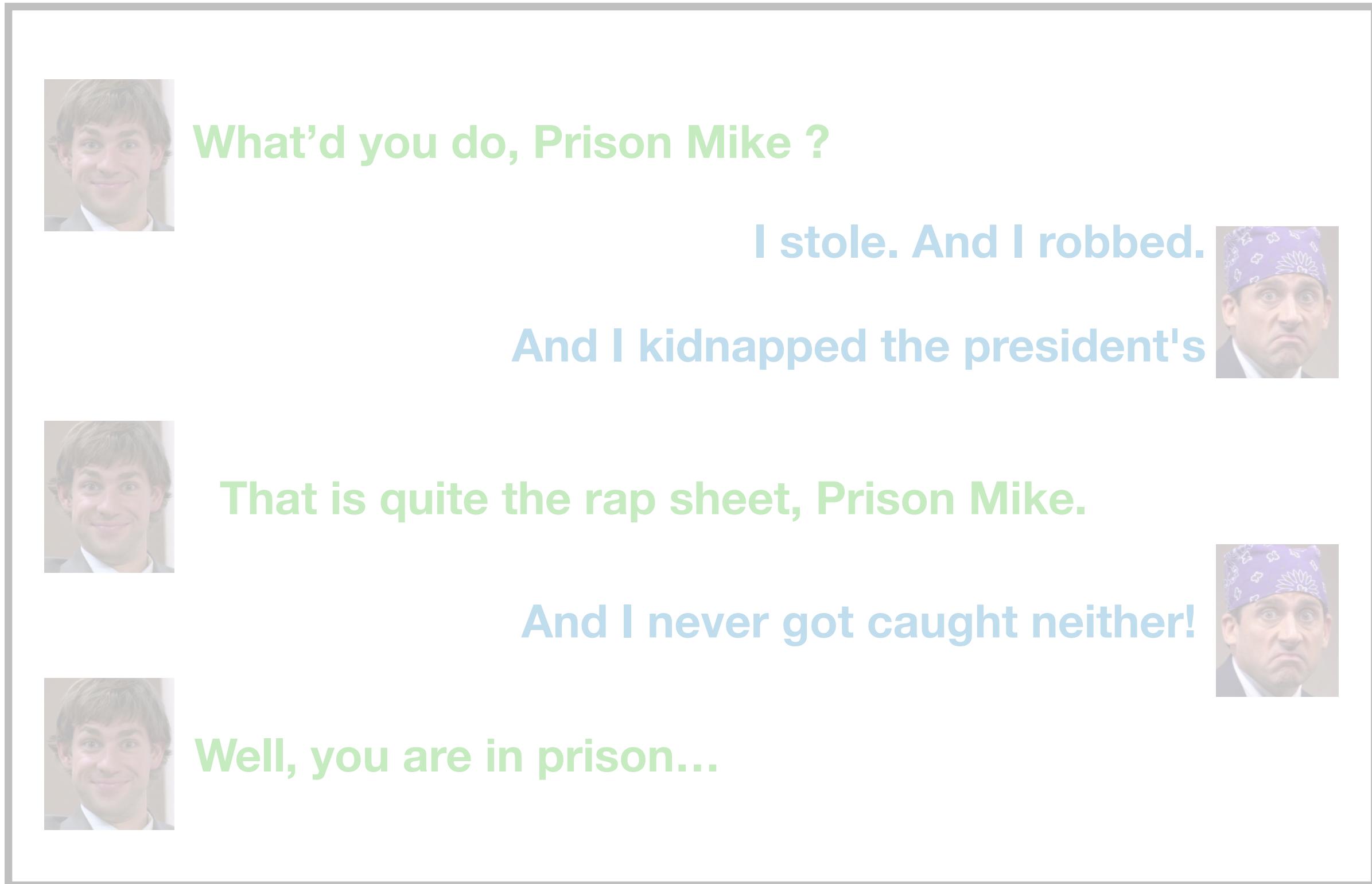


**Goal:** Learn a multi-purpose dialog embedding

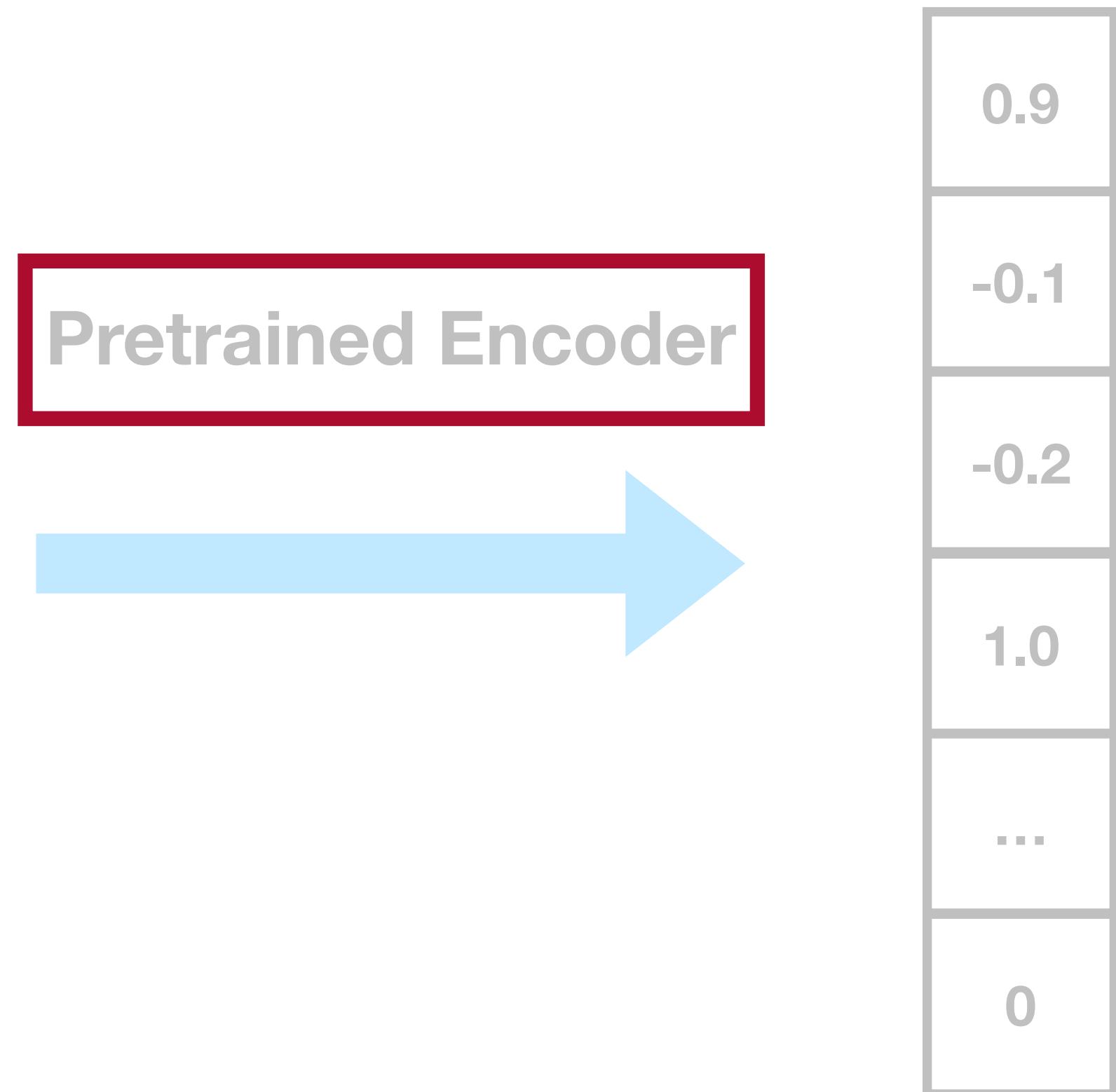


# Goals and Approach

## Input Conversation



## Multi-purpose embedding



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

**Propose new pertaining losses that are better suited for spoken conversations.**

# 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

Conversation  $C$

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

$u_1$



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

$u_1$



What'd you do, Prison Mike ?

I stole. And I robbed.

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

$u_3$



That is quite the rap sheet, Prison Mike.

$\omega_1^3 \quad \omega_2^3 \quad \omega_3^3 \quad \omega_4^3 \quad \omega_5^3 \quad \omega_6^3 \quad \omega_7^3 \quad \omega_8^3 \quad \omega_9^3 \quad \omega_{10}^3$

And I never got caught neither!

$u_5$



Well, you are in prison...

Conversation  $C$

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



$u_2$



$u_4$

## Formalisation

Conversation  $C$

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

$u_1$



What'd you do, Prison Mike ?

I stole. And I robbed.



$u_2$

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

$u_3$



That is quite the rap sheet, Prison Mike.

$$\omega_1^3 \quad \omega_2^3 \quad \omega_3^3 \quad \omega_4^3 \quad \omega_5^3 \quad \omega_6^3 \quad \omega_7^3 \quad \omega_8^3 \quad \omega_9^3 \quad \omega_{10}^3$$



$u_4$

And I never got caught neither!

$u_5$



Well, you are in prison...



Hierarchy is an important feature in dialog!

# Models

---

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

**Hierarchical Transformer Encoder:**  $f_{\theta}^u$  **and**  $f_{\theta}^d$

# Models

---

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

**Hierarchical Transformer Encoder:**  $f_\theta^u$  and  $f_\theta^d$

$$\begin{aligned}\mathcal{E}_{u_i} &= f_\theta^u(\omega_1^i, \dots, \omega_{|u_i|}^i), \\ \mathcal{E}_{C_j} &= f_\theta^d(\mathcal{E}_{u_1}, \dots, \mathcal{E}_{u_{|C_j|}}),\end{aligned}$$

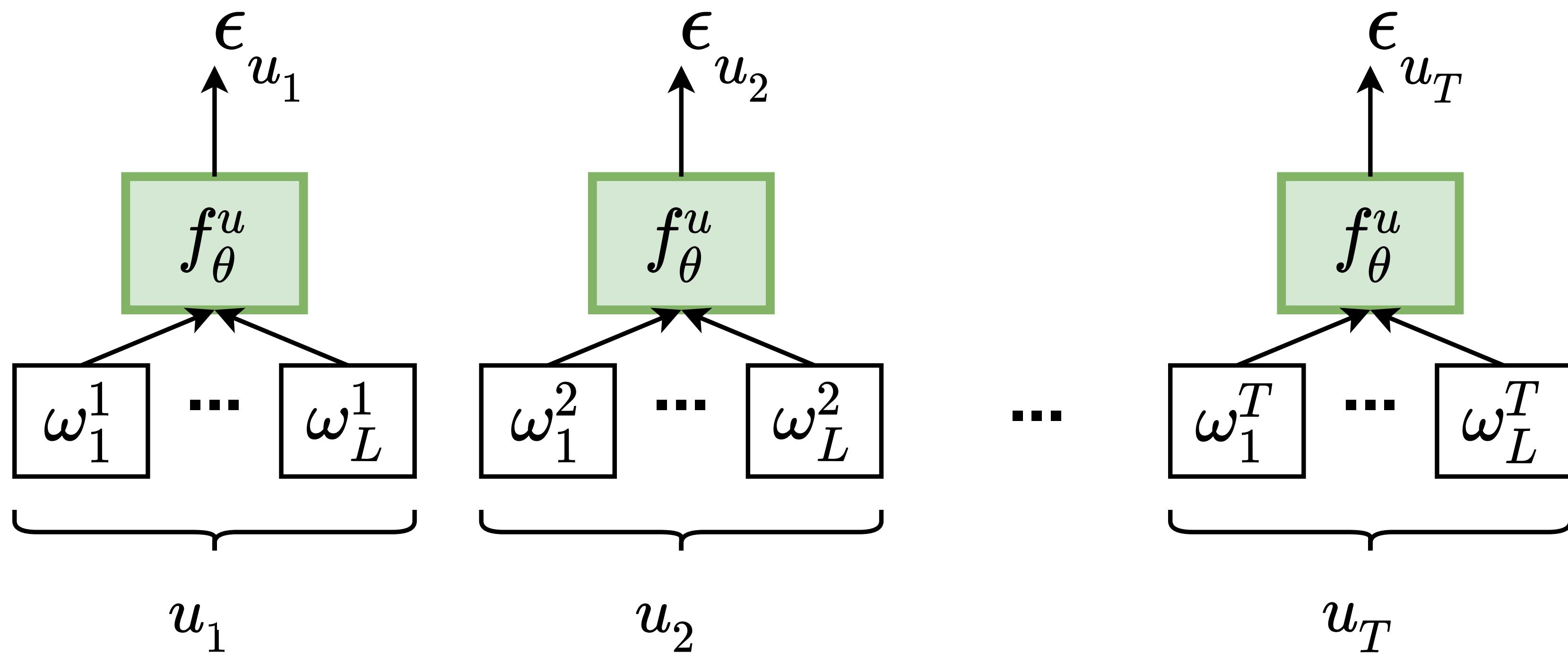
# Models

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

**Hierarchical Transformer Encoder:**  $f_\theta^u$  **and**  $f_\theta^d$

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

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



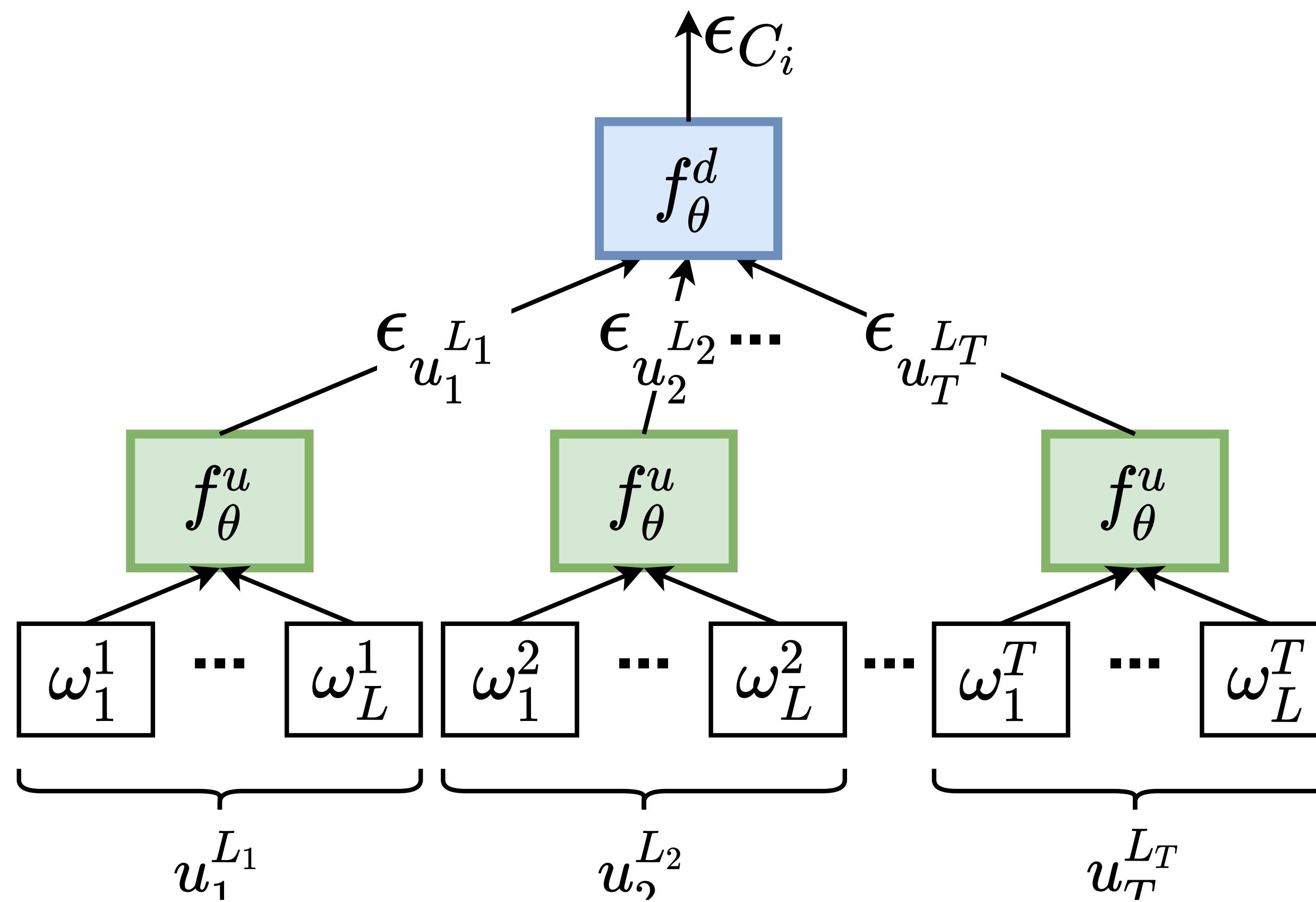
# Models

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

**Hierarchical Transformer Encoder:**  $f_\theta^u$  and  $f_\theta^d$

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

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



## Utterance Level Pretraining

---

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

## Utterance Level Pretraining

---

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

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

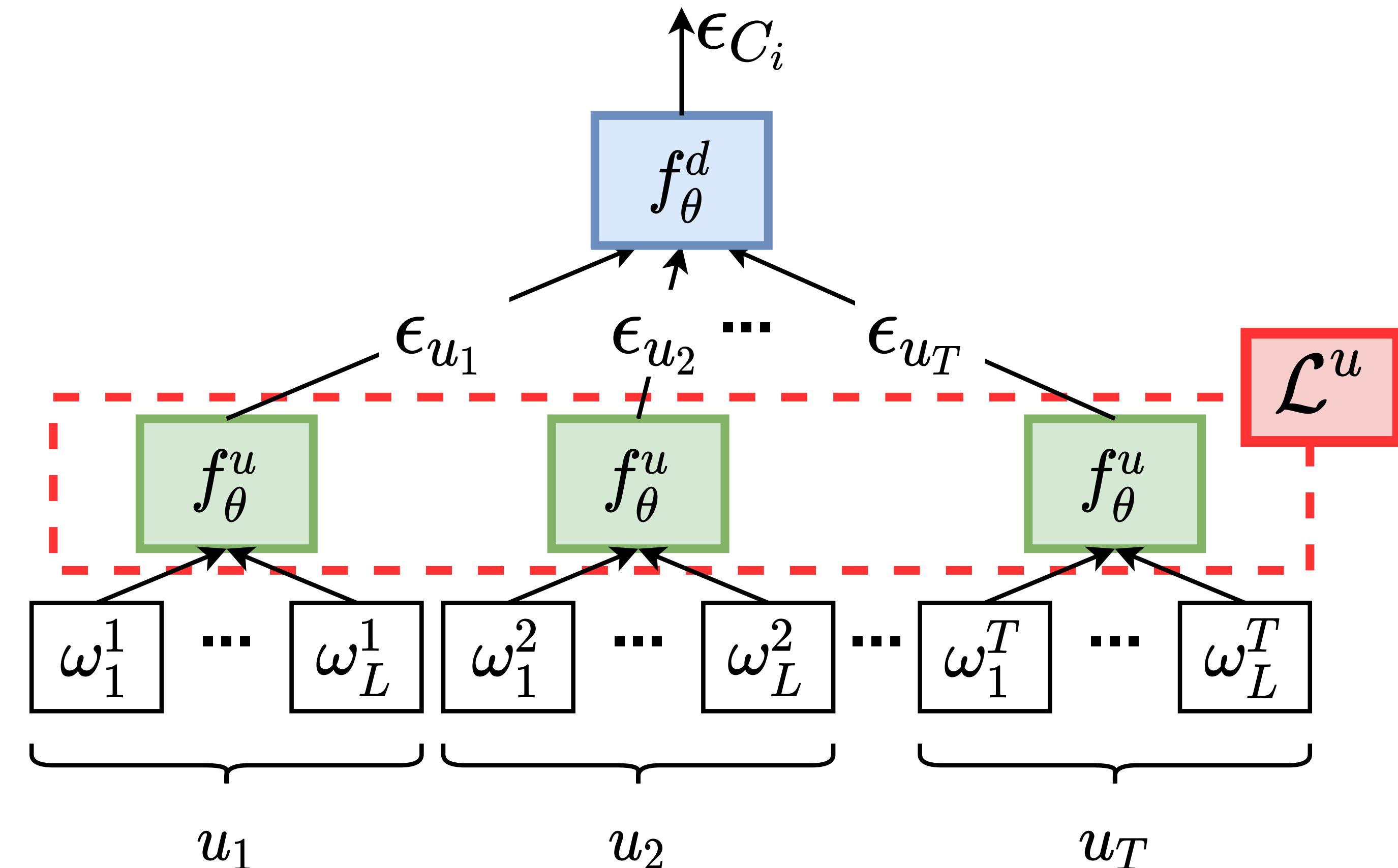
- Masked Word Pretraining

## 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



# Utterance Level Pretraining

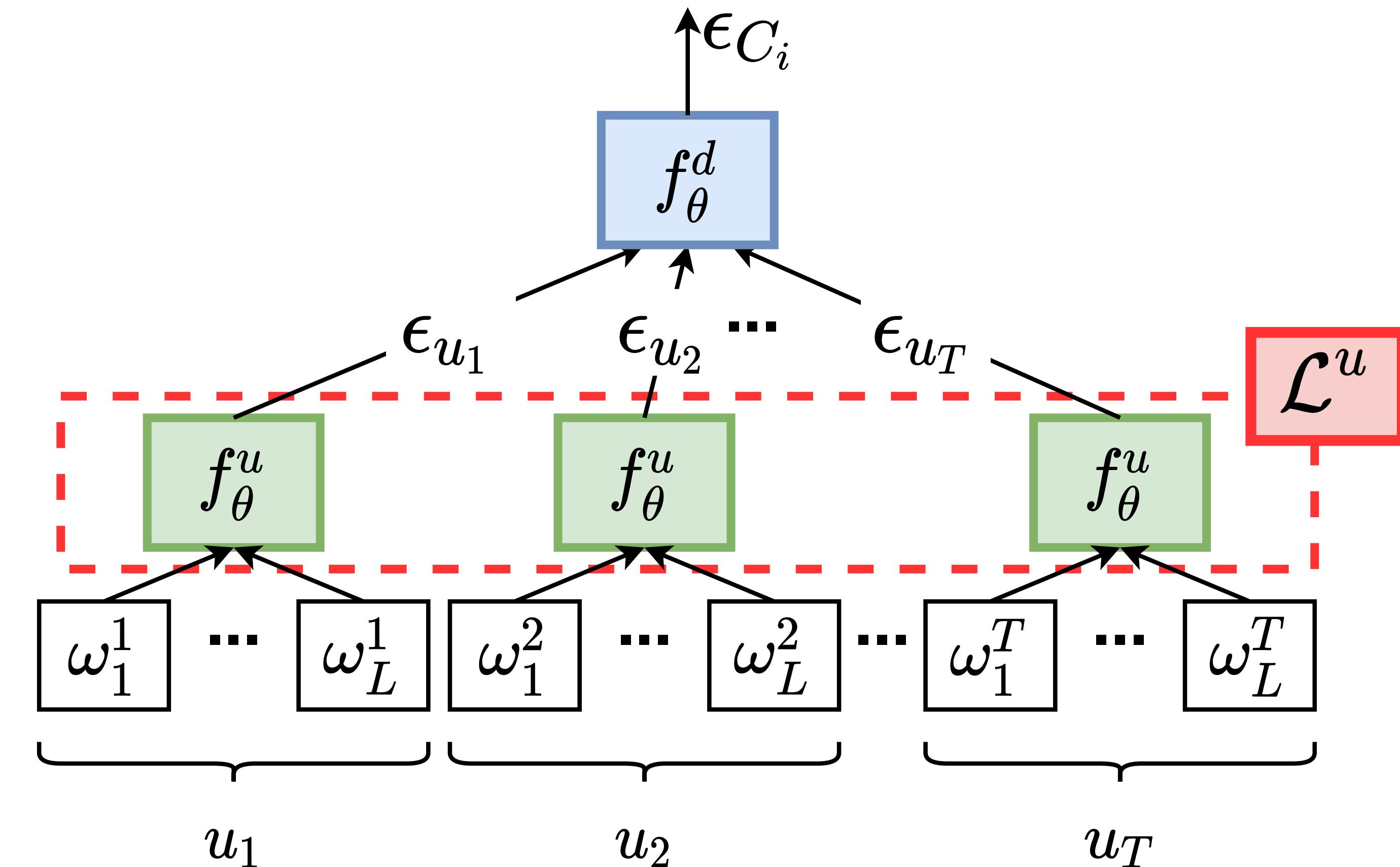
$$\mathcal{L}(\theta) = \lambda_u \mathcal{L}^u(\theta) + \lambda_d \mathcal{L}^d(\theta)$$

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

- Masked Word Pretraining

## Goal:

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



## Utterance Level Pretraining

---

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

## Utterance Level Pretraining

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

## Masked Utterance Modelling (MUM)

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

$u_3$



That|is|quite|the|rap|sheet|,|Prison|Mike|.

$\omega_1^4 \quad \omega_2^4 \quad \omega_3^4 \quad \omega_4^4 \quad \omega_5^4 \quad \omega_6^4 \quad \omega_7^4 \quad \omega_8^4 \quad \omega_9^4 \quad \omega_{10}^4$

# Utterance Level Pretraining

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

## Masked Utterance Modelling (MUM)

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

Set of masked tokens                          Set of masked indices

$u_3$



That|is|quite|the|rap|sheet|,|Prison|Mike|.

$\omega_1^4 \ \omega_2^4 \ \omega_3^4 \ \omega_4^4 \ \omega_5^4 \ \omega_6^4 \ \omega_7^4 \ \omega_8^4 \ \omega_9^4 \ \omega_{10}^4$

# Utterance Level Pretraining

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

## Masked Utterance Modelling (MUM)

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

Set of masked tokens    Set of masked indices

$u_3$



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

$\omega_1^4 \quad \omega_2^4 \quad \omega_3^4 \quad \omega_4^4 \quad \omega_5^4 \quad \omega_6^4 \quad \omega_7^4 \quad \omega_8^4 \quad \omega_9^4 \quad \omega_{10}^4$



# Utterance Level Pretraining

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

## Masked Utterance Modelling (MUM)

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

Set of masked tokens                                  Set of masked indices

$u_3$



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

$\omega_1^4 \ \omega_2^4 \ \omega_3^4 \ \omega_4^4 \ \omega_5^4 \ \omega_6^4 \ \omega_7^4 \ \omega_8^4 \ \omega_9^4 \ \omega_{10}^4$



## Dialog Level Pretraining

---

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

## **Dialog Level Pretraining**

---

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

### **Dialog Level Pretraining:** $\mathcal{L}^d(\theta)$

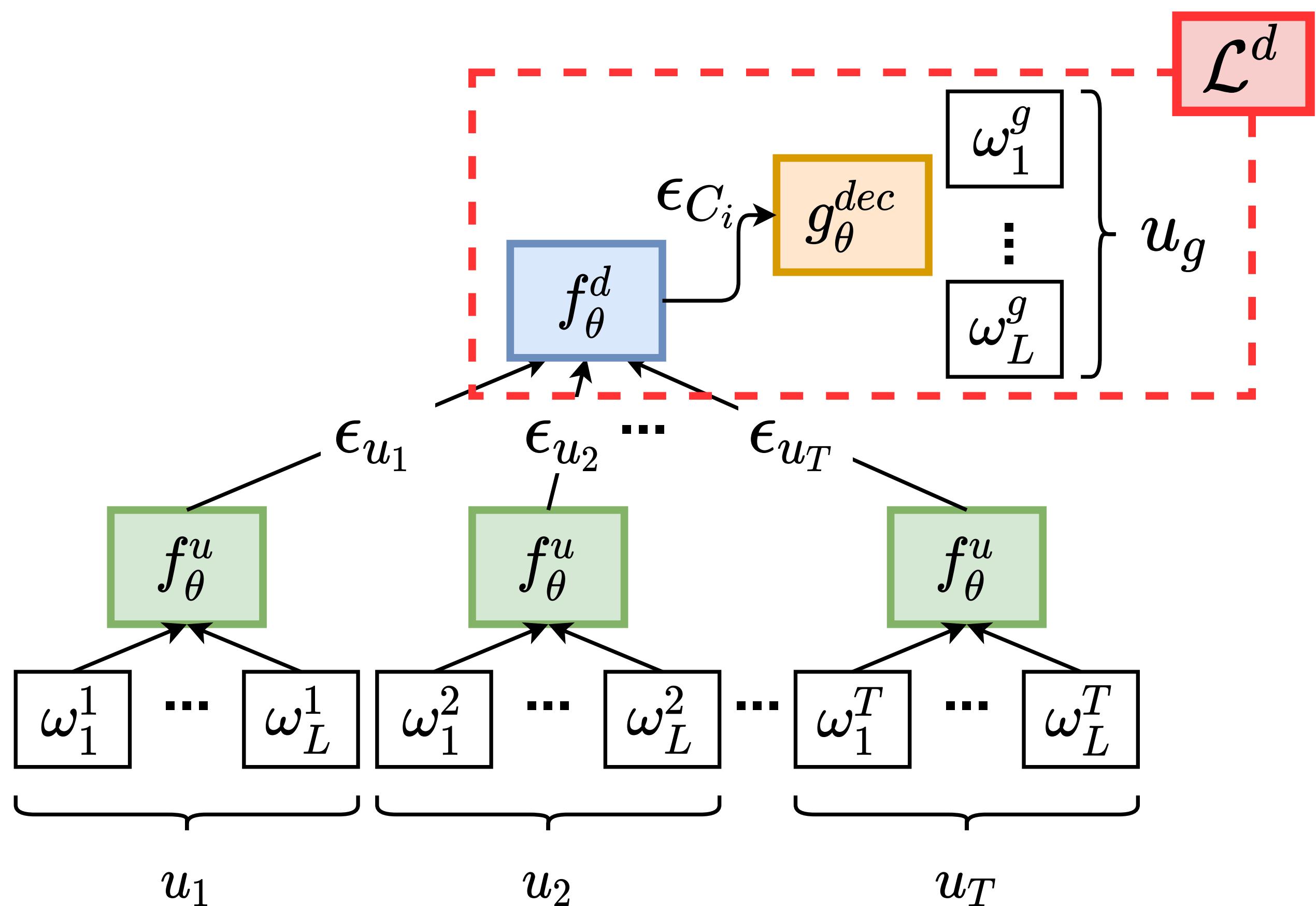
- **Masked Sequence Generation**
- Add a *causal decoder*

# Dialog Level Pretraining

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

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

- Masked Sequence Generation
- Add a causal decoder



# Dialog Level Pretraining

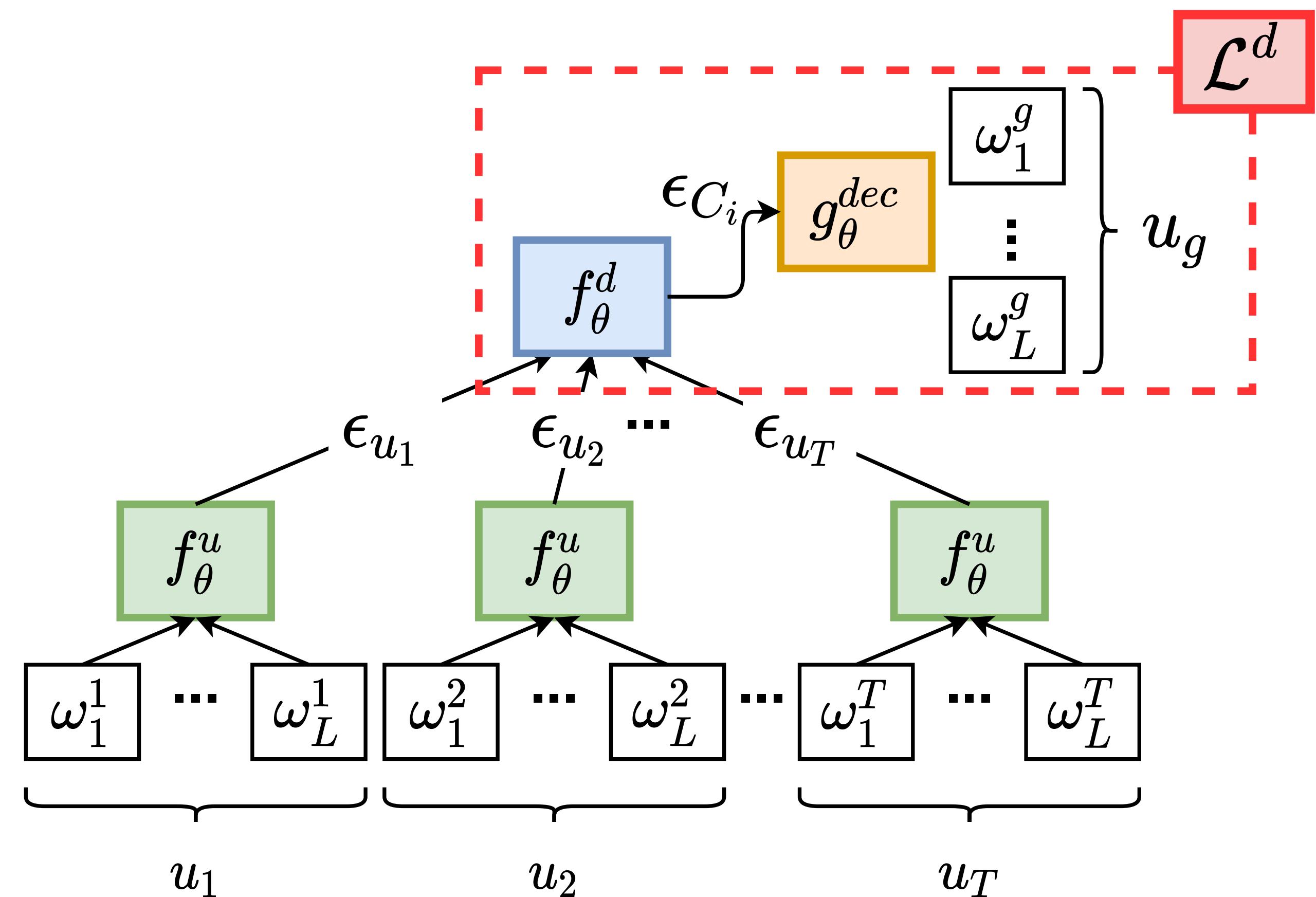
$$\mathcal{L}(\theta) = \lambda_u \mathcal{L}^u(\theta) + \lambda_d \mathcal{L}^d(\theta)$$

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

- 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

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

### Masked Sequence Generation (MSG)

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

## Dialog Level Pretraining

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

### Masked Sequence Generation (MSG)

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

# Dialog Level Pretraining

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

## Masked Sequence Generation (MSG)

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

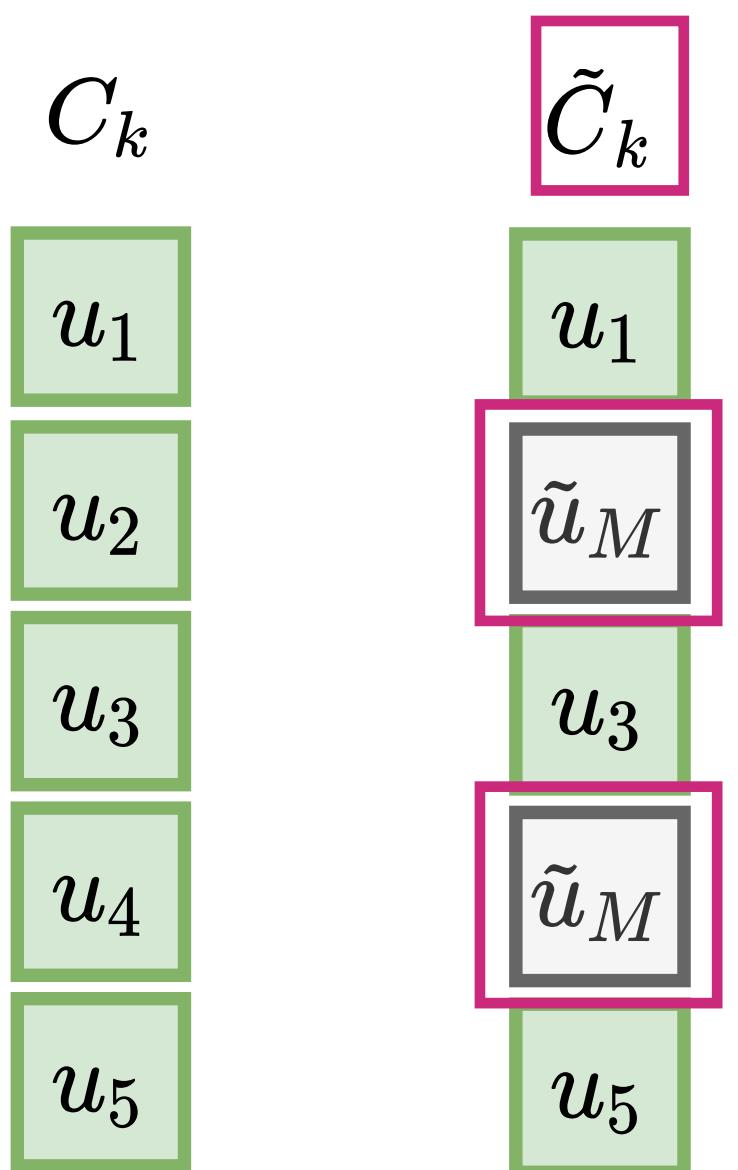


# Dialog Level Pretraining

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

## Masked Sequence Generation (MSG)

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

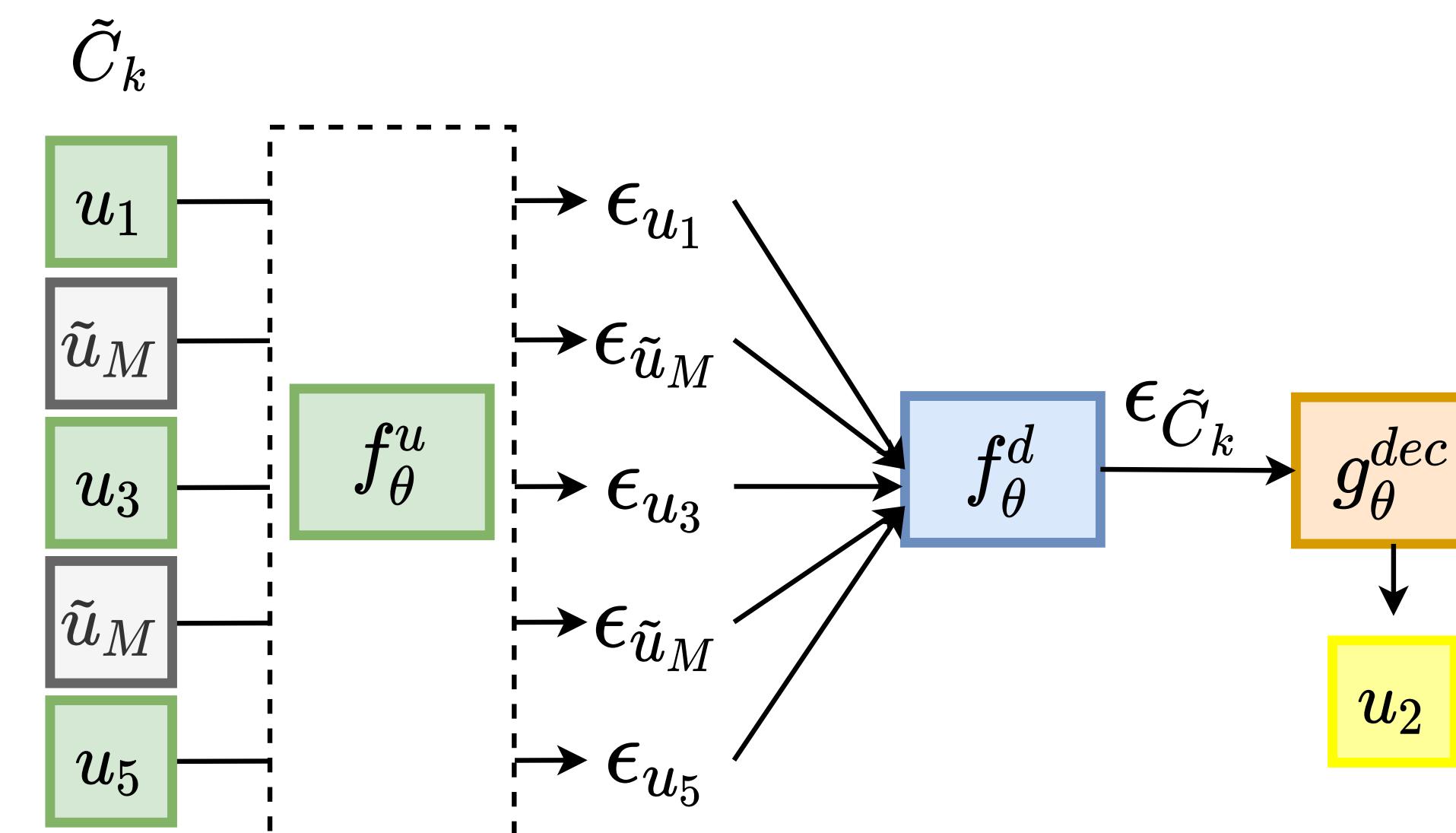
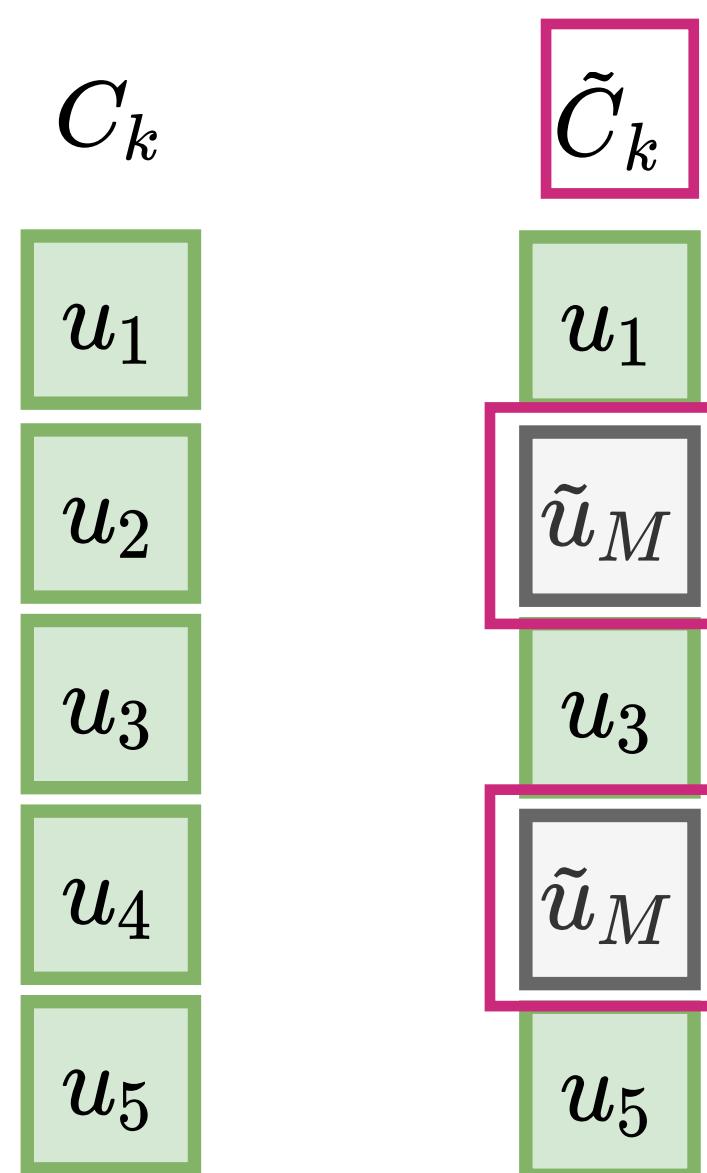


# Dialog Level Pretraining

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

## Masked Sequence Generation (MSG)

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

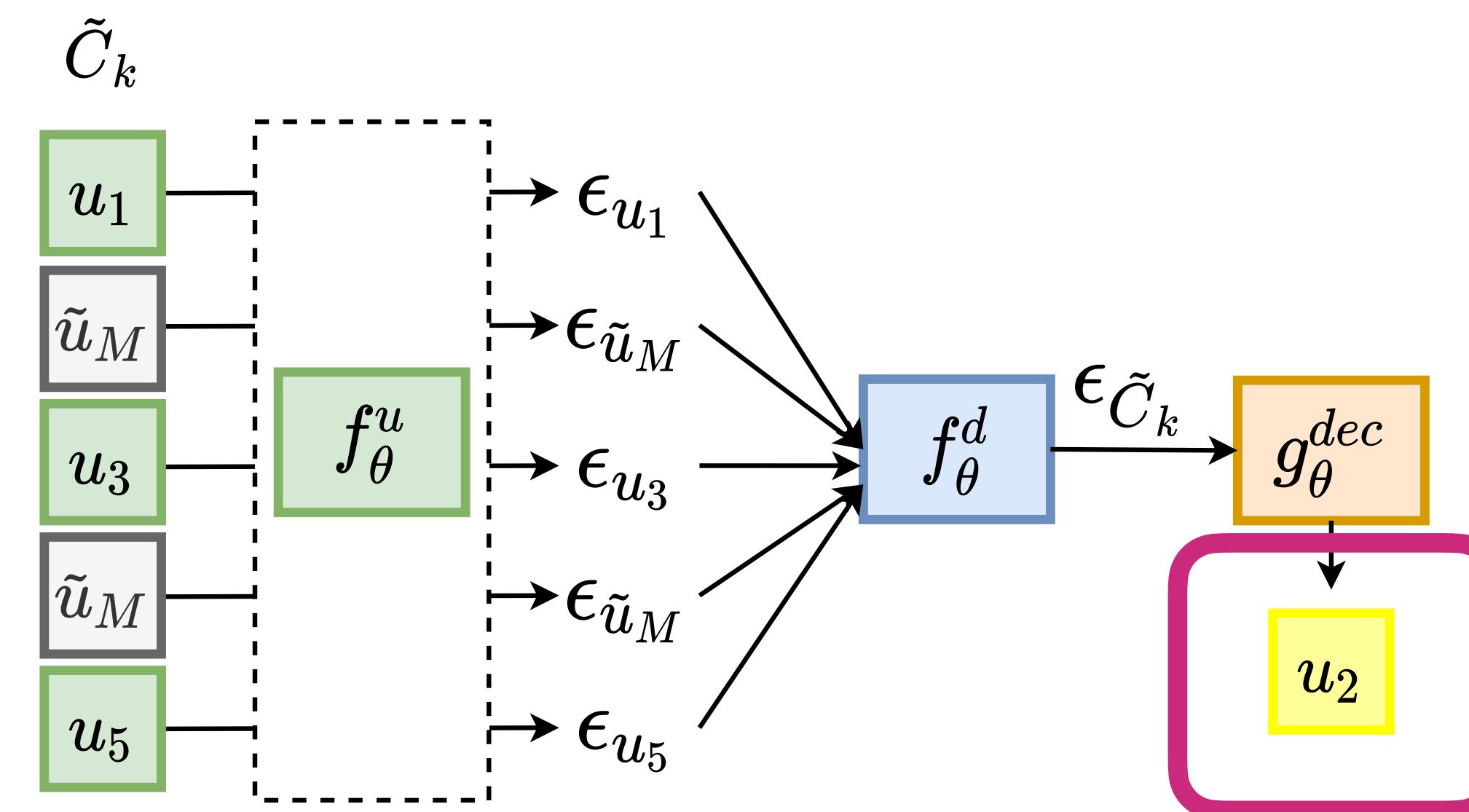
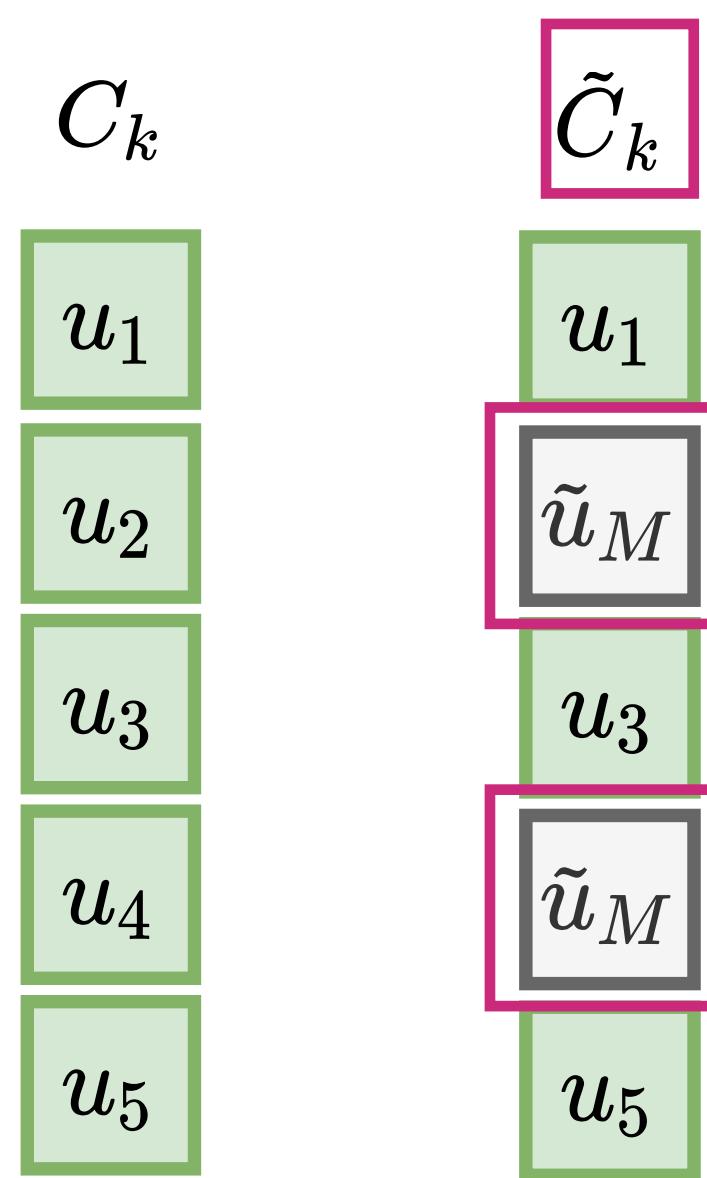


# Dialog Level Pretraining

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

## Masked Sequence Generation (MSG)

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



# Connection to Mutual Information

---

Kong et al 2020

## Mutual Information

Particularité des données

$$I(A; B) = H(A) - H(A | B)$$

# Connection to Mutual Information

Kong et al 2020

## Mutual Information

$$I(A; B) = H(A) - H(A | B)$$

**Cross entropy loss**  $E_{p(A,B)} \left[ f_\theta(a,b) - \log \sum_{\tilde{b} \in \mathcal{V}} \exp f_\theta(a, \tilde{b}) \right] - \log |\mathcal{V}| \quad \tilde{B} = \mathcal{B} \quad q(\tilde{B})$

$$f_\theta(a, b) = g_\psi(b)^T g_\phi(a)$$

Objective	a	b	$p(a, b)$	$g_\omega$	$g_\psi$
Skip-gram	word	word	word and its context	lookup	lookup
MLM	context	masked word	masked tokens probability	Transformer	lookup
NSP	sentence	sentence	(non-)consecutive sentences	Transformer	lookup
XLNet	context	masked word	factorization permutation	TXL++	lookup
DIM	context	masked $n$ -grams	sentence and its $n$ -grams	Transformer	not used

 $MSG$  $\tilde{C}_i$ 

Masked utterances

masked utterance probability

HTransformer

Lookup

# Connection to Mutual Information

Kong et al 2020

## Mutual Information

$$I(A; B) = H(A) - H(A | B)$$

Cross entropy loss  $E_{p(A,B)} \left[ f_\theta(a,b) - \log \sum_{\tilde{b} \in \mathcal{V}} \exp f_\theta(a, \tilde{b}) \right] - \log |\mathcal{V}| \quad \tilde{B} = \mathcal{B} \quad q(\tilde{B})$

$$f_\theta(a, b) = g_\psi(b)^T g_\phi(a)$$

Objective	$a$	$b$	$p(a, b)$	$g_\omega$	$g_\psi$
Skip-gram	word	word	word and its context	lookup	lookup
MLM	context	masked word	masked tokens probability	Transformer	lookup
NSP	sentence	sentence	(non-)consecutive sentences	Transformer	lookup
XLNet	context	masked word	factorization permutation	TXL++	lookup
DIM	context	masked $n$ -grams	sentence and its $n$ -grams	Transformer	not used
<b>MSG</b>	$\tilde{C}_i$	Masked utterances	masked utterance probability	HTransformer	Lookup

Same for GAP Yang et al. 2019

# 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

Target Task: Sequence Labelling

**SILICONE** (Sequence labellIng evaLuation  
benChmark fOr spoken laNguagE)

Sizes

Schemas

# Results on SILICONE

Pretraining Corpora

OpenSubtitles

Lison and Tiedemann, 2016

Spoken dialogs

> 2.3 billion

Target Task: Sequence Labelling

**SILICONE** (Sequence labellIng evaLuation  
benChmark fOr spoken laNguagE)

Sizes

Schemas

	Avg	SwDA	MRDA	DyDA <sub>DA</sub>	MT	Oasis	DyDA <sub>e</sub>	MELD <sub>s</sub>	MELD <sub>e</sub>	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	<b>82.6</b>	88.2	66.9	91.9	59.3	<b>61.4</b>	<b>45.0</b>	62.7
$\mathcal{H}\mathcal{R}$	69.8	77.5	90.9	80.1	82.8	64.3	91.5	59.3	59.9	40.3	51.1
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (TINY)	73.3	<b>79.3</b>	92.0	80.1	90.0	68.3	92.5	62.6	59.9	42.0	66.6
$\mathcal{H}\mathcal{T}(\theta_{GAP}^d)$ (TINY)	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (SMALL)	<b>74.3</b>	79.2	<b>92.4</b>	81.5	<b>90.6</b>	<b>69.4</b>	<b>92.7</b>	<b>64.1</b>	60.1	<b>45.0</b>	<b>68.2</b>

# Results on SILICONE

Pretraining Corpora

OpenSubtitles

Lison and Tiedemann, 2016

Spoken dialogs

> 2.3 billion

Target Task: Sequence Labelling

SILICONE (Sequence labellIng evaLuation  
benChmark fOr spoken laNguagE)

Sizes

Schemas

	Avg	SwDA	MRDA	DyDA <sub>DA</sub>	MT	Oasis	DyDA <sub>e</sub>	MELD <sub>s</sub>	MELD <sub>e</sub>	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	<b>82.6</b>	88.2	66.9	91.9	59.3	<b>61.4</b>	<b>45.0</b>	62.7
$\mathcal{H}\mathcal{R}$	69.8	77.5	90.9	80.1	82.8	64.3	91.5	59.3	59.9	40.3	51.1
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (TINY)	73.3	<b>79.3</b>	92.0	80.1	90.0	68.3	92.5	62.6	59.9	42.0	66.6
$\mathcal{H}\mathcal{T}(\theta_{GAP}^d)$ (TINY)	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (SMALL)	<b>74.3</b>	79.2	<b>92.4</b>	81.5	<b>90.6</b>	<b>69.4</b>	<b>92.7</b>	<b>64.1</b>	60.1	<b>45.0</b>	<b>68.2</b>

The proposed hierarchical pretraining helps.

# Results on SILICONE

Pretraining Corpora

OpenSubtitles

Lison and Tiedemann, 2016

Spoken dialogs

> 2.3 billion

Target Task: Sequence Labelling

SILICONE (Sequence labellIng evaLuation  
benChmark fOr spoken laNguagE)

Sizes

Schemas

	Avg	SwDA	MRDA	DyDA <sub>DA</sub>	MT	Oasis	DyDA <sub>e</sub>	MELD <sub>s</sub>	MELD <sub>e</sub>	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	<b>82.6</b>	88.2	66.9	91.9	59.3	<b>61.4</b>	<b>45.0</b>	62.7
$\mathcal{H}\mathcal{R}$	69.8	77.5	90.9	80.1	82.8	64.3	91.5	59.3	59.9	40.3	51.1
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (TINY)	73.3	<b>79.3</b>	92.0	80.1	90.0	68.3	92.5	62.6	59.9	42.0	66.6
$\mathcal{H}\mathcal{T}(\theta_{GAP}^d)$ (TINY)	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{H}\mathcal{T}(\theta_{MLM}^{u,d})$ (SMALL)	<b>74.3</b>	79.2	<b>92.4</b>	81.5	<b>90.6</b>	<b>69.4</b>	<b>92.7</b>	<b>64.1</b>	60.1	<b>45.0</b>	<b>68.2</b>

The proposed hierarchical pretraining helps.

Reduced model size thanks to hierarchy.

# **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**



**Emotion Predictor**



**Emotion**

**Positive +3**

.....

**Negative -3**

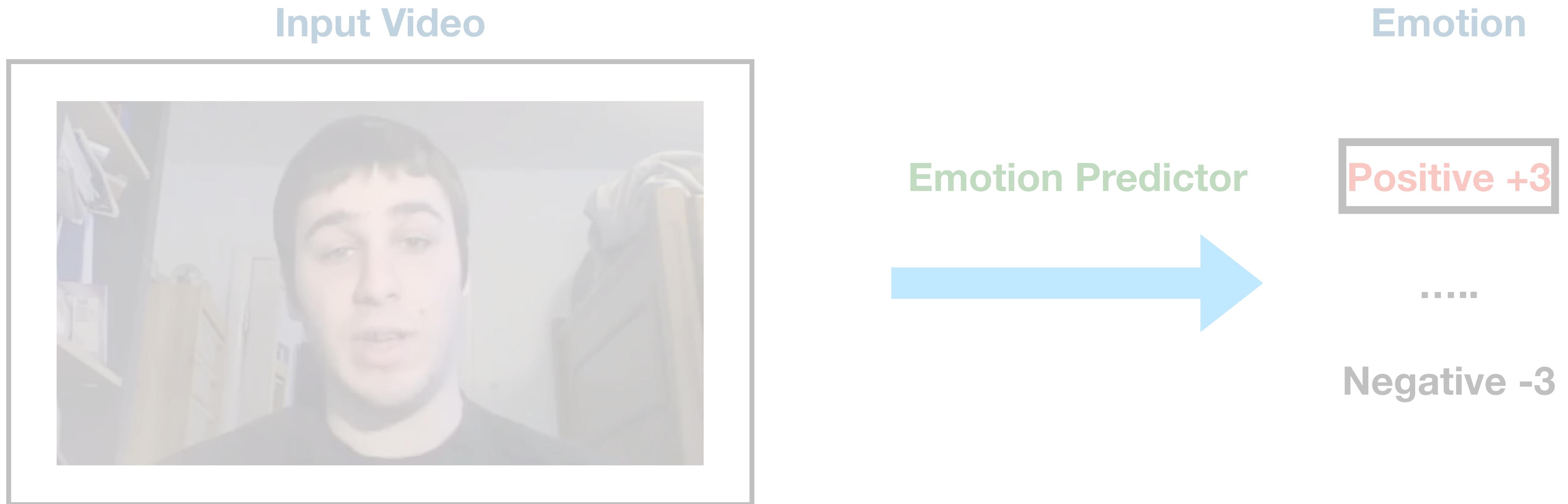
**The action is fucking awesome!**

# Multimodal Task Description



**Goal:** Include Multimodal Dimension in Representations of Spoken Transcripts

**Task:** Learn a multi-modal emotion predictor



# 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

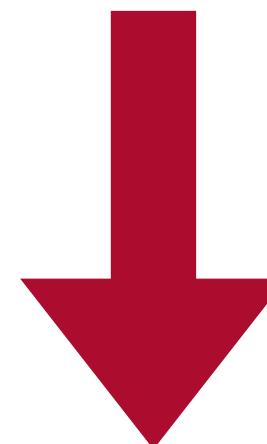
Representation

Alignment

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

Alignment

**Fusion**

Translation

Co-learning

Represent multimodal data (leverage complementarity, redundancy)

Identify relations between elements of different modalities

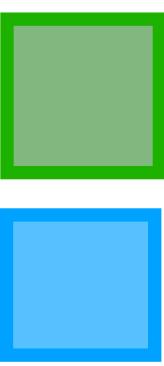
**Join information from modalities**

Translate one modality to another

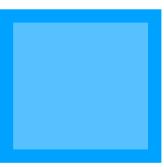
Transfer knowledge between modalities

# Multimodal sentiment analysis

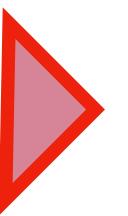
---



Fusion Block

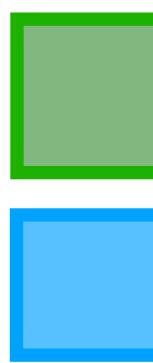


Embedding Block



Predictor block

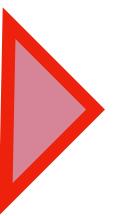
# Multimodal sentiment analysis



Fusion Block

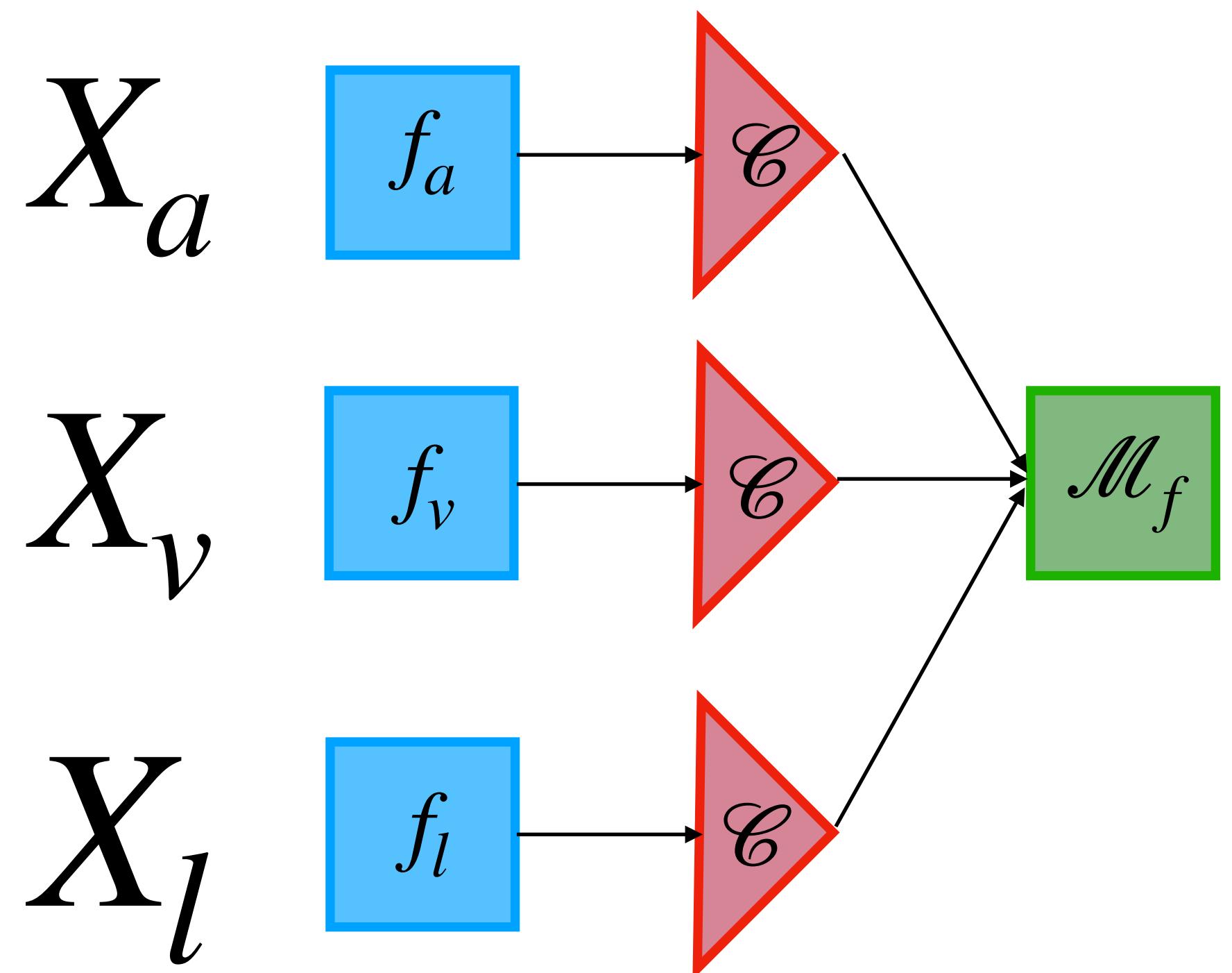


Embedding Block

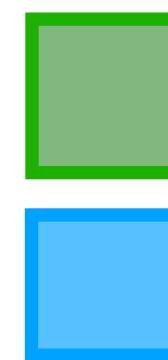


Predictor block

## Late Fusion

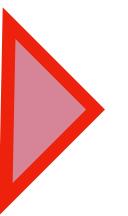


# Multimodal sentiment analysis



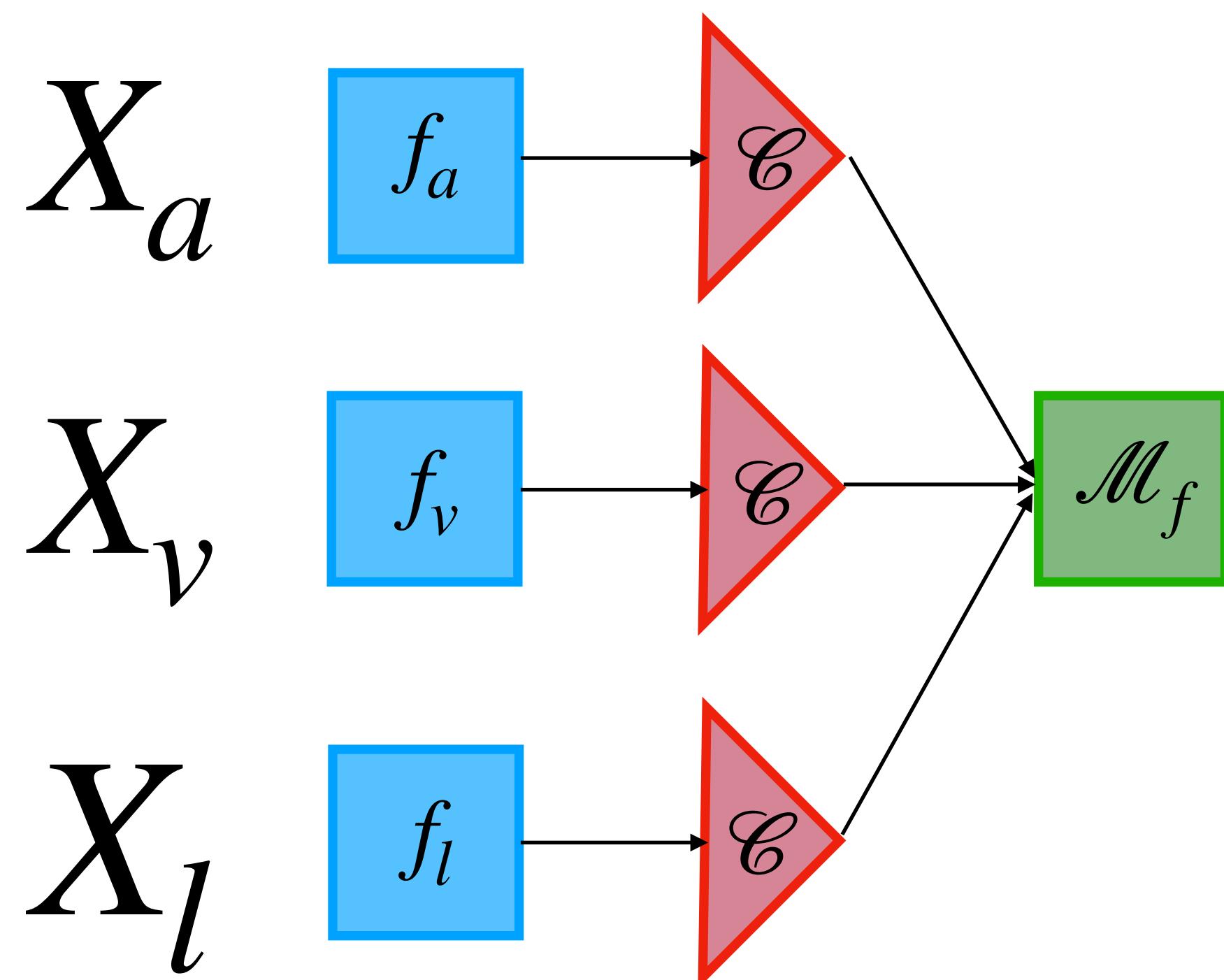
Fusion Block

Embedding Block

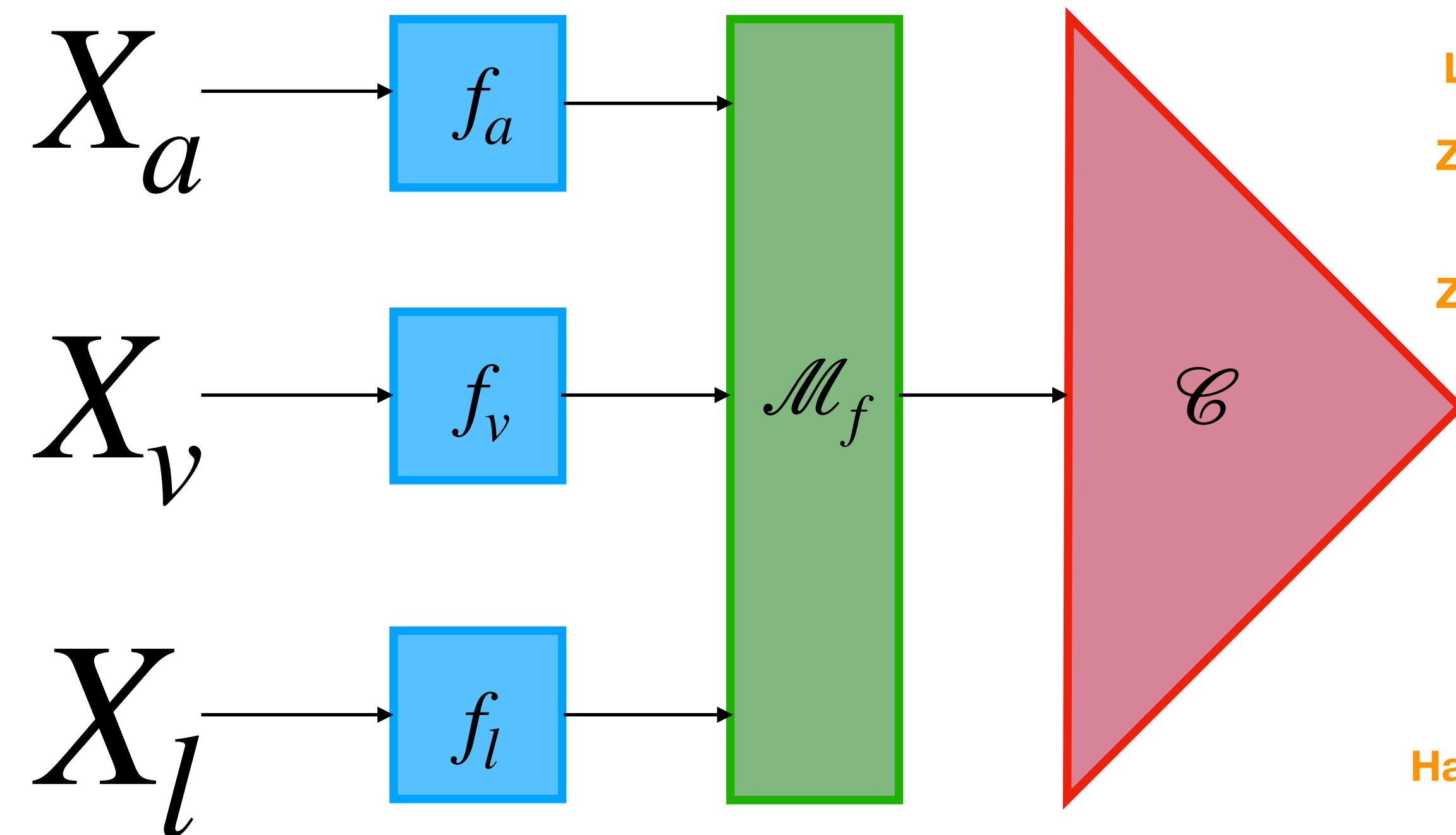


Predictor block

## Late Fusion



## Early Fusion



Liang et al 2019

Zadeh et al 2017

Zadeh et al 2018

Liu et al 2018

Hazarika et al 2020

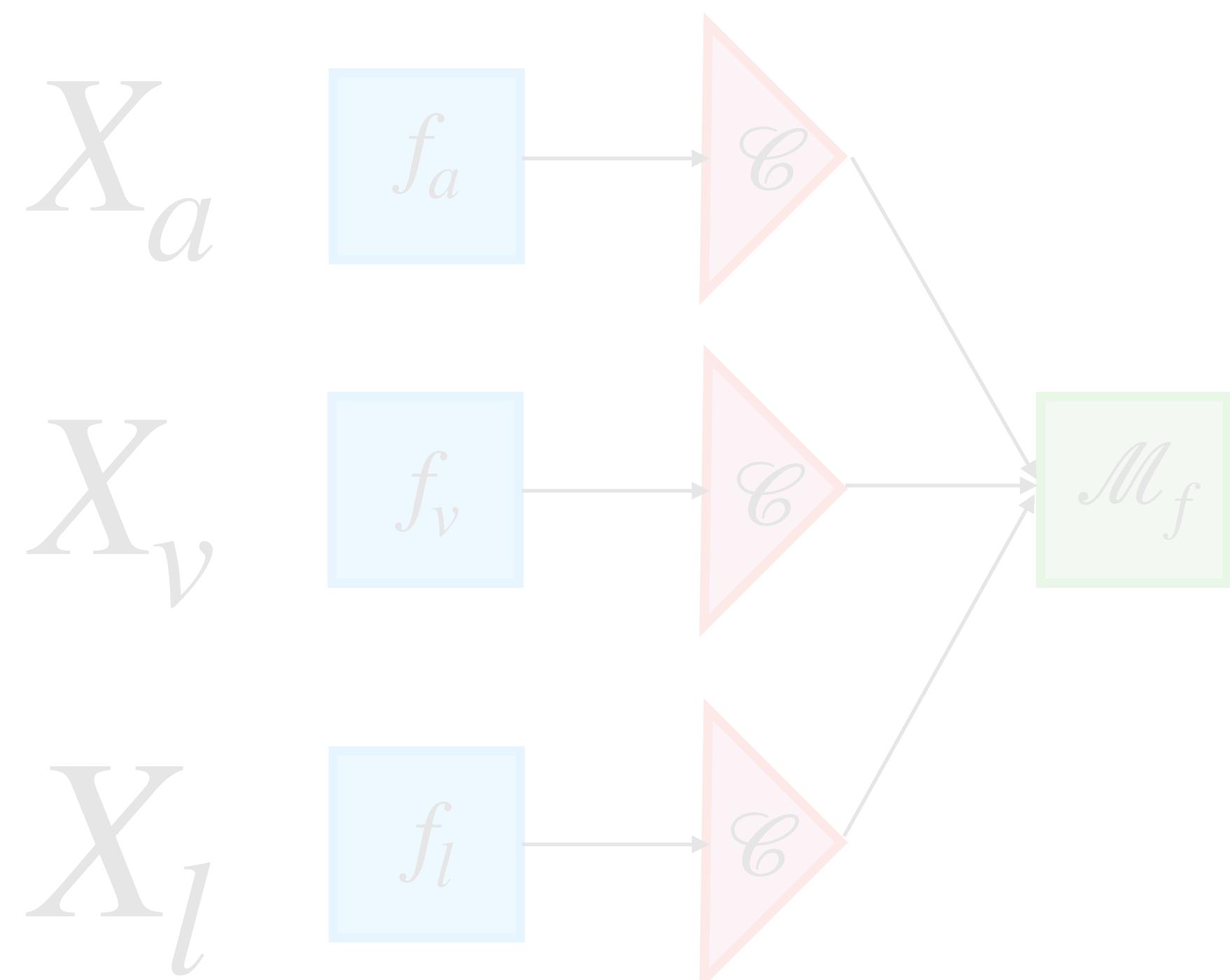
# Multimodal sentiment analysis

Fusion Block

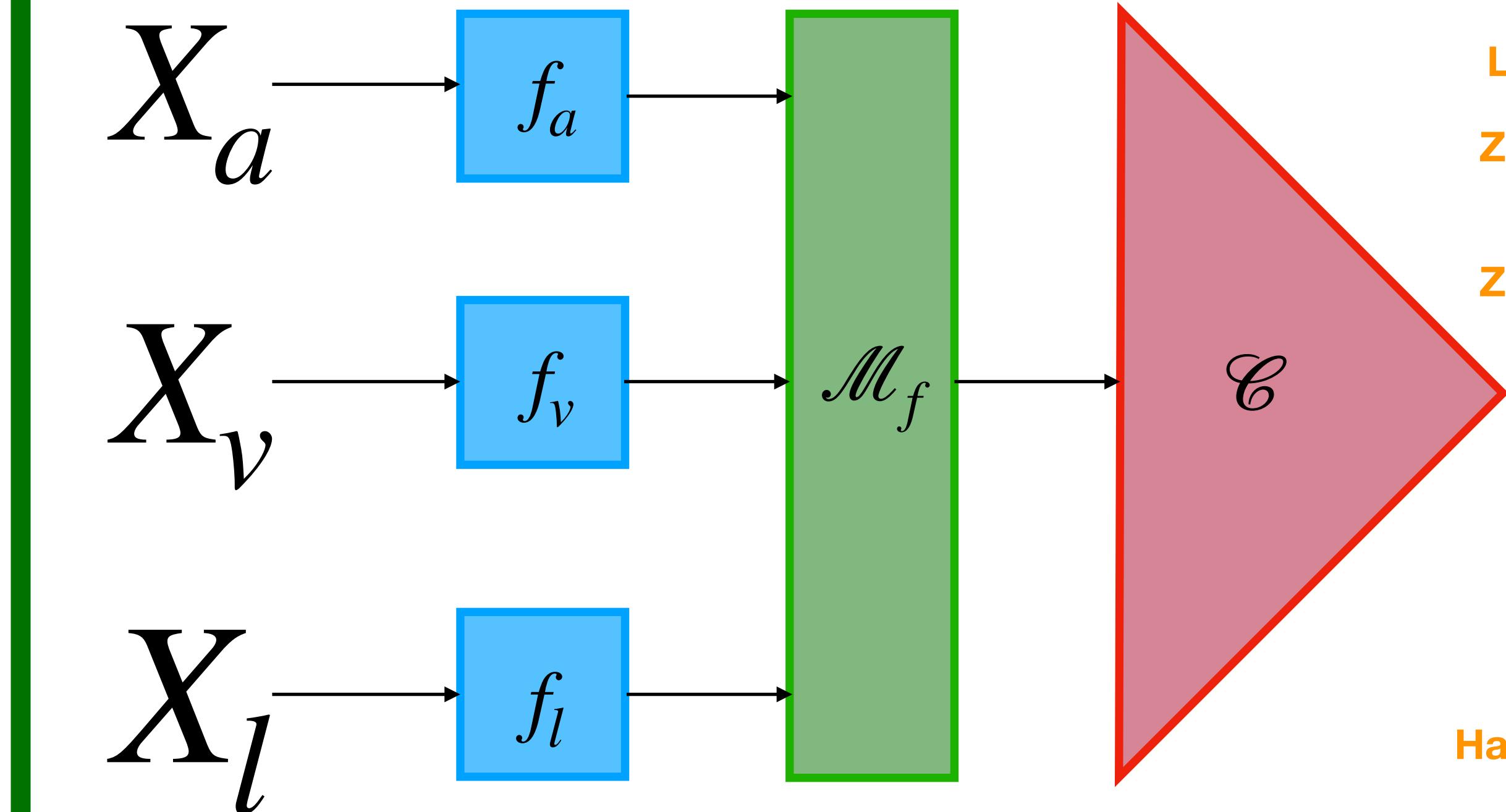
Embedding Block

Predictor block

## Late Fusion



## Early Fusion



Liang et al 2019

Zadeh et al 2017

Zadeh et al 2018

Liu et al 2018

Hazarika et al 2020

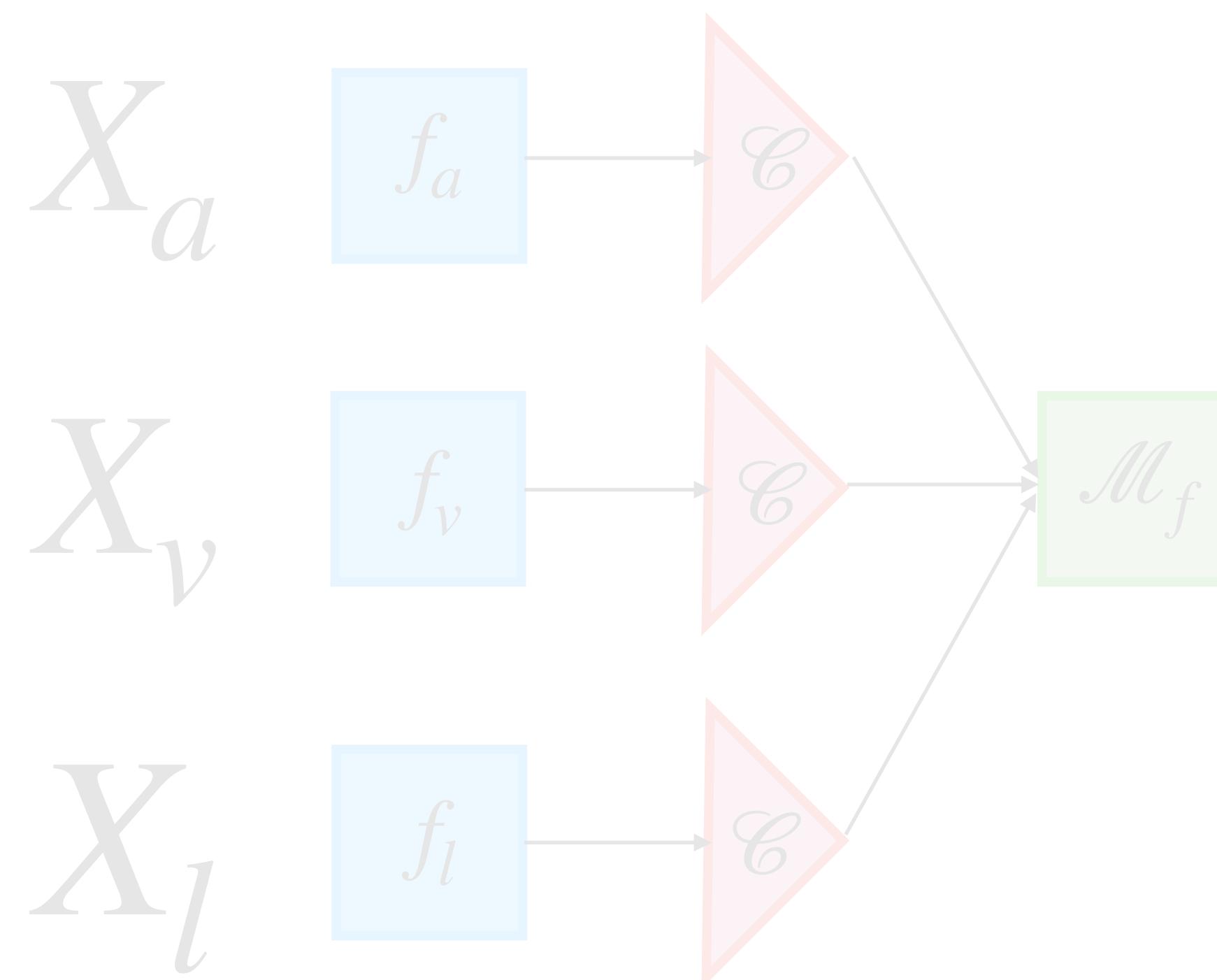
# Multimodal sentiment analysis

Fusion Block

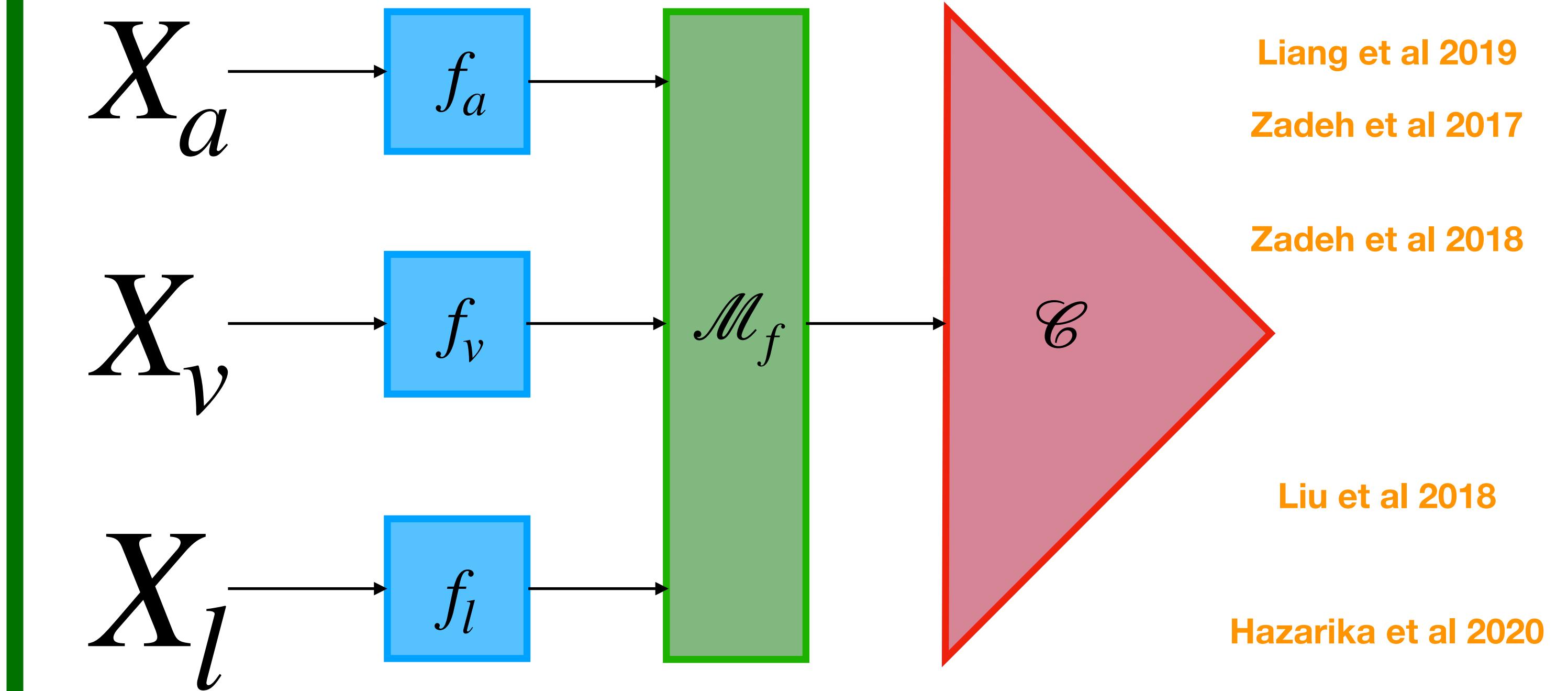
Embedding Block

Predictor block

## Late Fusion



## Early Fusion



Fusion in previous work mainly rely on complex neural networks  
Few previous works on improving fusion using loss function!

# Problem formulation

---

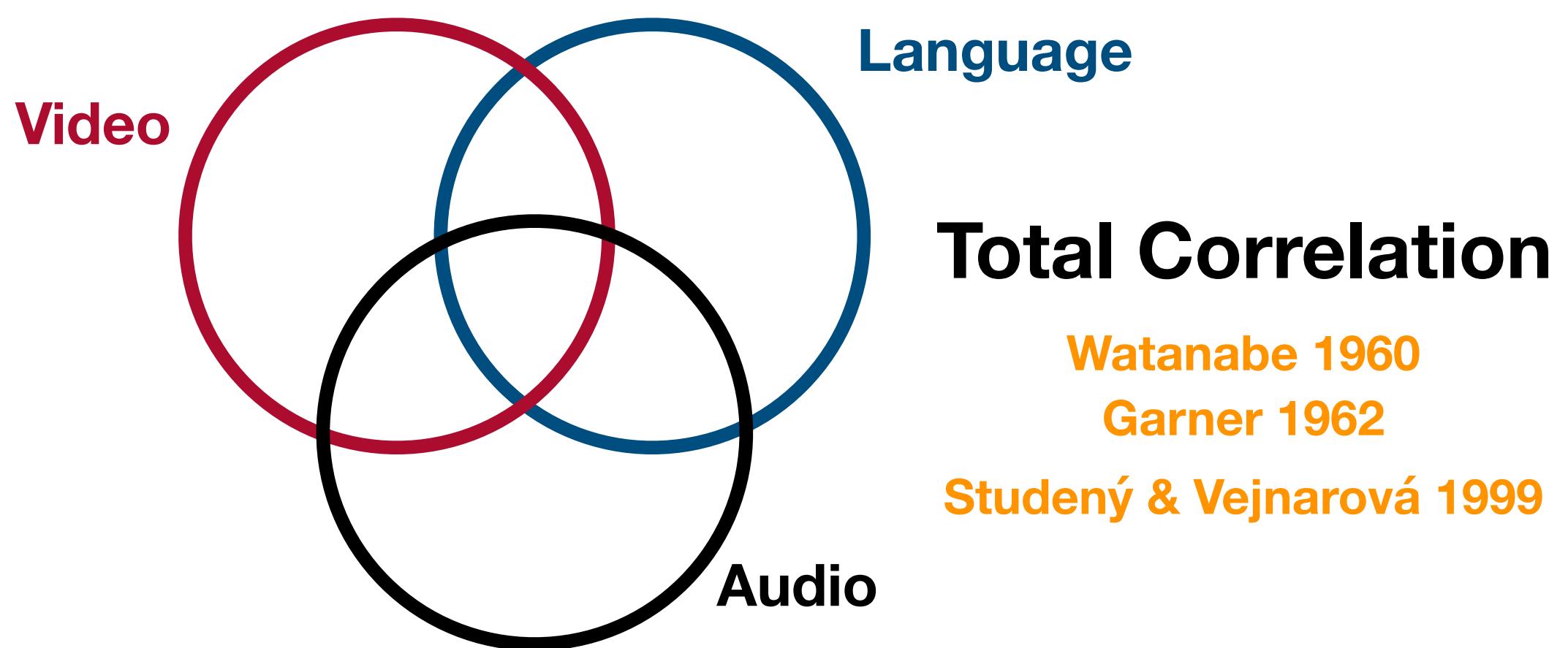
## Problem formulation

---

**Fusion Mechanism**  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$

# Problem formulation

Fusion Mechanism  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$



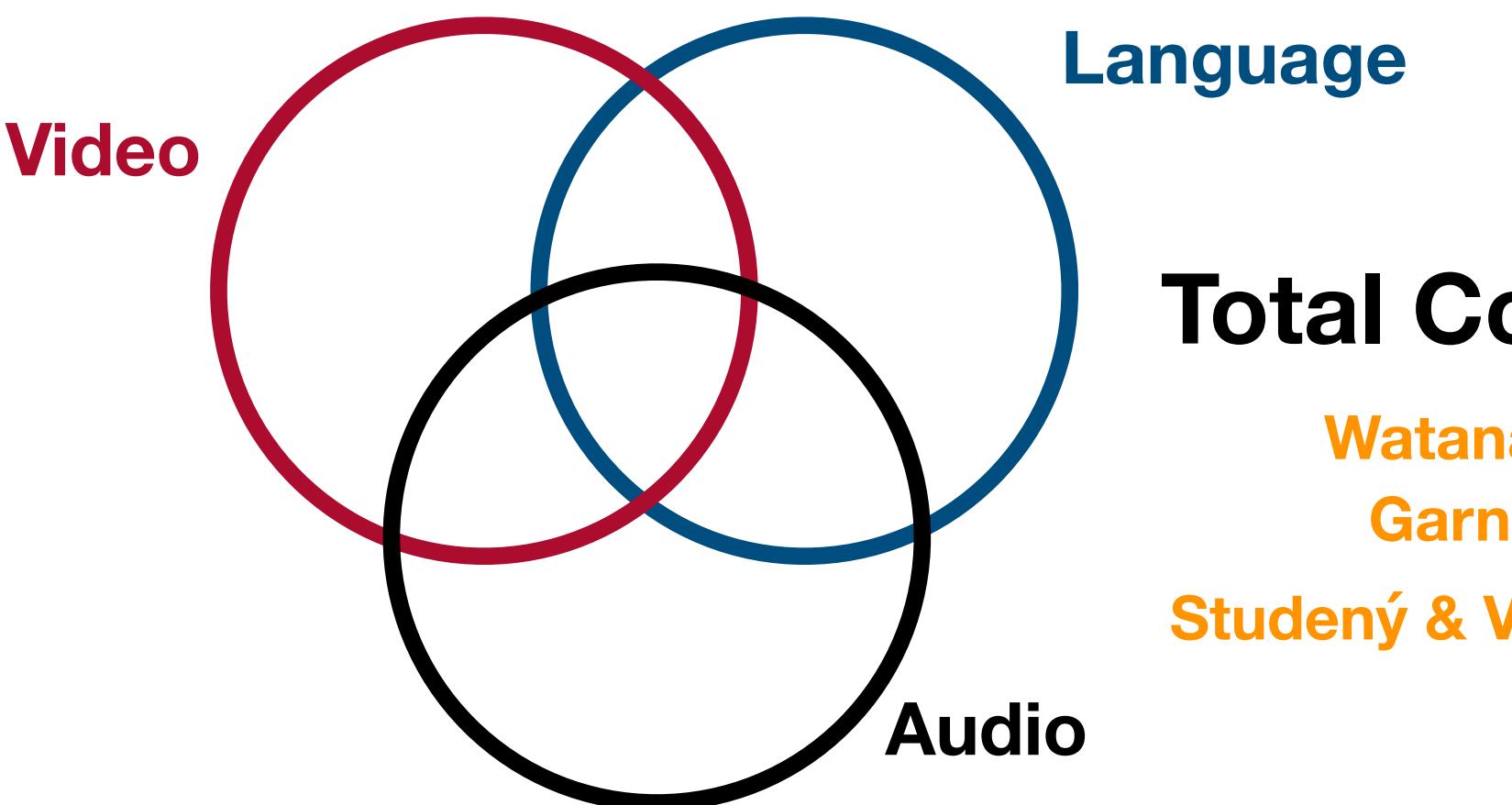
# Problem formulation

Fusion Mechanism  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$

What we want for  $\mathcal{M}_f$ :

- Retain modality-specific interaction
- Retain cross-view interaction
- Retain task specific information

$$\mathcal{L}_{MDM}$$
$$\mathcal{L}_{down.}$$



Total Correlation

Watanabe 1960

Garner 1962

Studený & Vejnarová 1999

# Problem formulation

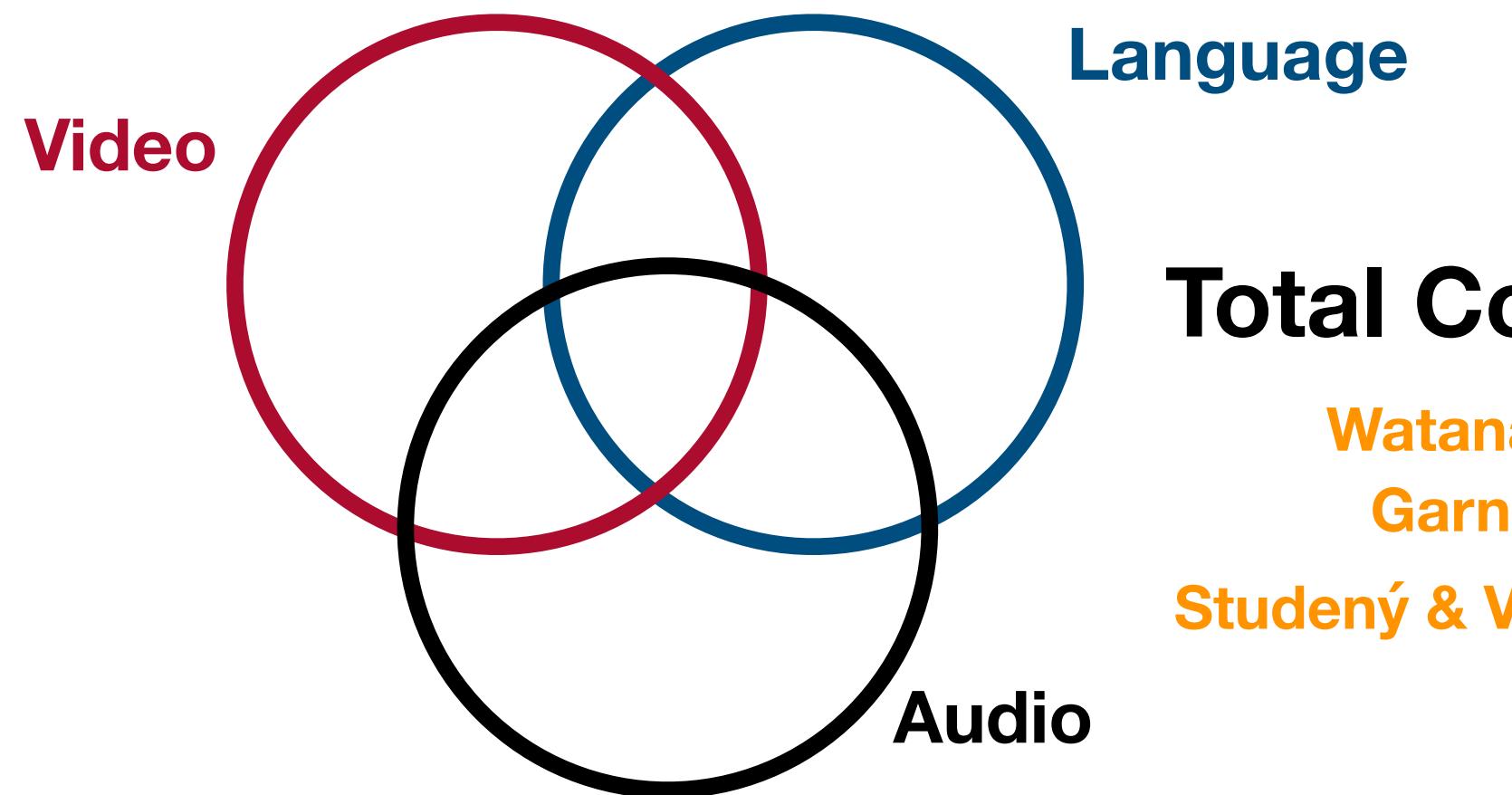
Fusion Mechanism  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$

What we want for  $\mathcal{M}_f$ :

- Retain modality-specific interaction
- Retain cross-view interaction
- Retain task specific information

$$\text{Total} = \underbrace{\mathcal{L}_{down.}}_{\text{main task}} - \underbrace{\lambda \cdot \mathcal{L}_{MDM}}_{\text{mutual dependency term}}$$

$$\mathcal{L}_{MDM}$$
  
$$\mathcal{L}_{down.}$$



## Total Correlation

Watanabe 1960

Garner 1962

Studený & Vejnarová 1999

$$\mathcal{L}_{MDM} \triangleq MDM \left( p_{X_a X_v X_l}(x_a, x_v, x_l), \prod_{j \in \{a, v, l\}} p_{X_j}(x_j) \right)$$

Belghazi et al 2018

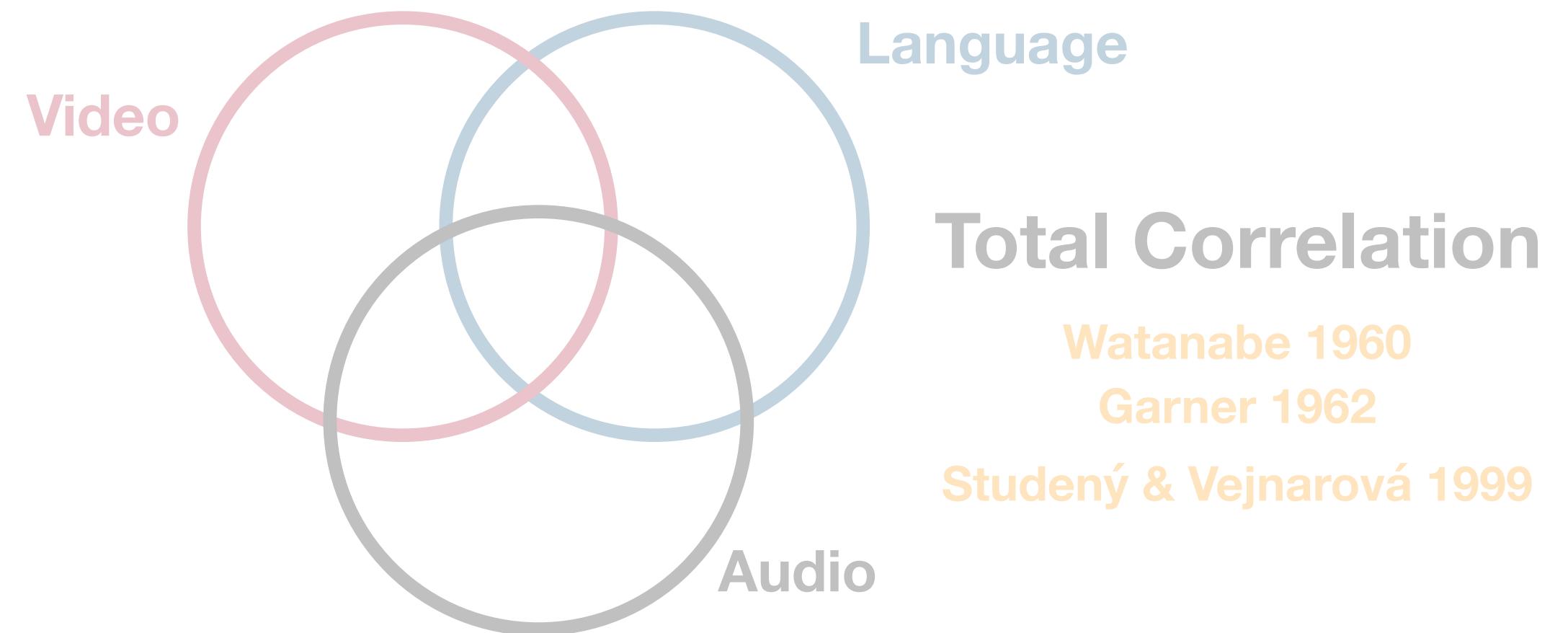
# Problem formulation

Fusion Mechanism  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$

What we want for  $\mathcal{M}_f$ :

- Retain modality-specific interaction
- Retain cross-view interaction
- Retain task specific information

$$\mathcal{L}_{MDM}$$
  
$$\mathcal{L}_{down.}$$



Total  $\underbrace{\mathcal{L}_{down.}}_{\text{main task}} - \underbrace{\lambda \cdot \mathcal{L}_{MDM}}_{\text{mutual dependency term}}$

$$\mathcal{L}_{MDM} \triangleq MDM \left( p_{X_a X_v X_l}(x_a, x_v, x_l), \prod_{j \in \{a, v, l\}} p_{X_j}(x_j) \right)$$

Belghazi et al 2018

3 different MDM : KL divergence, f-divergence, Wasserstein distance

# Overall Results

---

# Overall Results

---

CMU-MOSEI

CMU-MOSI

2,199/23,454 movie review videos

Sentiment Score in [-3,3]

# Overall Results

---

CMU-MOSEI	CMU-MOSI
<b>2,199/23,454 movie review videos</b>	
<b>Sentiment Score in [-3,3]</b>	

	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
--	-----------	-----------	---------	----------

CMU-MOSI

$\mathcal{L}_\emptyset$	31.1	76.1	1.00	0.65
$\mathcal{L}_{kl}$	<u>31.7</u>	<b>76.4</b>	1.00	<b>0.66</b>
$\mathcal{L}_f$	<b>33.7</b>	76.2	1.02	<b>0.66</b>
$\mathcal{L}_W$	<u>33.5</u>	<b>76.4</b>	<b>0.98</b>	<b>0.66</b>

CMU-MOSEI

$\mathcal{L}_\emptyset$	44.2	75.0	0.72	0.52
$\mathcal{L}_{kl}$	44.5	<u>75.6</u>	<u>0.70</u>	<u>0.53</u>
$\mathcal{L}_f$	<b>45.5</b>	75.2	<u>0.70</u>	0.52
$\mathcal{L}_W$	<u>45.3</u>	<b>75.9</b>	<b>0.68</b>	<b>0.54</b>

	CMU-MOSI				CMU-MOSEI			
	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
MFN								
$\mathcal{L}_\emptyset$	31.3	76.6	1.01	0.62	44.4	74.7	0.72	0.53
$\mathcal{L}_{kl}$	<u>32.5</u>	76.7	<b>0.96</b>	<b>0.65</b>	44.2	74.7	0.72	<b>0.57</b>
$\mathcal{L}_f$	<u>35.7</u>	<u>77.4</u>	<b>0.96</b>	<b>0.65</b>	<u>46.1</u>	<b>75.4</b>	<b>0.69</b>	<u>0.56</u>
$\mathcal{L}_W$	<b>35.9</b>	<b>77.6</b>	<b>0.96</b>	<b>0.65</b>	<u>46.2</u>	75.1	<b>0.69</b>	0.56
LFN								
$\mathcal{L}_\emptyset$	31.9	76.9	1.00	0.63	45.2	74.2	0.70	0.54
$\mathcal{L}_{kl}$	<u>32.6</u>	<b>77.7</b>	0.97	0.63	<u>46.1</u>	75.3	0.68	<b>0.57</b>
$\mathcal{L}_f$	<b>35.6</b>	77.1	0.97	0.63	45.8	<b>75.4</b>	0.69	<u>0.57</u>
$\mathcal{L}_W$	<b>35.6</b>	<b>77.7</b>	<b>0.96</b>	<b>0.67</b>	<u>46.2</u>	<b>75.4</b>	<b>0.67</b>	<b>0.57</b>
MAGBERT								
$\mathcal{L}_\emptyset$	40.2	84.7	0.79	0.80	46.8	84.9	<b>0.59</b>	0.77
$\mathcal{L}_{kl}$	<b>42.0</b>	<b>85.6</b>	<b>0.76</b>	<b>0.82</b>	47.1	85.4	<b>0.59</b>	<b>0.79</b>
$\mathcal{L}_f$	<u>41.7</u>	<b>85.6</b>	0.78	<b>0.82</b>	46.9	<b>85.6</b>	<b>0.59</b>	<b>0.79</b>
$\mathcal{L}_W$	<u>41.8</u>	85.3	<b>0.76</b>	<b>0.82</b>	<u>47.8</u>	85.5	<b>0.59</b>	<b>0.79</b>
MAGXLNET								
$\mathcal{L}_\emptyset$	43.0	86.2	0.76	<b>0.82</b>	46.7	84.4	<b>0.59</b>	0.79
$\mathcal{L}_{kl}$	<b>44.5</b>	86.1	<b>0.74</b>	<b>0.82</b>	<u>47.5</u>	<b>85.4</b>	<b>0.59</b>	0.81
$\mathcal{L}_f$	<u>43.9</u>	86.6	<b>0.74</b>	<b>0.82</b>	47.4	85.0	<b>0.59</b>	0.81
$\mathcal{L}_W$	<u>44.4</u>	<b>86.9</b>	<b>0.74</b>	<b>0.82</b>	<u>47.9</u>	<b>85.8</b>	<b>0.59</b>	<b>0.82</b>

# Overall Results

---

CMU-MOSEI	CMU-MOSI
<b>2,199/23,454 movie review videos</b>	
<b>Sentiment Score in [-3,3]</b>	

## Simple fusion mechanism

	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
--	-----------	-----------	---------	----------

	CMU-MOSI			
$\mathcal{L}_\emptyset$	31.1	76.1	1.00	0.65
$\mathcal{L}_{kl}$	<u>31.7</u>	<b>76.4</b>	1.00	<b>0.66</b>
$\mathcal{L}_f$	<b>33.7</b>	76.2	1.02	<b>0.66</b>
$\mathcal{L}_W$	33.5	<b>76.4</b>	<b>0.98</b>	<b>0.66</b>

	CMU-MOSEI			
$\mathcal{L}_\emptyset$	44.2	75.0	0.72	0.52
$\mathcal{L}_{kl}$	44.5	<u>75.6</u>	<u>0.70</u>	<u>0.53</u>
$\mathcal{L}_f$	<b>45.5</b>	75.2	<u>0.70</u>	0.52
$\mathcal{L}_W$	<u>45.3</u>	<b>75.9</b>	<u>0.68</u>	<b>0.54</b>

## Complex fusion mechanism

	CMU-MOSI				CMU-MOSEI			
	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
MFN								
$\mathcal{L}_\emptyset$	31.3	76.6	1.01	0.62	44.4	74.7	0.72	0.53
$\mathcal{L}_{kl}$	<u>32.5</u>	76.7	<u>0.96</u>	<u>0.65</u>	44.2	74.7	0.72	<u>0.57</u>
$\mathcal{L}_f$	35.7	77.4	<u>0.96</u>	<u>0.65</u>	46.1	<b>75.4</b>	<b>0.69</b>	0.56
$\mathcal{L}_W$	<b>35.9</b>	<b>77.6</b>	<b>0.96</b>	<b>0.65</b>	<b>46.2</b>	75.1	<b>0.69</b>	0.56
LFN								
$\mathcal{L}_\emptyset$	31.9	76.9	1.00	0.63	45.2	74.2	0.70	0.54
$\mathcal{L}_{kl}$	<u>32.6</u>	<b>77.7</b>	0.97	0.63	<u>46.1</u>	75.3	0.68	<u>0.57</u>
$\mathcal{L}_f$	<b>35.6</b>	77.1	0.97	0.63	45.8	<b>75.4</b>	0.69	<u>0.57</u>
$\mathcal{L}_W$	<b>35.6</b>	<b>77.7</b>	<b>0.96</b>	<b>0.67</b>	<b>46.2</b>	<b>75.4</b>	<b>0.67</b>	<b>0.57</b>
MAGBERT								
$\mathcal{L}_\emptyset$	40.2	84.7	0.79	0.80	46.8	84.9	<b>0.59</b>	0.77
$\mathcal{L}_{kl}$	<b>42.0</b>	<b>85.6</b>	<u>0.76</u>	<b>0.82</b>	47.1	85.4	<b>0.59</b>	<u>0.79</u>
$\mathcal{L}_f$	41.7	<b>85.6</b>	0.78	<b>0.82</b>	46.9	<b>85.6</b>	<b>0.59</b>	<u>0.79</u>
$\mathcal{L}_W$	<u>41.8</u>	85.3	<b>0.76</b>	<b>0.82</b>	<b>47.8</b>	85.5	<b>0.59</b>	<u>0.79</u>
MAGXLNET								
$\mathcal{L}_\emptyset$	43.0	86.2	0.76	<b>0.82</b>	46.7	84.4	<b>0.59</b>	0.79
$\mathcal{L}_{kl}$	<b>44.5</b>	86.1	<u>0.74</u>	<b>0.82</b>	<u>47.5</u>	<u>85.4</u>	<b>0.59</b>	0.81
$\mathcal{L}_f$	43.9	86.6	<u>0.74</u>	<b>0.82</b>	47.4	85.0	<b>0.59</b>	0.81
$\mathcal{L}_W$	<u>44.4</u>	<b>86.9</b>	<u>0.74</u>	<b>0.82</b>	<b>47.9</b>	<u>85.8</u>	<b>0.59</b>	<b>0.82</b>

# Towards explainable representations

---

# Towards explainable representations

---

**Goal: Use low/high values or  $MDM$  to explain representations**

# Towards explainable representations

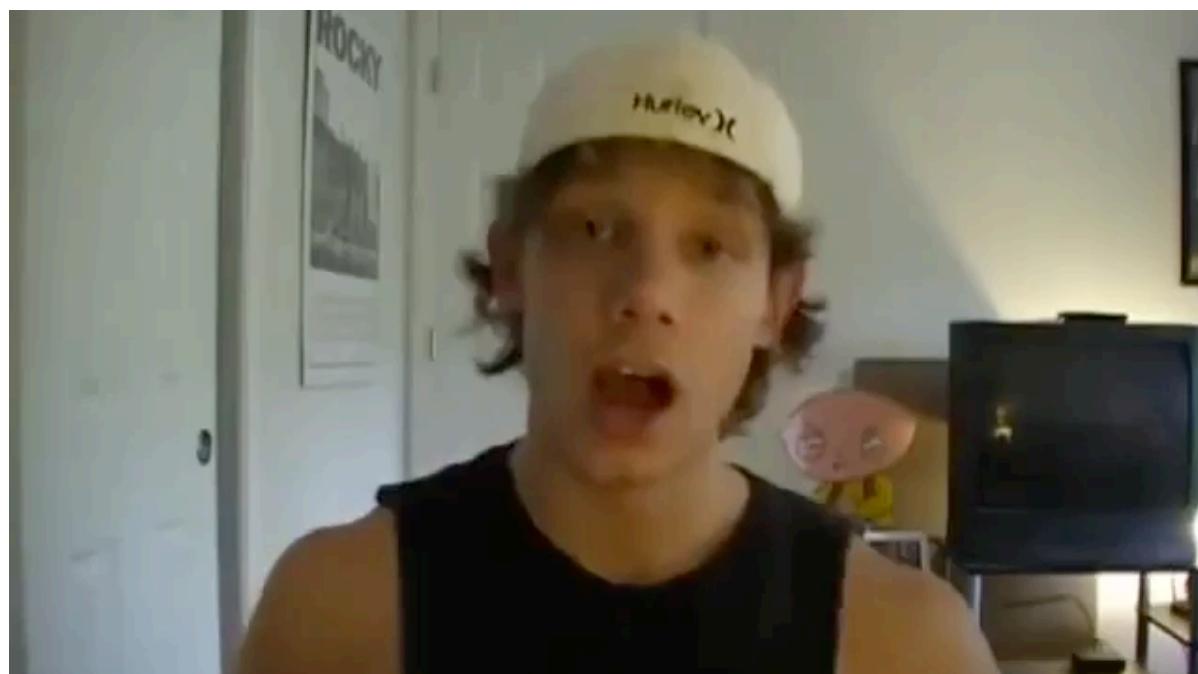
**Goal: Use low/high values or MDM to explain representations**

Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L

# Towards explainable representations

**Goal: Use low/high values or MDM to explain representations**

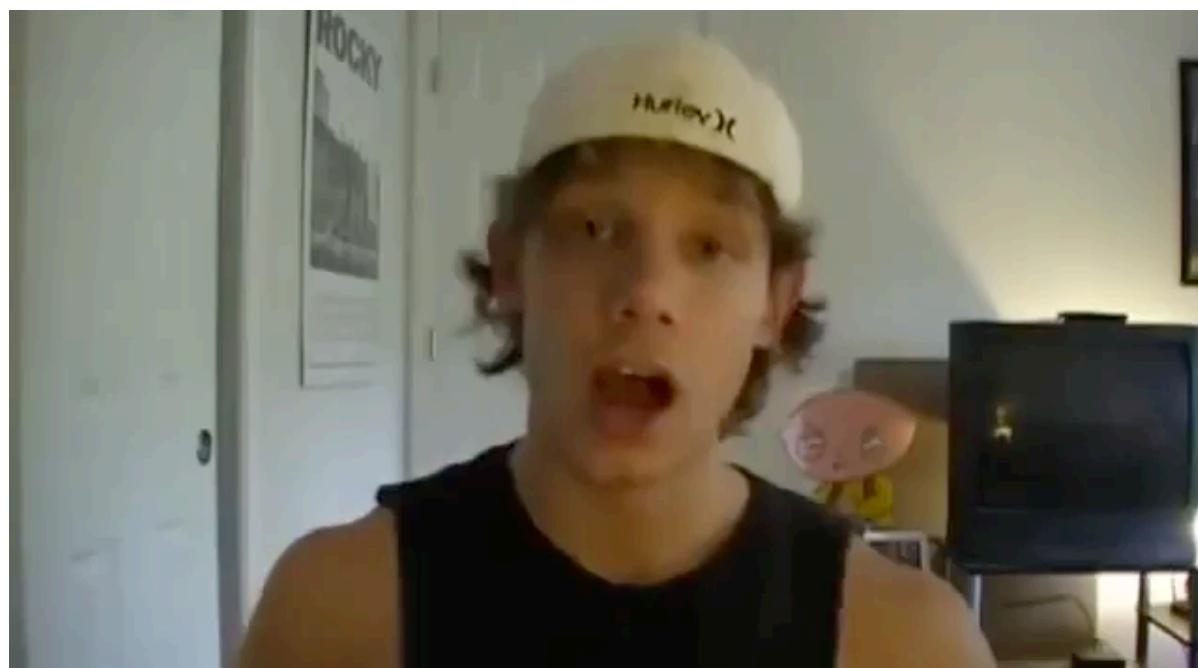
Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L



# Towards explainable representations

**Goal: Use low/high values or MDM to explain representations**

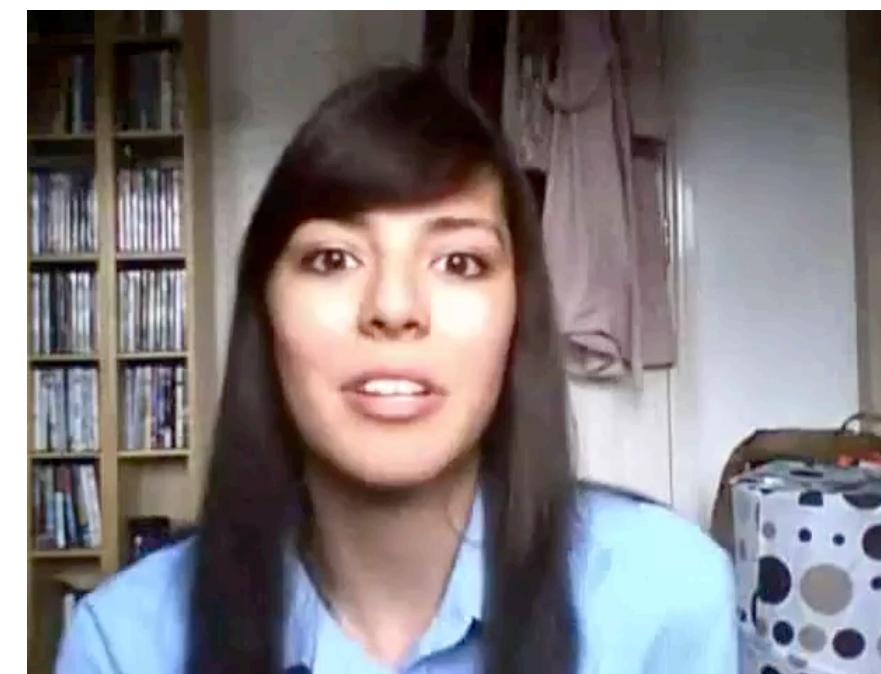
Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L



# Towards explainable representations

**Goal: Use low/high values or MDM to explain representations**

Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L



# Towards explainable representations

**Goal: Use low/high values or MDM to explain representations**

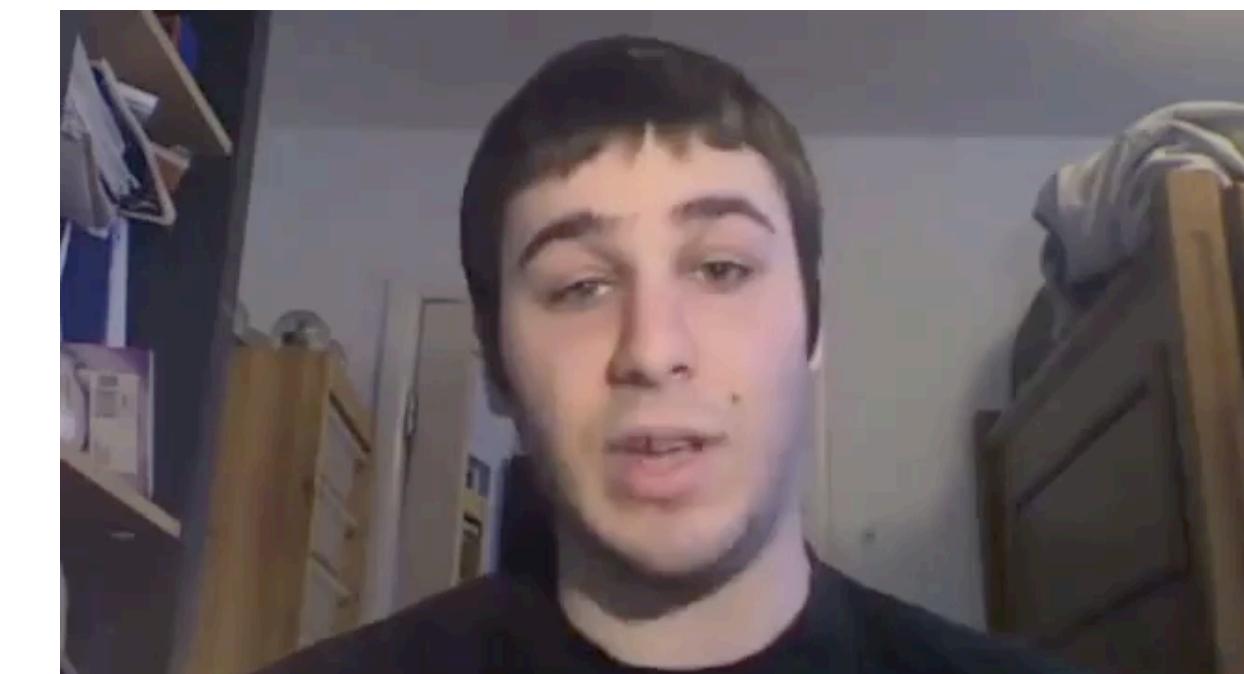
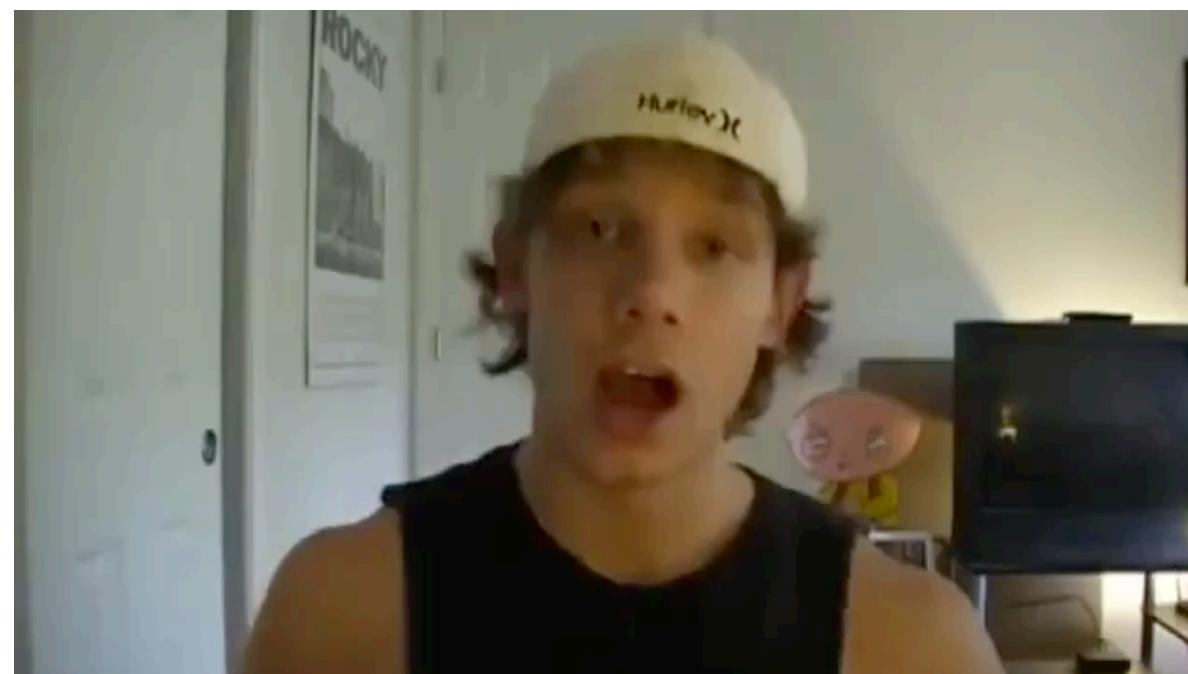
Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L
the action is fucking awesome	head nod + excited voice	H



# Towards explainable representations

Goal: Use low/high values or MDM to explain representations

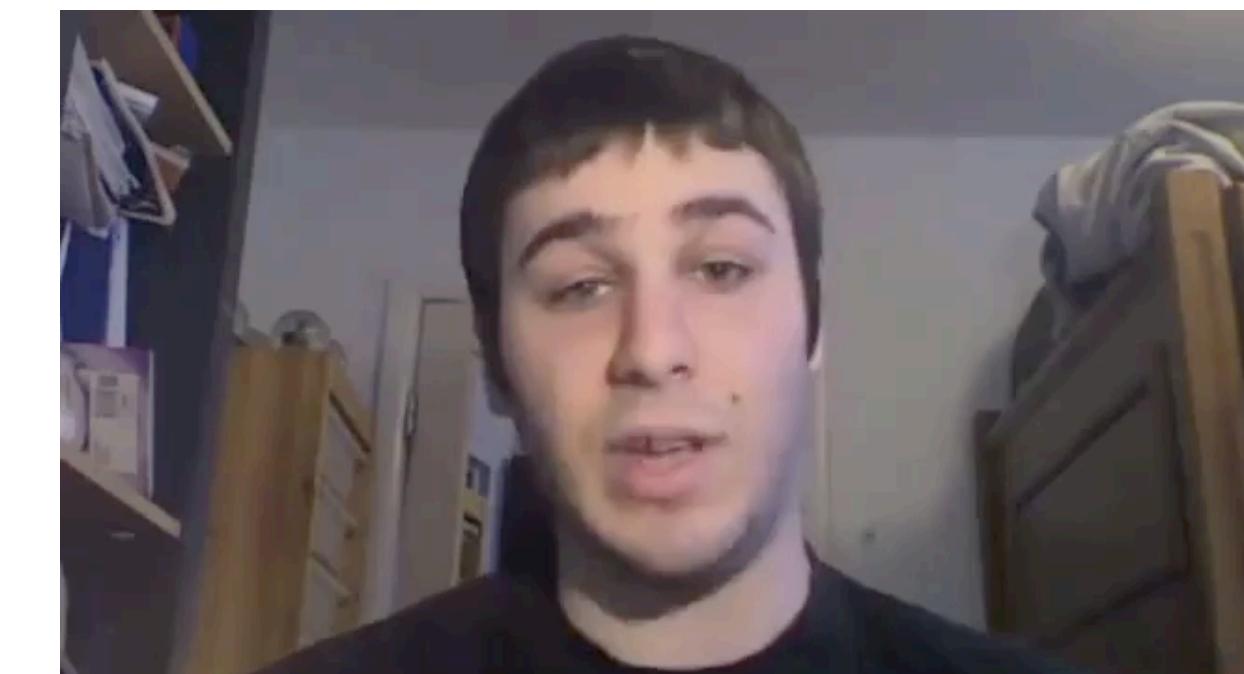
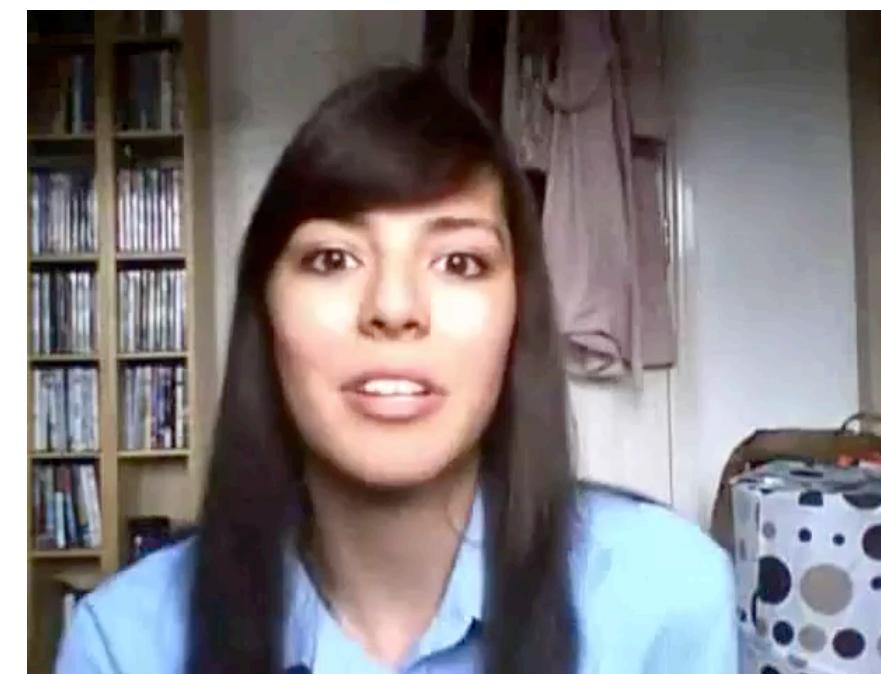
Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L
the action is fucking awesome	head nod + excited voice	H



# Towards explainable representations

Goal: Use low/high values or MDM to explain representations

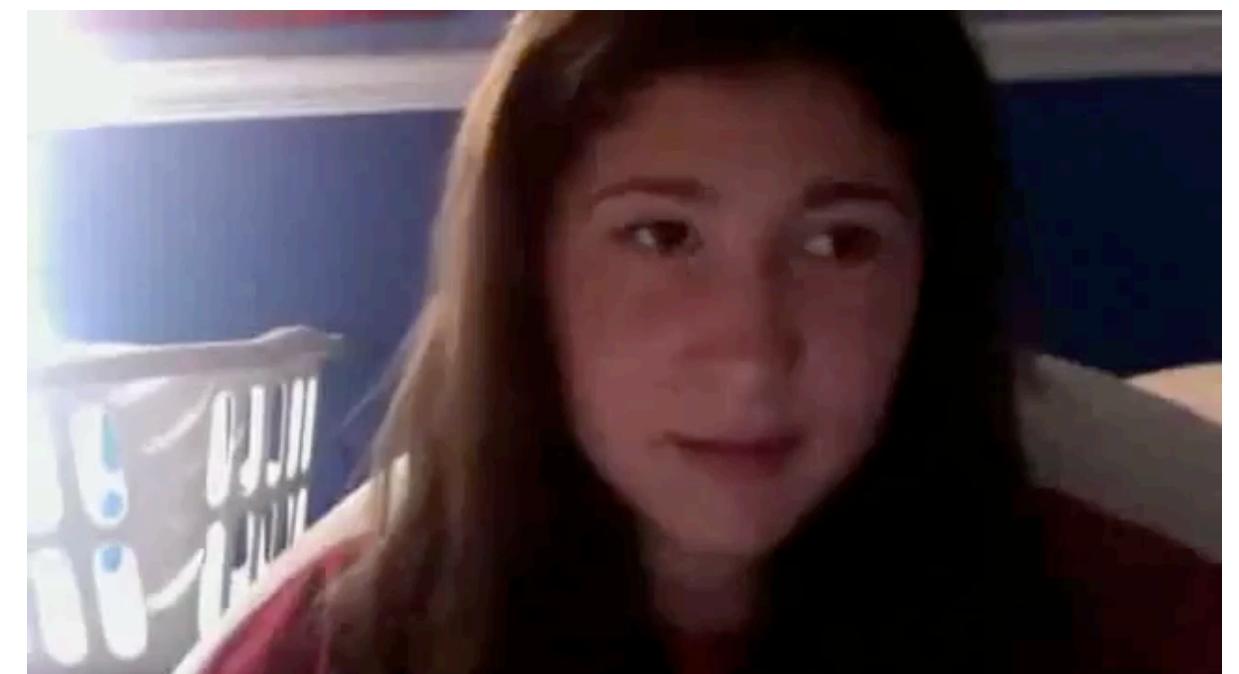
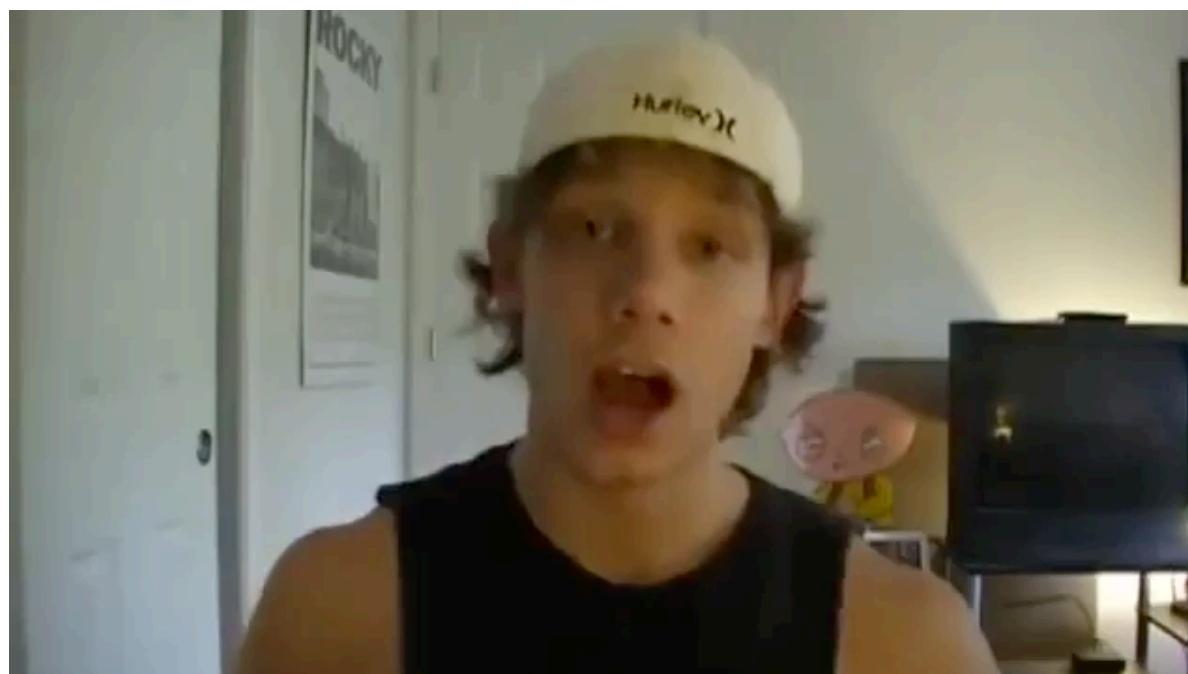
Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L
the action is fucking awesome	head nod + excited voice	H
[l] it was cute you know the actors did a great job bringing the smurfs to life such as joe george lopez neil patrick harris katy perry and a fourth	multiple smiles	H



# Towards explainable representations

**Goal: Use low/high values or MDM to explain representations**

Spoken Transcripts	Acoustic and visual behaviour	MDM
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L
the action is fucking awesome	head nod + excited voice	H
[l] it was cute you know the actors did a great job bringing the smurfs to life such as joe george lopez neil patrick harris katy perry and a fourth	multiple smiles	H



## From NLU to NLG

---



## From NLU to NLG

---

To handle an interaction, conversational agents are required to **understand the user**.



## From NLU to NLG

---

To handle an interaction, conversational agents are required to **understand the user**.



**When only language modality is available.**

## From NLU to NLG

---

To handle an interaction, conversational agents are required to **understand the user**.



**When only language modality is available.**

**When only multiple modalities are available.**

## From NLU to NLG

---

To handle an interaction, conversational agents are required to **understand the user**.



**When only language modality is available.**

**InfoMax Principle**

Linsker et al 1988

**When only multiple modalities are available.**

## From NLU to NLG

---

To handle an interaction, conversational agents are required to **understand the user**.



When **only language modality** is available.

**InfoMax Principle**

Linsker et al 1988

When **only multiple modalities** are available.

**To interact with the user the Conversational Agent need also to produce text.**

## **2. How to Use the Geometrical Properties of the Measures of Information to Generate and Evaluate Generated Text?**

# Estimation of Mutual Information

---

# Estimation of Mutual Information

---

**InfoMax Principle**

**Find**  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

# Estimation of Mutual Information

---

**InfoMax Principle**

**Find**  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

**MI is hard to compute in high dimension!**

Paninski 2003

Pichler et al 2020

# Estimation of Mutual Information

InfoMax Principle

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

MI is hard to compute in high dimension!

Paninski 2003

Pichler et al 2020

Existing estimators of MI

MINE

Belghazi et al 2018

Lower bounds

InfoNCE

Oord et al 2018

# Estimation of Mutual Information

InfoMax Principle

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

MI is hard to compute in high dimension!

Paninski 2003

Pichler et al 2020

Existing estimators of MI

Lower bounds

MINE

Belghazi et al 2018

InfoNCE

Oord et al 2018

Lower bounds can be maximised!

# Estimation of Mutual Information

InfoMax Principle

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

MI is hard to compute in high dimension!

Paninski 2003

Pichler et al 2020

Existing estimators of MI

MINE

Belghazi et al 2018

Lower bounds

InfoNCE

Oord et al 2018

Lower bounds can be maximised!

What if we want to minimise the Mutual Information?

# Estimation of Mutual Information

InfoMax Principle

Find  $\theta \in \Theta$  such that  $\theta = \operatorname{argmax} I(A, f_\theta(A))$

MI is hard to compute in high dimension!

Paninski 2003

Pichler et al 2020

Existing estimators of MI

MINE

Belghazi et al 2018

Lower bounds

InfoNCE

Oord et al 2018

Lower bounds can be maximised!

What if we want to minimise the Mutual Information?

We would prefer to rely on an upper bound!

CLUB

Cheng et al 2020

# Learning to Disentangle representations

# Learning to Disentangle representations

---

**Goal** Learn  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$  that retains as much as possible information of the original content from the input sentence  $X \in \mathcal{X}$  but as little as possible about the sensitive attribute  $Y \in \mathcal{Y}$

# Learning to Disentangle representations

---

**Goal** Learn  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$  that retains as much as possible information of the original content from the input sentence  $X \in \mathcal{X}$  but as little as possible about the sensitive attribute  $Y \in \mathcal{Y}$

**Minimize**  $I(f_\theta(X); Y)$

# Learning to Disentangle representations

---

**Goal** Learn  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$  that retains as much as possible information of the original content from the input sentence  $X \in \mathcal{X}$  but as little as possible about the sensitive attribute  $Y \in \mathcal{Y}$

Minimize  $I(f_\theta(X); Y)$

**Hypothesis**  $Y$  is a discrete attribute or concept.

# Learning to Disentangle representations

**Goal** Learn  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$  that retains as much as possible information of the original content from the input sentence  $X \in \mathcal{X}$  but as little as possible about the sensitive attribute  $Y \in \mathcal{Y}$

Minimize  $I(f_\theta(X); Y)$

**Hypothesis**  $Y$  is a discrete attribute or concept.

## Applications

### Audio & Video processing

Hsieh et al., 2018   Sanchez et al., 2019

### Style transfer

Fu et al., 2017

### Visual Reasoning

van Steenkiste et al., 2018

### Fair Classification

Elazar et al., 2018

### Conditional Generation

Denton et al., 2020

# **Application to Polarity Transfer**

---

# Application to Polarity Transfer

## Task

*X* I really hate these stupid cats

*Y* Negative

# Application to Polarity Transfer

## Task

$X$  I really hate these stupid cats

$Y$  Negative



$\hat{X}$  I love this wonderful cat!

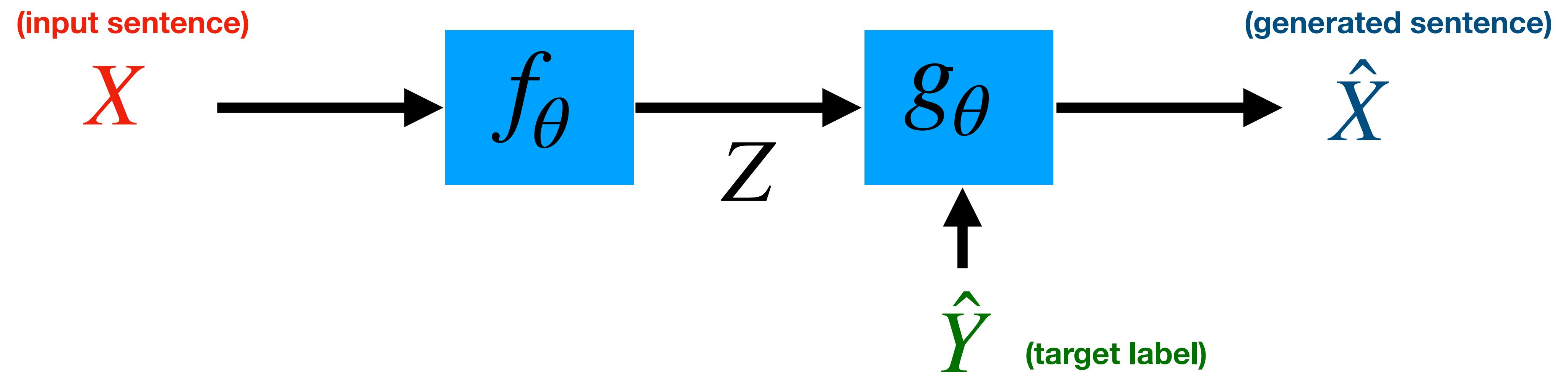
$\hat{Y}$  Positive

# Application to Polarity Transfer

## Task

$X$  I really hate these stupid cats →  $\hat{X}$  I love this wonderful cat!  
 $Y$  Negative →  $\hat{Y}$  Positive

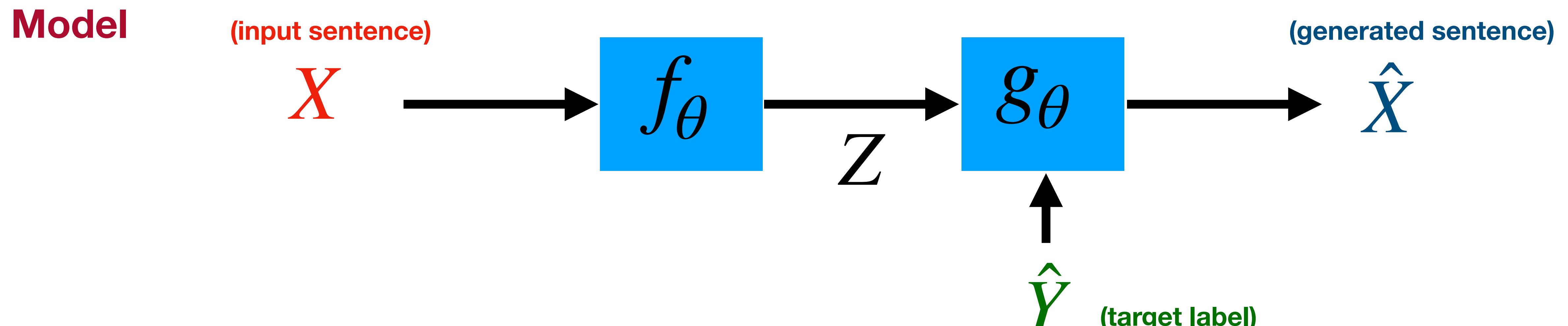
## Model



# Application to Polarity Transfer

## Task

$X$  I really hate these stupid cats →  $\hat{X}$  I love this wonderful cat!  
 $Y$  Negative →  $\hat{Y}$  Positive



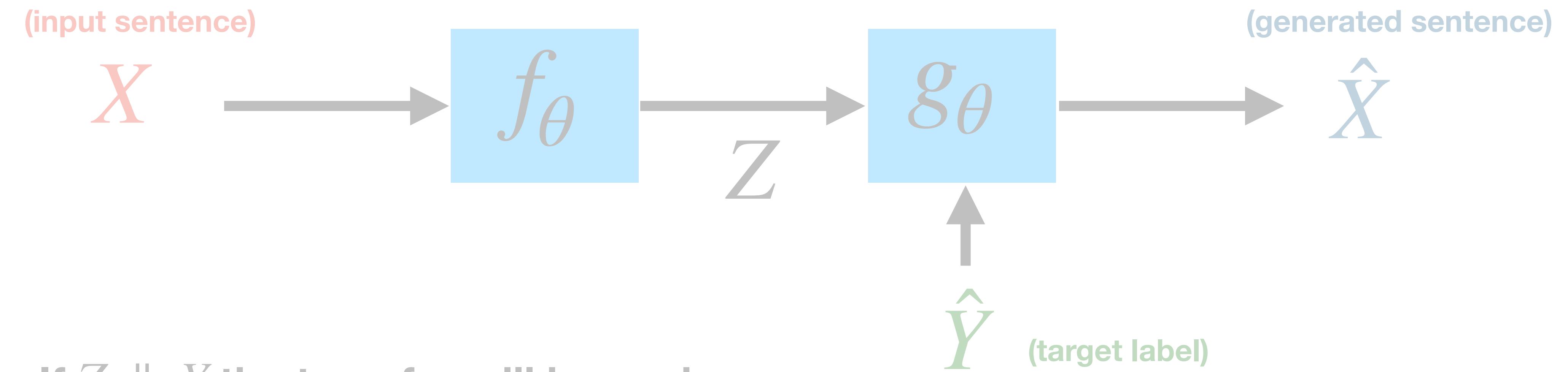
**Assumption:** If  $Z \perp\!\!\!\perp Y$  the transfer will be easier

# Application to Polarity Transfer

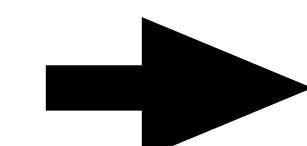
## Task

$X$  I really hate these stupid cats →  $\hat{X}$  I love this wonderful cat!  
 $Y$  Negative →  $\hat{Y}$  Positive

## Model

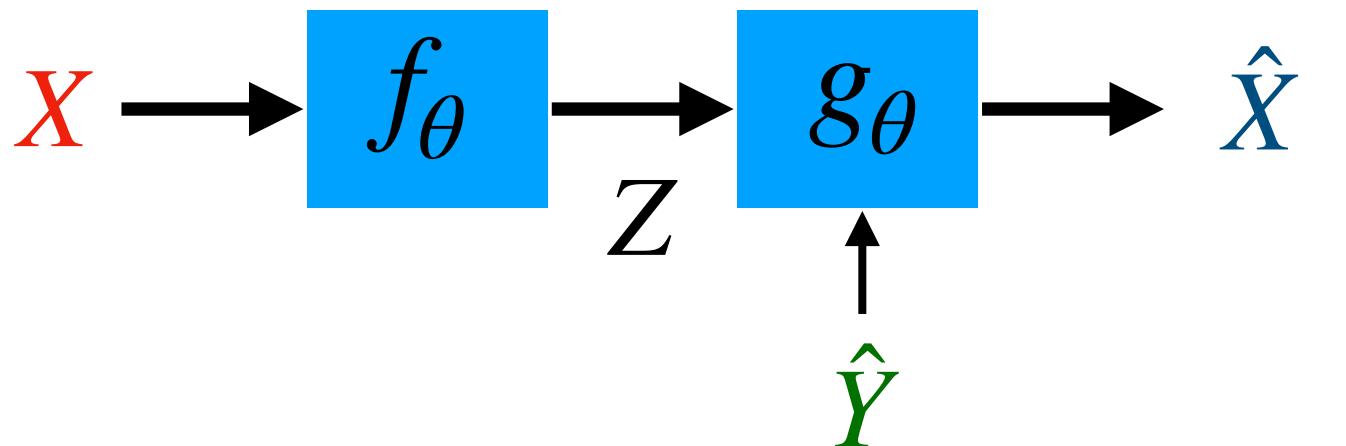


**Assumption:** If  $Z \perp\!\!\!\perp Y$  the transfer will be easier



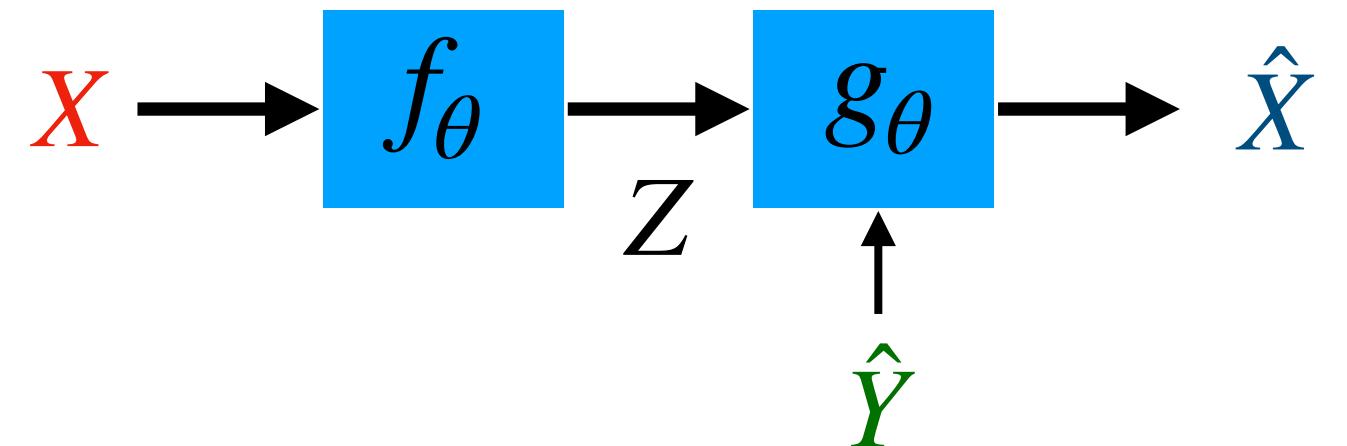
**Learn to disentangle  $Z$  and  $Y$**

## MI & Training Objective



## MI & Training Objective

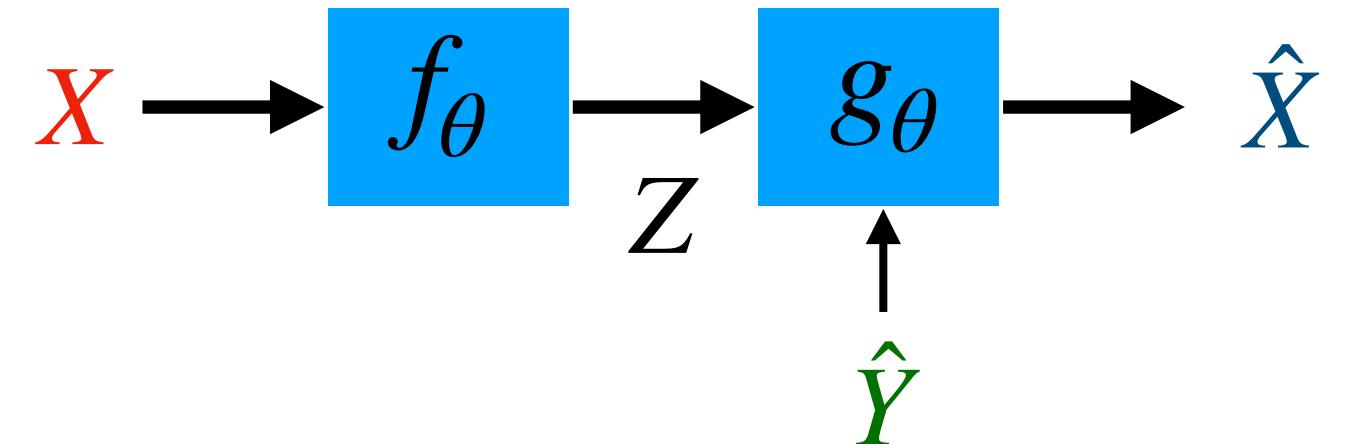
**Objective**  $\mathcal{L} = \underbrace{\mathcal{L}_{down.}(f_\theta, g_{\theta_d})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_\theta(X); Y)}_{\text{disentangled}}$



## MI & Training Objective

Objective

$$\mathcal{L} = \underbrace{\mathcal{L}_{down.}(f_\theta, g_{\theta_d})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_\theta(X); Y)}_{\text{disentangled}}$$



New upper bound on MI

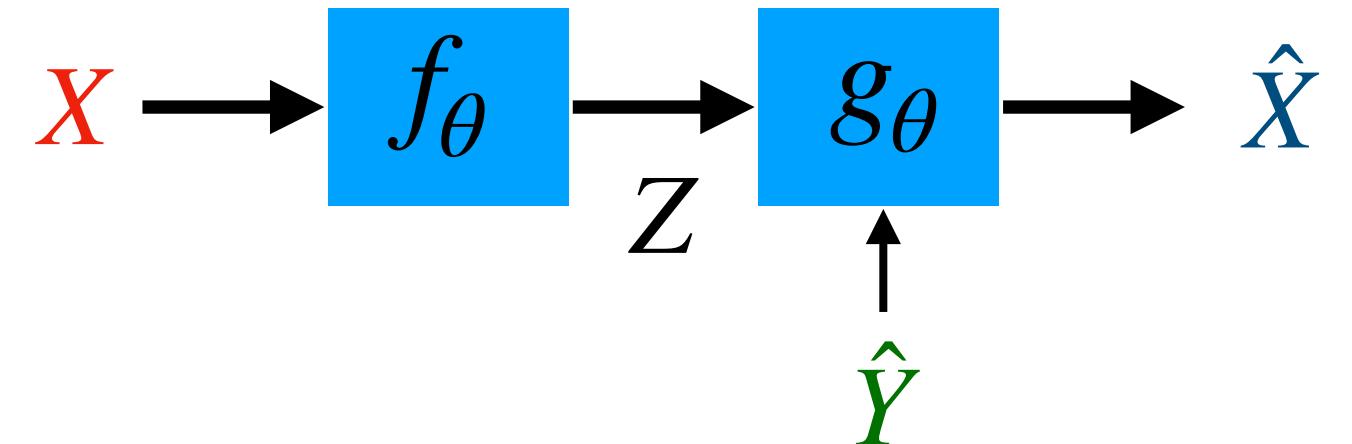
$$I(Z; Y) \leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z) p_Z(z) dz \right] + \mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] + D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}),$$

Renyi, 1960

## MI & Training Objective

Objective

$$\mathcal{L} = \underbrace{\mathcal{L}_{down.}(f_\theta, g_{\theta_d})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_\theta(X); Y)}_{\text{disentangled}}$$



New upper bound on MI

$$I(Z; Y) \leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z)p_Z(z)dz \right] + \mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] + D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}),$$

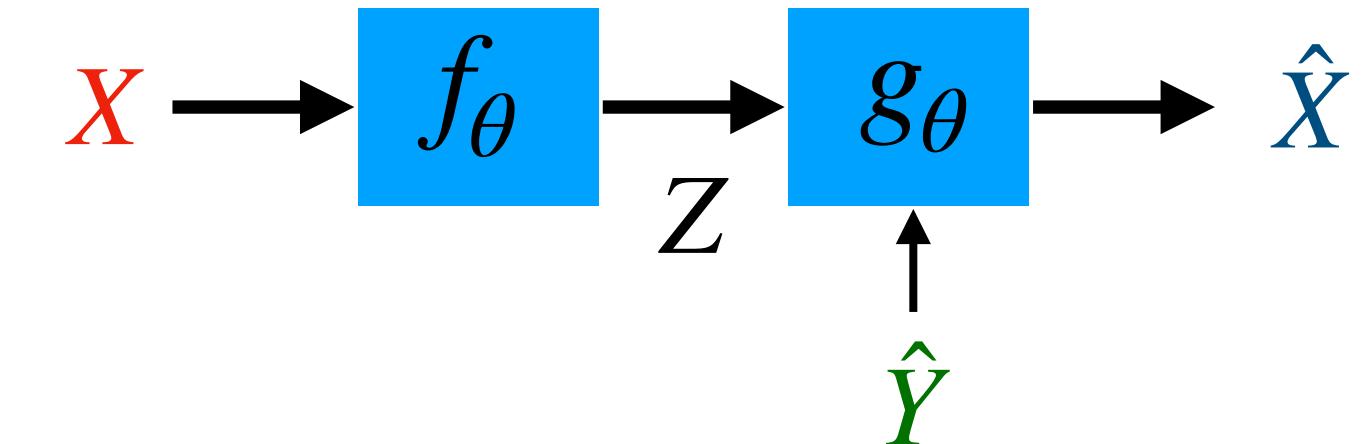
Renyi, 1960

Upper bound on  $h(Y)$

## MI & Training Objective

**Objective**

$$\mathcal{L} = \underbrace{\mathcal{L}_{down.}(f_\theta, g_{\theta_d})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_\theta(X); Y)}_{\text{disentangled}}$$



## New upper bound on MI

$$I(Z; Y) \leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z)p_Z(z)dz \right]$$

$$+ \mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] + D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}),$$

Renyi, 1960

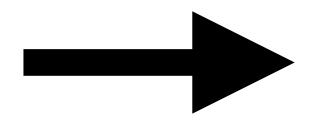
**Upper bound on  $h(Y)$**

**Lower bound on  $h(Z|Y)$**

## **Experimental Setting**

---

I really hate these stupid cats



I love this wonderful cat!

# Experimental Setting

---

I really hate these stupid cats → I love this wonderful cat!

## Baselines

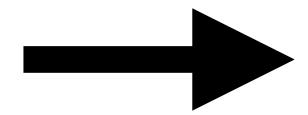
### Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect  
disentanglement

# Experimental Setting

I really hate these stupid cats



I love this wonderful cat!

## Baselines

### Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect disentanglement

### MI Estimators

#### NWJ / MINE

Belgazi et al 2018

Do not converge

#### CLUB

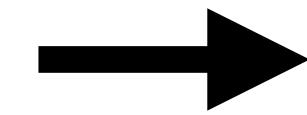
Cheng et al 2020

No control over the degree of disentanglement

Feutry et al 2017

# Experimental Setting

I really hate these stupid cats



I love this wonderful cat!

## Baselines

### Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect disentanglement

### MI Estimators

#### NWJ / MINE

Belgazi et al 2018

Do not converge

#### CLUB

Cheng et al 2020

No control over the degree of disentanglement

Feutry et al 2017

## Evaluation

### Disentanglement

Adversary trained from scratch

How disentangled is Z from Y?

Accuracy (lower is better)

# Experimental Setting

I really hate these stupid cats



I love this wonderful cat!

## Baselines

### Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect disentanglement

### MI Estimators

### NWJ / MINE

Belgazi et al 2018

Do not converge

### CLUB

Cheng et al 2020

No control over the degree of disentanglement

Feutry et al 2017

## Evaluation

### Disentanglement

Adversary trained from scratch

How disentangled is  $Z$  from  $Y$ ?

Accuracy (lower is better)

### Content Preservation

N-gram overlap between  $X$  and  $\hat{X}$

How much content is preserved between  $X$  and  $\hat{X}$

BLEU score (higher is better)

# Experimental Setting

I really hate these stupid cats → I love this wonderful cat!

## Baselines

Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect disentanglement

MI Estimators

NWJ / MINE

Belgazi et al 2018

Do not converge

CLUB

Cheng et al 2020

No control over the degree of disentanglement

Feutry et al 2017

## Evaluation

### Disentanglement

Adversary trained from scratch

How disentangled is  $Z$  from  $Y$ ?

Accuracy (lower is better)

### Content Preservation

N-gram overlap between  $X$  and  $\hat{X}$

How much content is preserved between  $X$  and  $\hat{X}$

BLEU score (higher is better)

### Fluency

Perplexity of  $\hat{X}$

How fluent (i.e. well formed) is  $\hat{X}$  ?

Perplexity score (lower is better)

# Experimental Setting

I really hate these stupid cats → I love this wonderful cat!

## Baselines

### Adversarial Losses (ADV)

Lample et al 2019

Can not achieve perfect disentanglement

### MI Estimators

### NWJ / MINE

Belgazi et al 2018

Do not converge

### CLUB

Cheng et al 2020

No control over the degree of disentanglement

Feutry et al 2017

## Evaluation

### Disentanglement

Adversary trained from scratch

How disentangled is  $Z$  from  $Y$ ?

Accuracy (lower is better)

### Content Preservation

N-gram overlap between  $X$  and  $\hat{X}$

How much content is preserved between  $X$  and  $\hat{X}$

BLEU score (higher is better)

### Fluency

Perplexity of  $\hat{X}$

How fluent (i.e. well formed) is  $\hat{X}$  ?

Perplexity score (lower is better)

### Transfert accuracy

Sentiment predictor

Is the target style ( $\hat{Y}$ ) present in  $\hat{X}$  ?

Accuracy (higher is better)

# **Text Style Transfert**

---

# **Text Style Transfert**

---

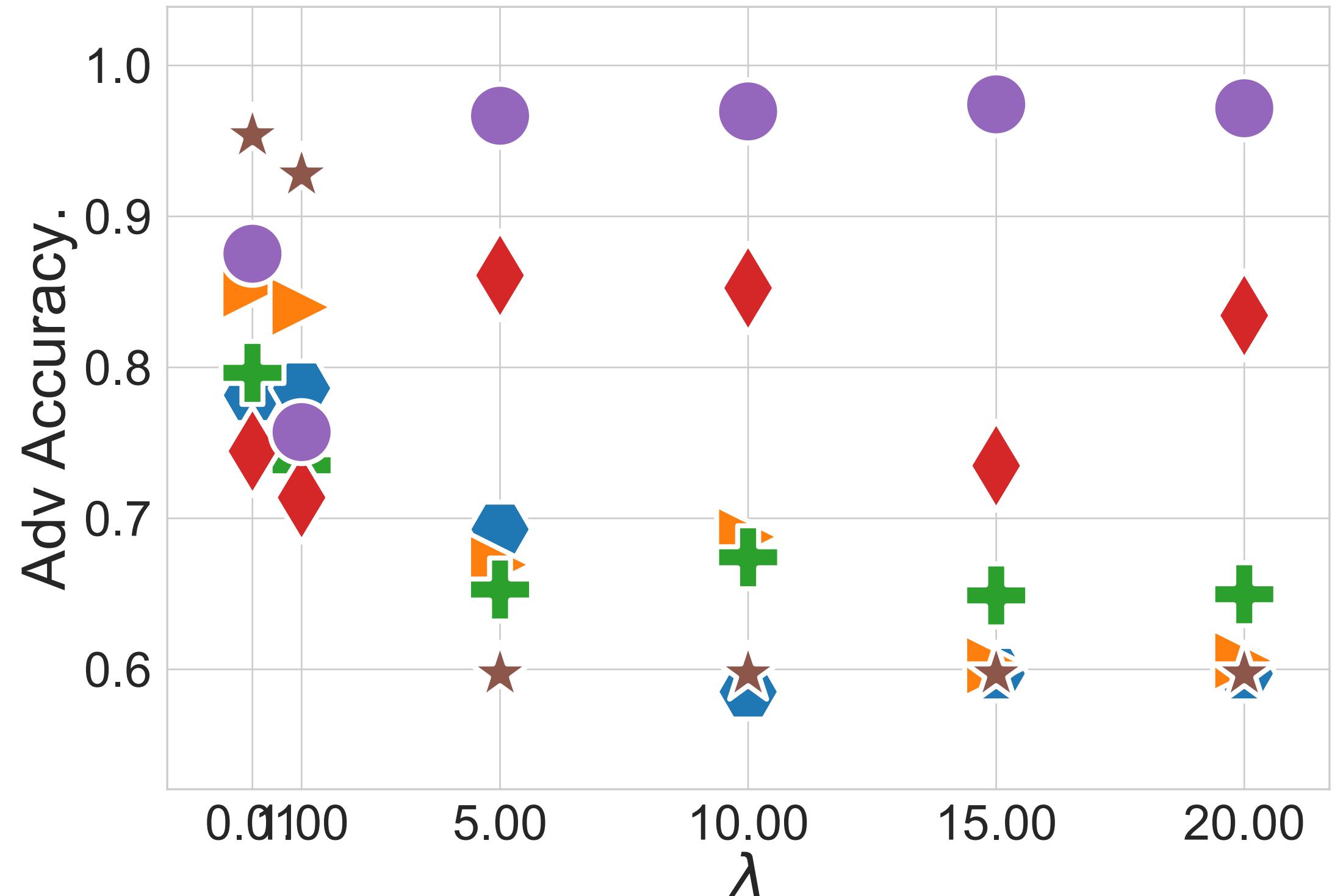
## **Task**

**I really hate these stupid cats** → **I love this wonderful cat !**

**Negative**                                   **Positive**

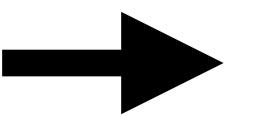
# Text Style Transfert

# Disentanglement



# Task

# I really hate these stupid cats



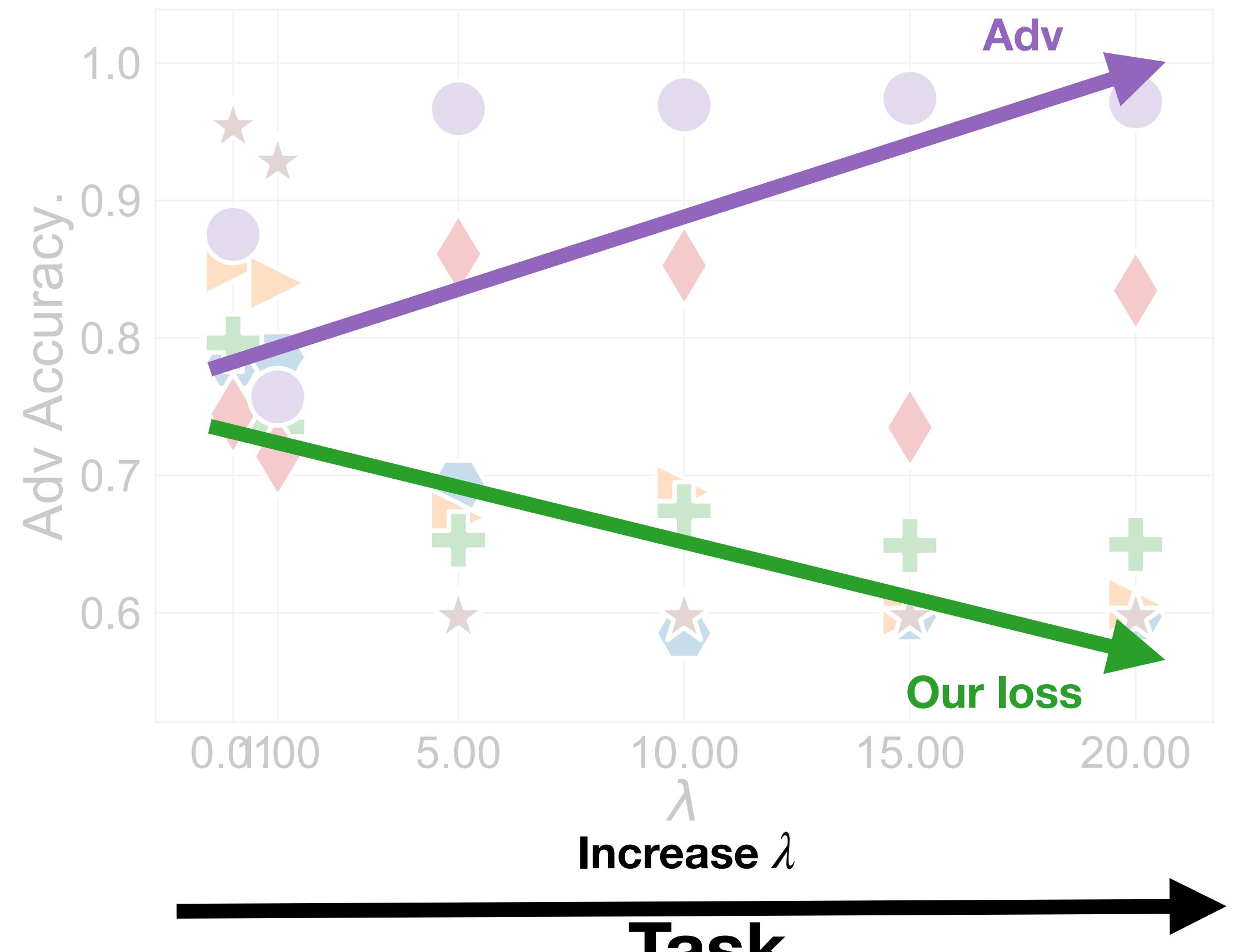
# I love this wonderful cat !

# Negative

# Positive

# Text Style Transfert

## Disentanglement

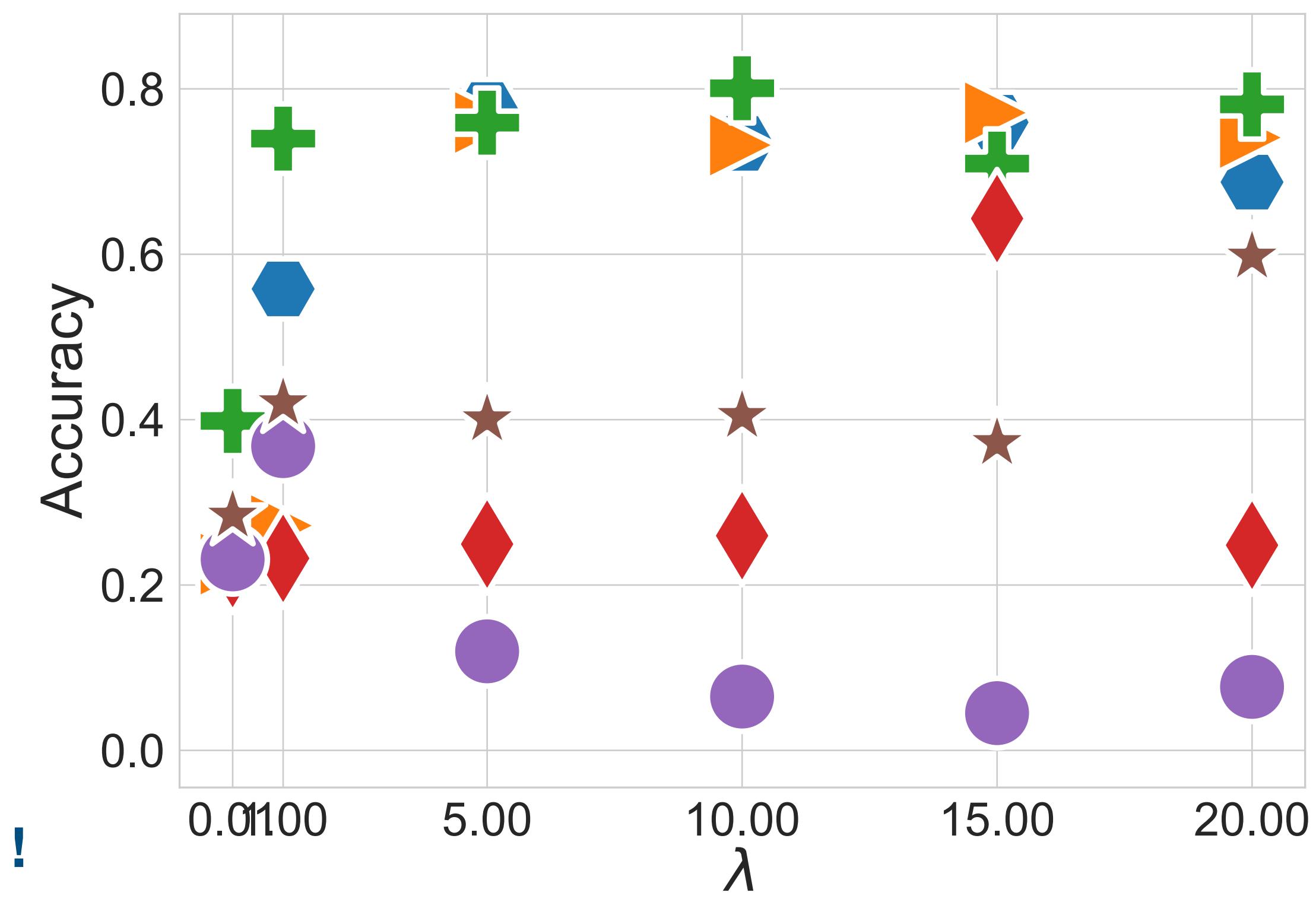


I really hate these stupid cats → I love this wonderful cat !

Negative                              Positive

# Text Style Transfert

## Disentanglement



# Text Style Transfert

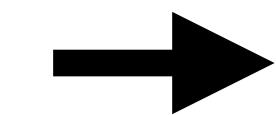
## Disentanglement



## Style Accuracy



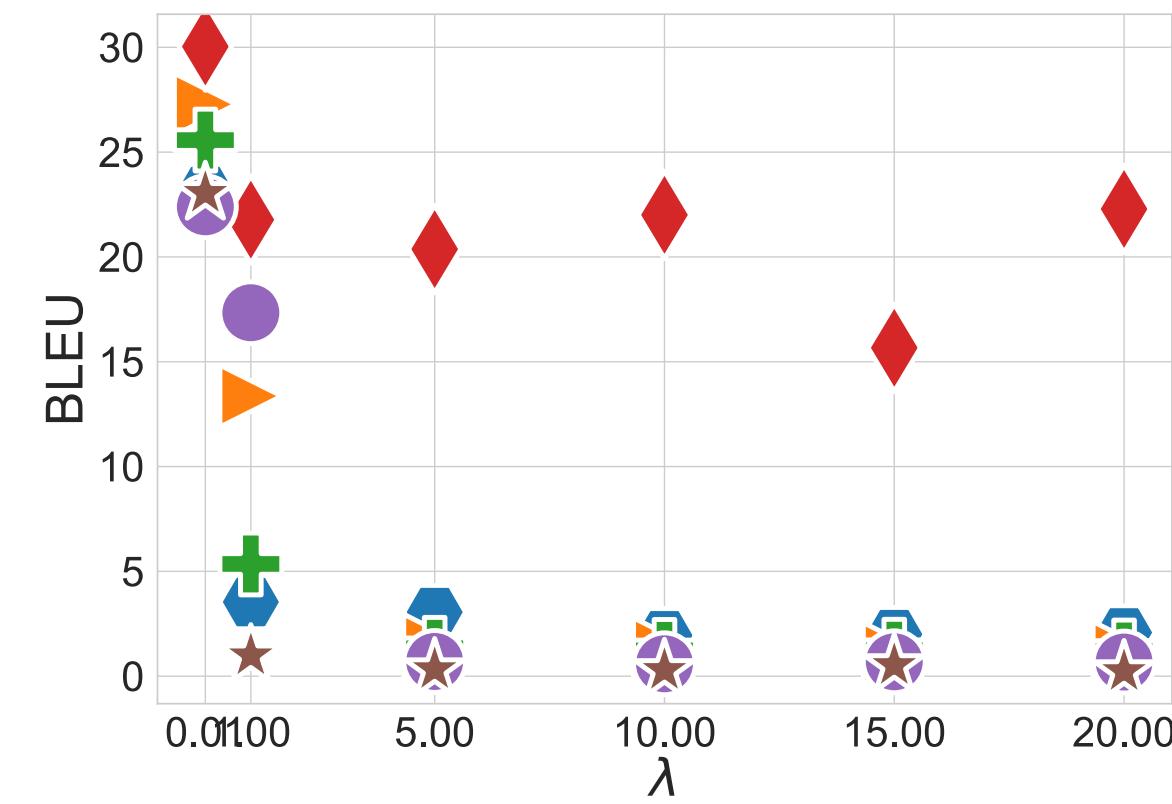
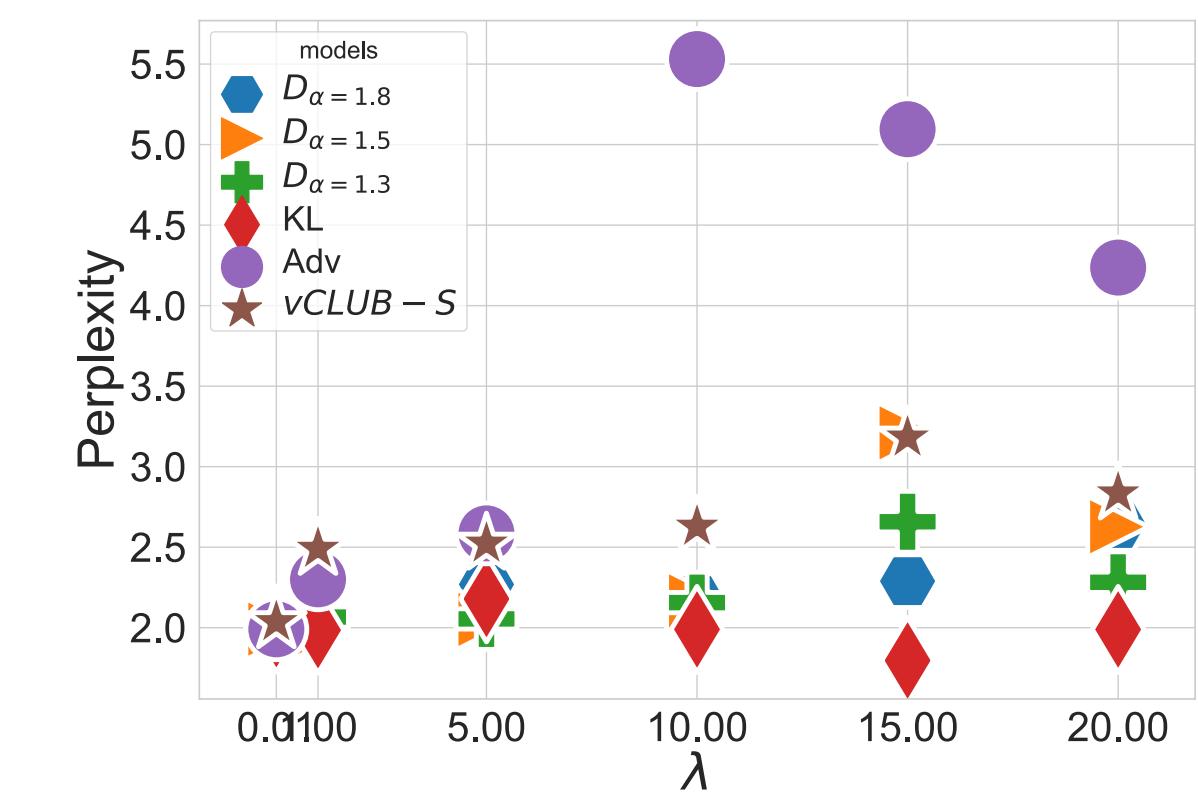
I really hate these stupid cats  
Negative



I love this wonderful cat !  
Positive

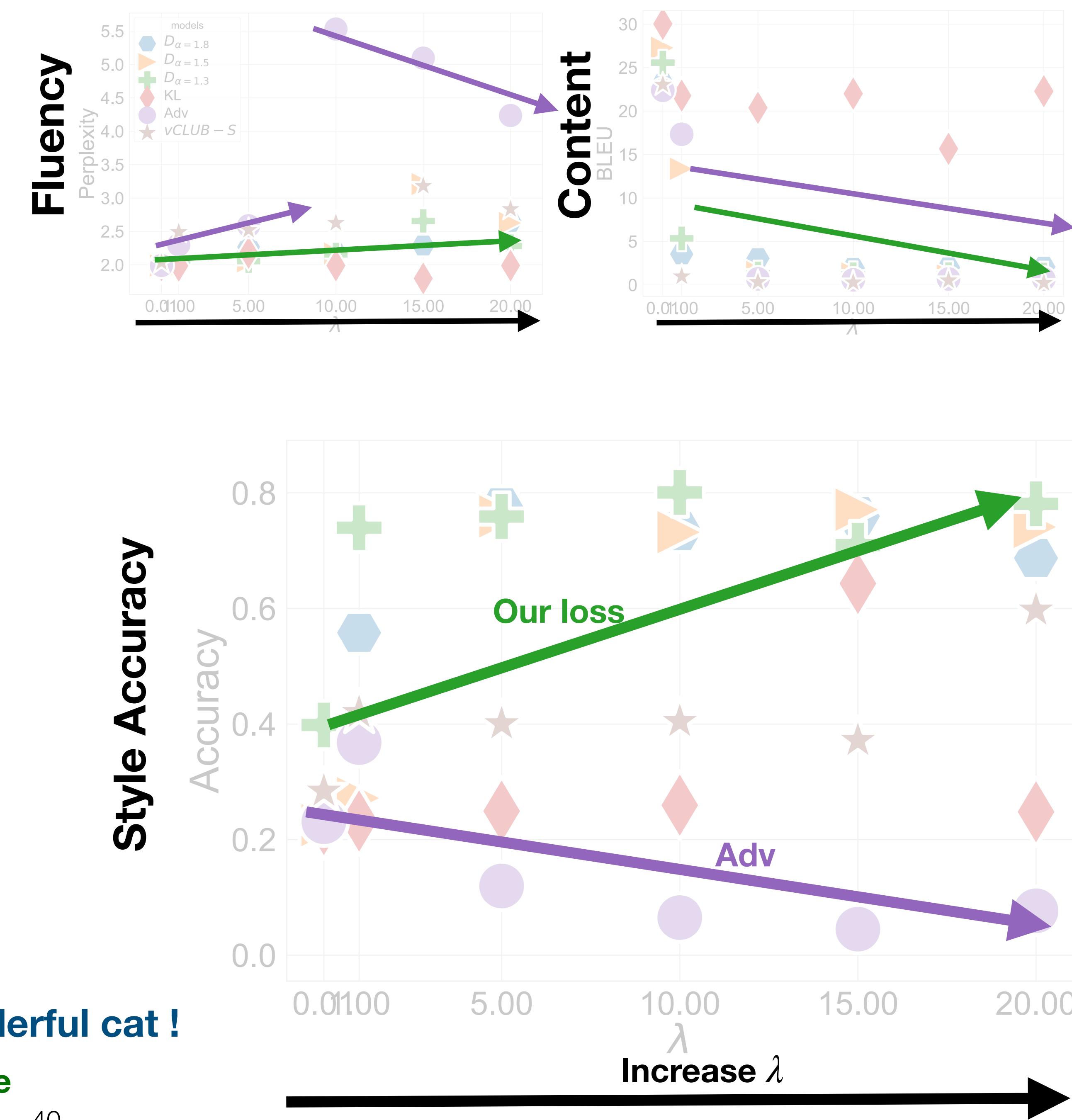
# Text Style Transfert

## Disentanglement

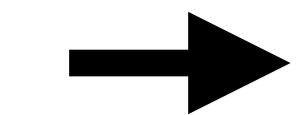


# Text Style Transfert

## Disentanglement



I really hate these stupid cats  
Negative



I love this wonderful cat !  
Positive

# Towards Better Evaluation of NLG

---

# Towards Better Evaluation of NLG

---

**Transferring style is easier with  
disentangled representations**

# Towards Better Evaluation of NLG

---

Transferring style is easier with  
disentangled representations

There is no free lunch!

# Towards Better Evaluation of NLG

---

Transferring style is easier with disentangled representations

There is no free lunch!

Disentangling also removes important information about the content.

# Towards Better Evaluation of NLG

---

Transferring style is easier with  
disentangled representations

There is no free lunch!

Disentangling also removes  
important information about the  
content.

**Evaluation of content preservation based on BLEU score.**

# Towards Better Evaluation of NLG

---

Transferring style is easier with disentangled representations

There is no free lunch!

Disentangling also removes important information about the content.

**Evaluation of content preservation based on BLEU score.**

This is not a robust evaluation as BLEU does not handle synonyms!

# Towards Better Evaluation of NLG

---

Transferring style is easier with disentangled representations

There is no free lunch!

Disentangling also removes important information about the content.

**Evaluation of content preservation based on BLEU score.**

This is not a robust evaluation as BLEU does not handle synonyms!

How to draw the right conclusions if we cannot properly evaluate NLG?

# Towards Better Evaluation of NLG

Transferring style is easier with disentangled representations

There is no free lunch!

Disentangling also removes important information about the content.

Evaluation of content preservation based on BLEU score.

This is not a robust evaluation as BLEU does not handle synonyms!

How to draw the right conclusions if we cannot properly evaluate NLG?

## What is automatic evaluation?

R: The weather is cold today.



C: It is freezing today

R: I like those cats.



C: It is freezing today

# **Importance of Evaluation of NLG**

---

## **Importance of Evaluation of NLG**

---

**Why do we rely on the automatic evaluation?**

# **Importance of Evaluation of NLG**

---

**Why do we rely on the automatic evaluation?**

- 1. Cheap:** compared to human evaluation.

# **Importance of Evaluation of NLG**

---

**Why do we rely on the automatic evaluation?**

- 1. Cheap:** compared to human evaluation.
- 2. Fast:** you can label “instantaneously”.

# **Importance of Evaluation of NLG**

---

## **Why do we rely on the automatic evaluation?**

- 1. Cheap:** compared to human evaluation.
- 2. Fast:** you can label “instantaneously”.
- 3. Reproducible:** two sentence always get the same score.

# **Importance of Evaluation of NLG**

---

## **Why do we rely on the automatic evaluation?**

- 1. Cheap: compared to human evaluation.**
- 2. Fast: you can label “instantaneously”.**
- 3. Reproducible: two sentence always get the same score.**
- 4. Easy to use: don’t need to train annotators, ask the right questions .....**

# **Importance of Evaluation of NLG**

---

**Why do we rely on the automatic evaluation?**

- 1. Cheap:** compared to human evaluation.
- 2. Fast:** you can label “instantaneously”.
- 3. Reproducible:** two sentence always get the same score.
- 4. Easy to use: don’t need to train annotators, ask the right questions .....**

**In which case do we use them?**

# **Importance of Evaluation of NLG**

---

Why do we rely on the automatic evaluation?

1. Cheap: compared to human evaluation.
2. Fast: you can label “instantaneously”.
3. Reproducible: two sentence always get the same score.
4. Easy to use: don’t need to train annotators, ask the right questions .....

In which case do we use them?

1. **Debug NLG systems without annotators.**

# **Importance of Evaluation of NLG**

---

Why do we rely on the automatic evaluation?

1. Cheap: compared to human evaluation.
2. Fast: you can label “instantaneously”.
3. Reproducible: two sentence always get the same score.
4. Easy to use: don’t need to train annotators, ask the right questions .....

In which case do we use them?

1. Debug NLG systems without annotators.
2. Improve learning of systems by deriving new losses.

# **Importance of Evaluation of NLG**

---

Why do we rely on the automatic evaluation?

1. **Cheap**: compared to human evaluation.
2. **Fast**: you can label “instantaneously”.
3. **Reproducible**: two sentence always get the same score.
4. **Easy** to use: don’t need to train annotators, ask the right questions .....

In which case do we use them?

1. **Debug NLG systems without annotators.**
2. **Improve learning of systems by deriving new losses.**
3. **Compare different systems.**

# Existing Methods

---

# Existing Methods

---

## Edit Based

Snover et al. 2006

### Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)

sailor -> sailir (S)

sailir -> sailin (S)

sailin\_ -> sailing (I)

Distance is 4 !

# Existing Methods

---

## Edit Based

Snover et al. 2006

### Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)

sailor -> sailir (S)

sailir -> sailin (S)

sailin\_ -> sailing (I)

Distance is 4 !

## N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

### Unigrams

C : I like these very nice pies !

R : I like those cakes !

### Bigrams

C : I like these very nice pies !

R : I like those cakes !

# Existing Methods

## Edit Based

Snover et al. 2006

### Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (**S**)

sailor -> sailir (**S**)

sailir -> sailin (**S**)

sailin\_ -> sailing (**I**)

Distance is 4 !

## N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

### Unigrams

C : I like these very nice pies !

R : I like those cakes !

### Bigrams

C : I like these very nice pies !

R : I like those cakes !

## Embedding Based

### Word Mover distance

Kusner et al. 2015

### BertScore

Zhang et al. 2019

### MoverScore

Zhao et al. 2019

### Sentence Mover

Clark et al. 2019

# Existing Methods

## Edit Based

Snover et al. 2006

### Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)

sailor -> sailir (S)

sailir -> sailin (S)

sailin -> sailing (I)

Distance is 4 !

## InfoLM

## N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

### Unigrams

C : I like these very nice pies !

R : I like those cakes !

### Bigrams

C : I like these very nice pies !

R : I like those cakes !

## Embedding Based

### Word Mover distance

Kusner et al. 2015

### BertScore

Zhang et al. 2019

### MoverScore

Zhao et al. 2019

### Sentence Mover

Clark et al. 2019

# Intuition of InfoLM

---

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

**Equivalence for masked contexts**  $\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$  **MLM predicts a distribution over  $\Omega$**   
 $p_{\Omega}(\cdot | [R]^i)$

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

<b>Equivalence for masked contexts</b>	$\mathcal{I} : [0,1]^{ \Omega } \times [0,1]^{ \Omega }$	<b>MLM predicts a distribution over <math>\Omega</math></b>
		$p_{\Omega}(\cdot   [R]^i)$
	MLM	

# Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

MLM

R: It is [MASK] today.

C: It is [MASK] this morning !

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

MLM

$$p_{\Omega}(\cdot | [R]^2)$$

$$p_{\Omega}(\cdot | [C]^2)$$



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

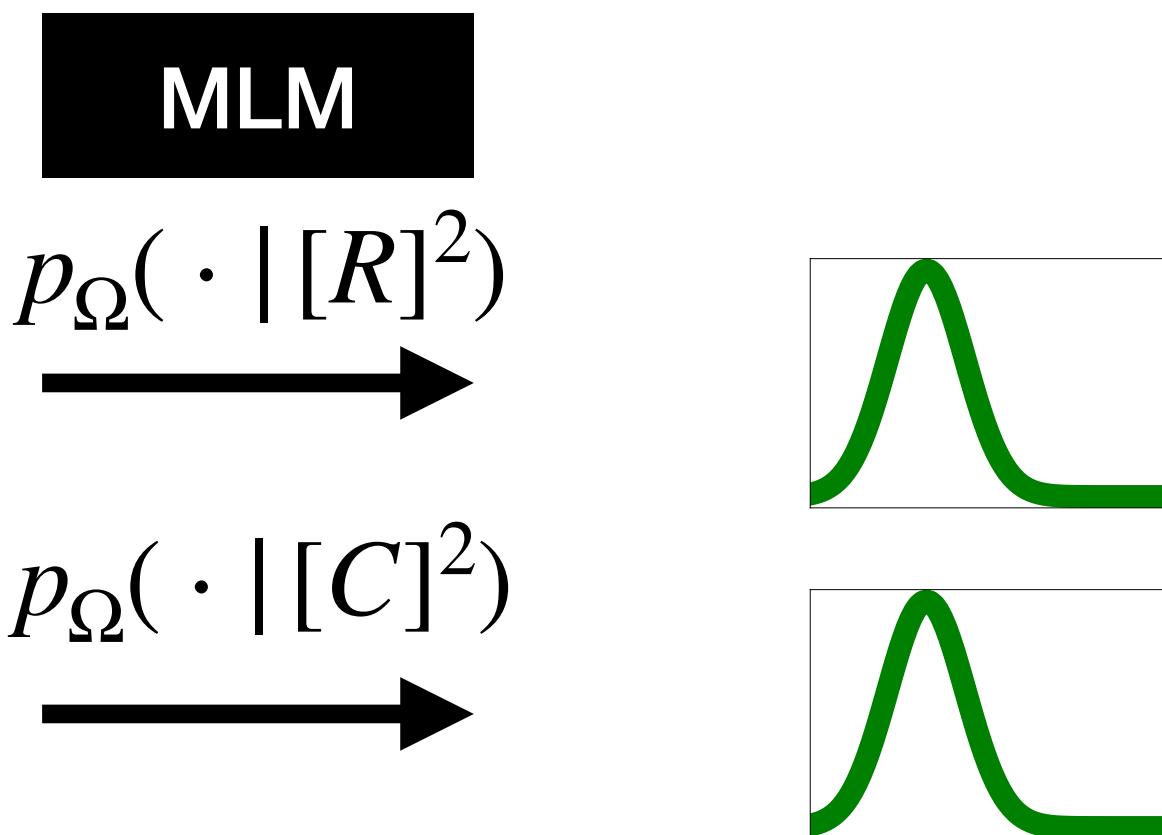
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

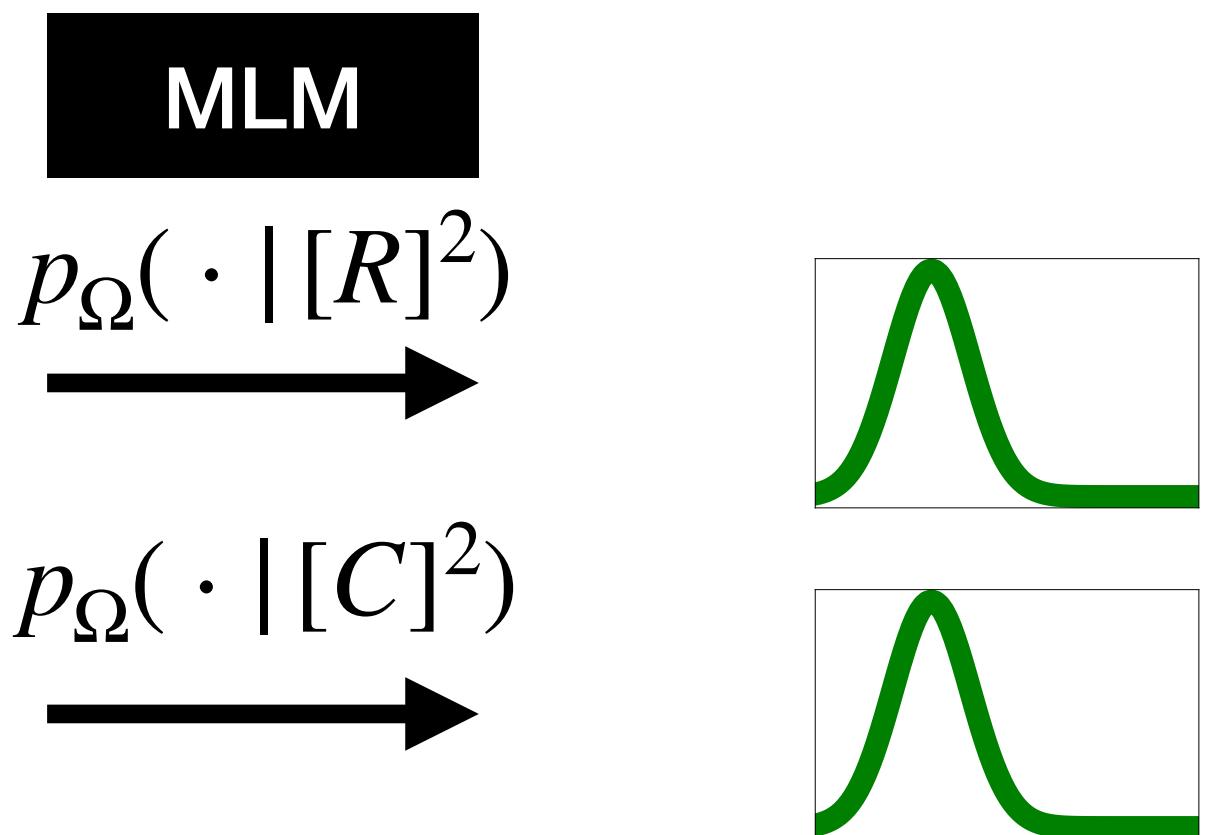
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



$$\mathcal{I}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

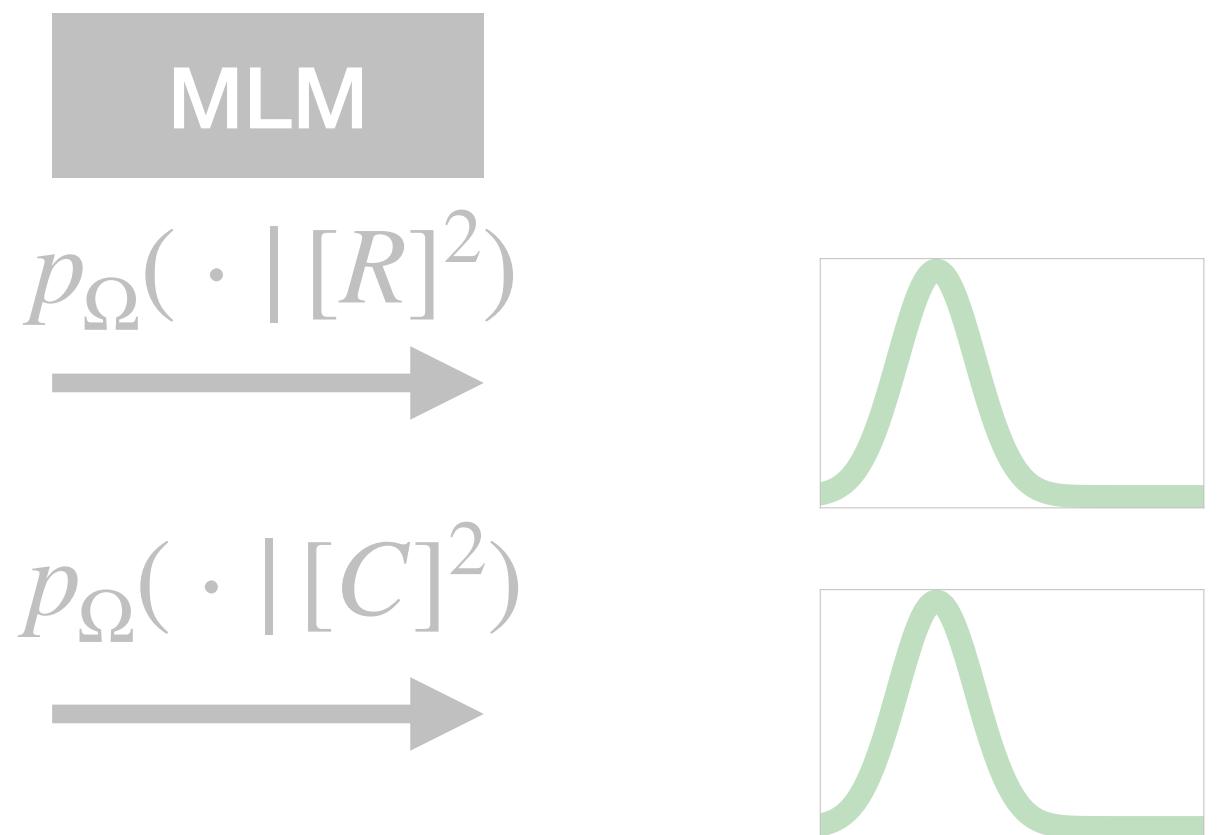
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



$$\mathcal{J}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

Aggregate the similar context to evaluate sentence similarity based on different measures of informations

Fisher Rao  
Rao 1987

AB-Divergences  
Cichocki, Cruces, and Amari 2011

$\alpha$ -divergences  
Csiszar 1967

$L_p$ -distances

# **Summary of NLG & Conclusions**

---

## **Summary of NLG & Conclusions**

---

**This thesis has addressed the interplay between NLP and the measures of informations**

## **Summary of NLG & Conclusions**

---

**This thesis has addressed the interplay between NLP and the measures of informations**

**Application of the measures to include multimodal and conversation dimensions in pretrained representations.**

## **Summary of NLG & Conclusions**

---

**This thesis has addressed the interplay between NLP and the measures of informations**

**Application of the measures to include multimodal and conversation dimensions in pretrained representations.**

**Motivation for our losses**

**Robustness**

**Prediction Accuracy**

**Interpretability**

# **Summary of NLG & Conclusions**

---

This thesis has addressed the interplay between NLP and the measures of informations

**Application of the measures to include multimodal and conversation dimensions in pretrained representations.**

Motivation for our losses

Robustness

Prediction Accuracy

Interpretability

**Application of the measures to natural language generation task.**

# **Summary of NLG & Conclusions**

---

This thesis has addressed the interplay between NLP and the measures of informations

**Application of the measures to include multimodal and conversation dimensions in pretrained representations.**

Motivation for our losses

Robustness

Prediction Accuracy

Interpretability

**Application of the measures to natural language generation task.**

New estimator of MI

Control the style of a sentence

Offers better trade-offs

# **Summary of NLG & Conclusions**

---

This thesis has addressed the interplay between NLP and the measures of informations

**Application of the measures to include multimodal and conversation dimensions in pretrained representations.**

Motivation for our losses

Robustness

Prediction Accuracy

Interpretability

**Application of the measures to natural language generation task.**

New estimator of MI

Control the style of a sentence

InfoLM

Offers better trade-offs

Measure similarity of between subwords

# Perspectives

---

## Perspectives

---

**Many Future works can be drawn from the works I presented**

# Perspectives

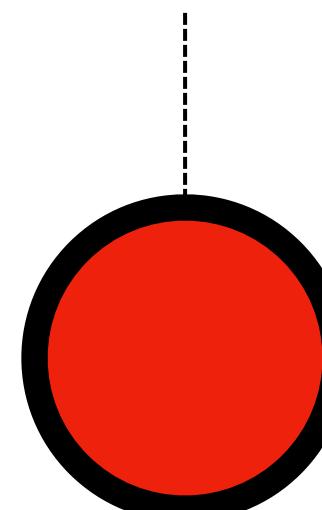
---

**Many Future works can be drawn from the works I presented**

## Applications

**A new loss for learning multilingual representations inspired by MI.**

**Chapuis\*, Colombo\* et al 2021**



# Perspectives

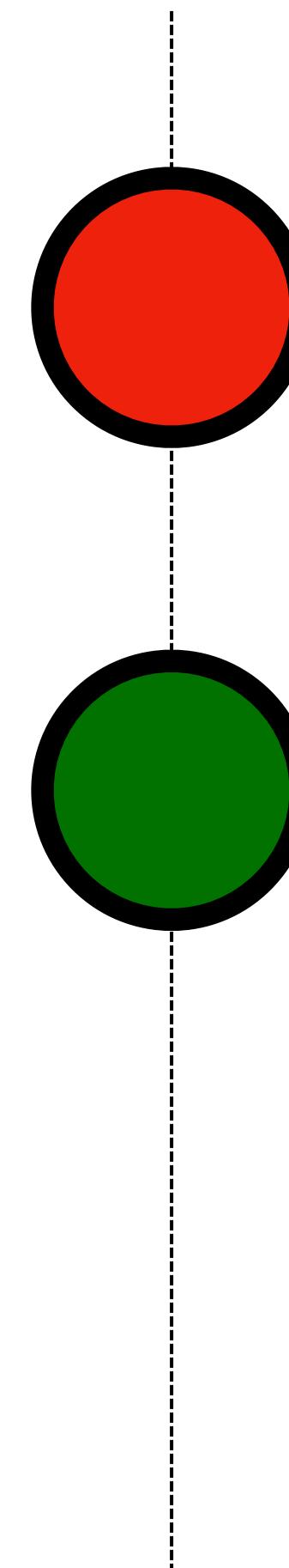
---

**Many Future works can be drawn from the works I presented**

## Applications

**A new loss for learning multilingual representations inspired by MI.**

**Chapuis\*, Colombo\* et al 2021**



## Estimation

**How about estimating other information theoretic quantities?**

**Pichlet\*, Colombo\* et al 2021**

# Perspectives

---

**Many Future works can be drawn from the works I presented**

## Applications

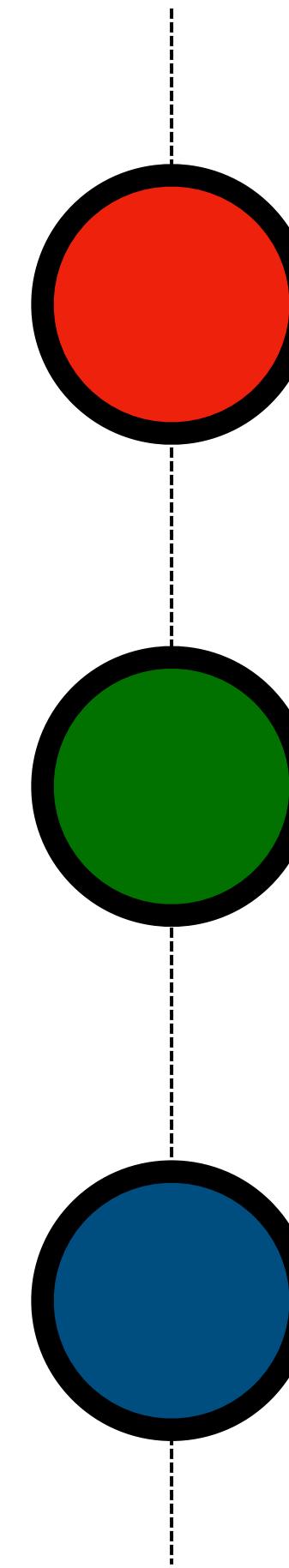
**A new loss for learning multilingual representations inspired by MI.**

**Chapuis\*, Colombo\* et al 2021**

## Evaluation

**Achieving disentangled representations with guarantees.**

**Extending InfoLM to other tasks including translation and Reference free evaluation**



## Estimation

**How about estimating other information theoretic quantities?**

**Pichlet\*, Colombo\* et al 2021**

# Acknowledgements

---

# Acknowledgements

---

This PhD work is the result of a CIFRE collaboration. It has been funded by IBM France.

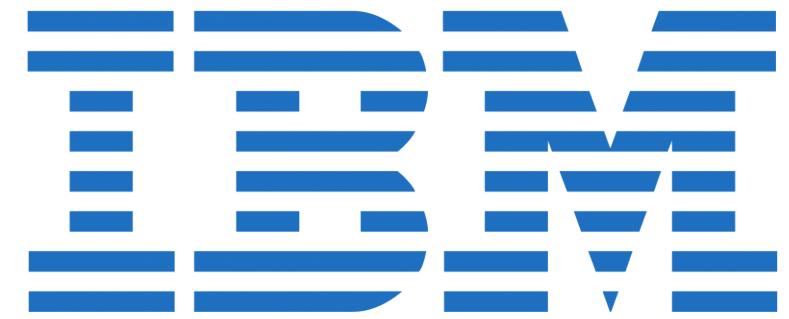
Computing Grant from GENCI.

Chloé Clavel



Giovanna Varni

Emmanuel Vignon



Joffrey Martinez



# Acknowledgements

---

This PhD work is the result of a CIFRE collaboration. It has been **funded by IBM France.**

Computing Grant from GENCI.

Chloé Clavel



Giovanna Varni



Emmanuel Vignon



Joffrey Martinez



**This work would no have been possible without my co-authors**

**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 Clémenton**

# **Learning to Represent and Generate Text using Information Measures**

**Thanks for your attention!**

# Intuition of InfoLM

---

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

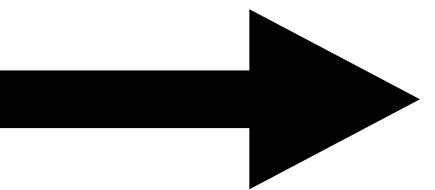
How to aggregate contexts?

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?

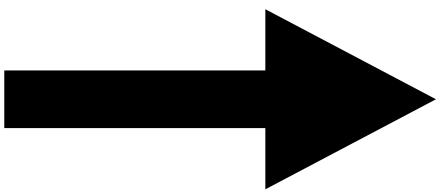


## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?



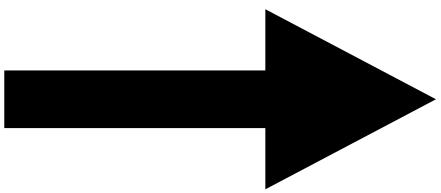
Weighted Sum!

# Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?



Weighted Sum!

Reference

[MASK] is cold today.

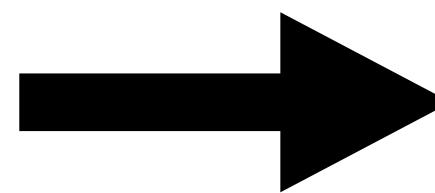
It is [MASK] today.  
...

It is cold today [MASK]  
...

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?



Weighted Sum!

Reference

[MASK] is cold today.

It is [MASK] today.

It is cold today [MASK]

Candidate

[MASK] is freezing this morning !

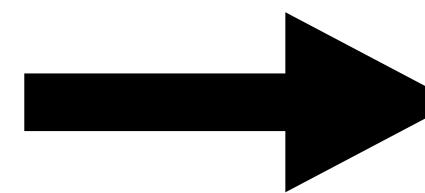
It is [MASK] this morning !

It is freezing this morning [MASK]

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?



Weighted Sum!

Reference

[MASK] is cold today.

It is [MASK] today.

...

It is cold today [MASK]

$$P \triangleq \frac{1}{5} \sum_{k=0}^4 \gamma_k \times p_{\Omega}(\cdot | [R]^k)$$

Candidate

[MASK] is freezing this morning !

It is [MASK] this morning !

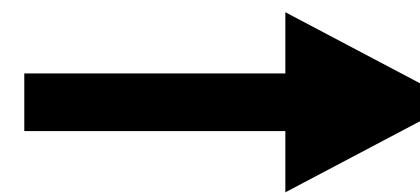
...

It is freezing this morning [MASK]

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

How to aggregate contexts?



Weighted Sum!

Reference

[MASK] is cold today.

It is [MASK] today.

...

It is cold today [MASK]

$$P \triangleq \frac{1}{5} \sum_{k=0}^4 \gamma_k \times p_{\Omega}(\cdot | [R]^k)$$

Candidate

[MASK] is freezing this morning !

It is [MASK] this morning !

...

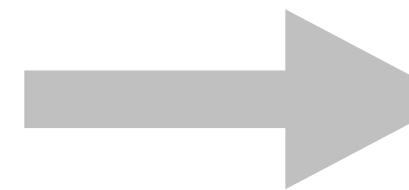
It is freezing this morning [MASK]

$$Q \triangleq \frac{1}{6} \sum_{k=0}^5 \gamma_k \times p_{\Omega}(\cdot | [C]^k)$$

# Intuition of InfoLM

Goal    Compute a similarity score between R and C.

How to aggregate contexts?



Weighted Sum!

Reference

[MASK] is cold today.

It is [MASK] today.

...

It is cold today [MASK]

Candidate

[MASK] is freezing this morning !

It is [MASK] this morning !

...

It is freezing this morning [MASK]

$$P \triangleq \frac{1}{5} \sum_{k=0}^4 \gamma_k \times p_{\Omega}(\cdot | [R]^k)$$

$$\text{InfoLM}(R, C) \triangleq \mathcal{J}(P, Q)$$

$$Q \triangleq \frac{1}{6} \sum_{k=0}^5 \gamma_k \times p_{\Omega}(\cdot | [C]^k)$$

# Intuition of InfoLM

---

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

## Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

**Equivalence for masked contexts**  $\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$  **MLM predicts a distribution over  $\Omega$**   
 $p_{\Omega}(\cdot | [R]^i)$

MLM

# Intuition of InfoLM

---

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

MLM

R: It is [MASK] today.

C: It is [MASK] this morning !

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is **[MASK]** today.

C: It is **[MASK]** this morning !

MLM

$$p_{\Omega}(\cdot | [R]^2)$$

$$p_{\Omega}(\cdot | [C]^2)$$



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

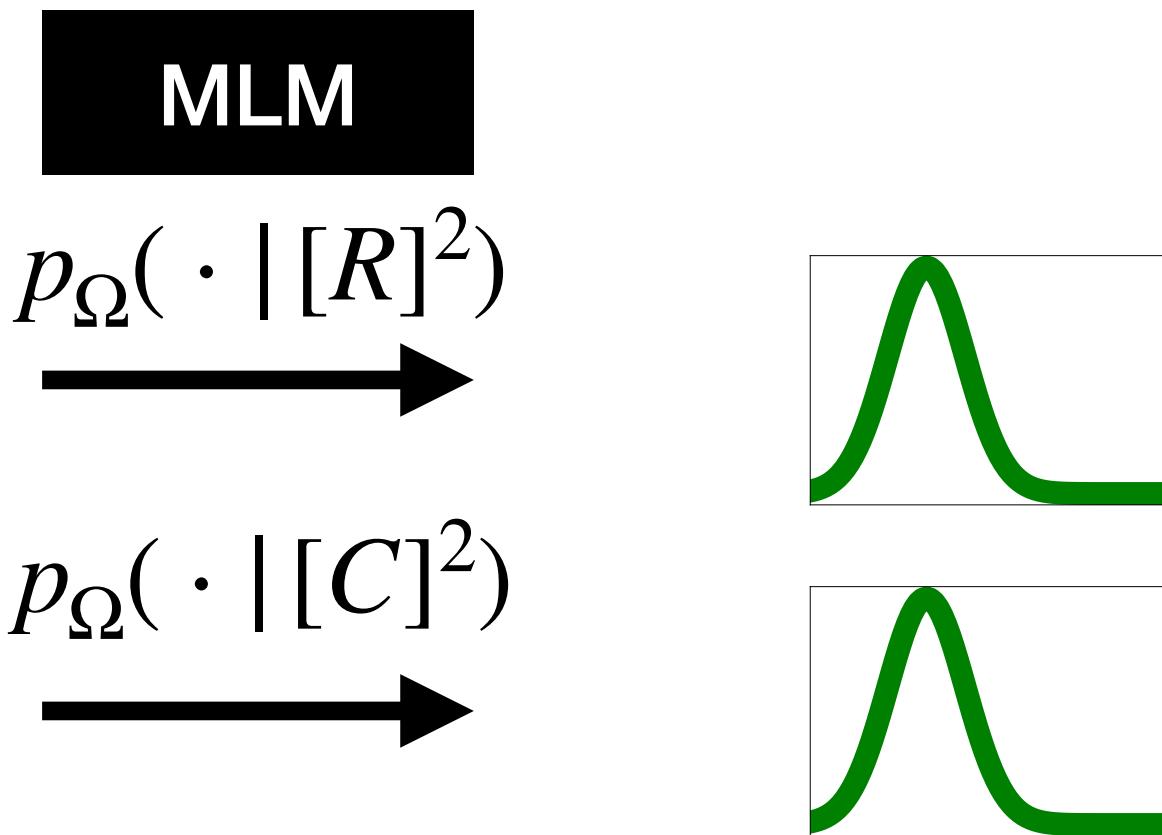
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

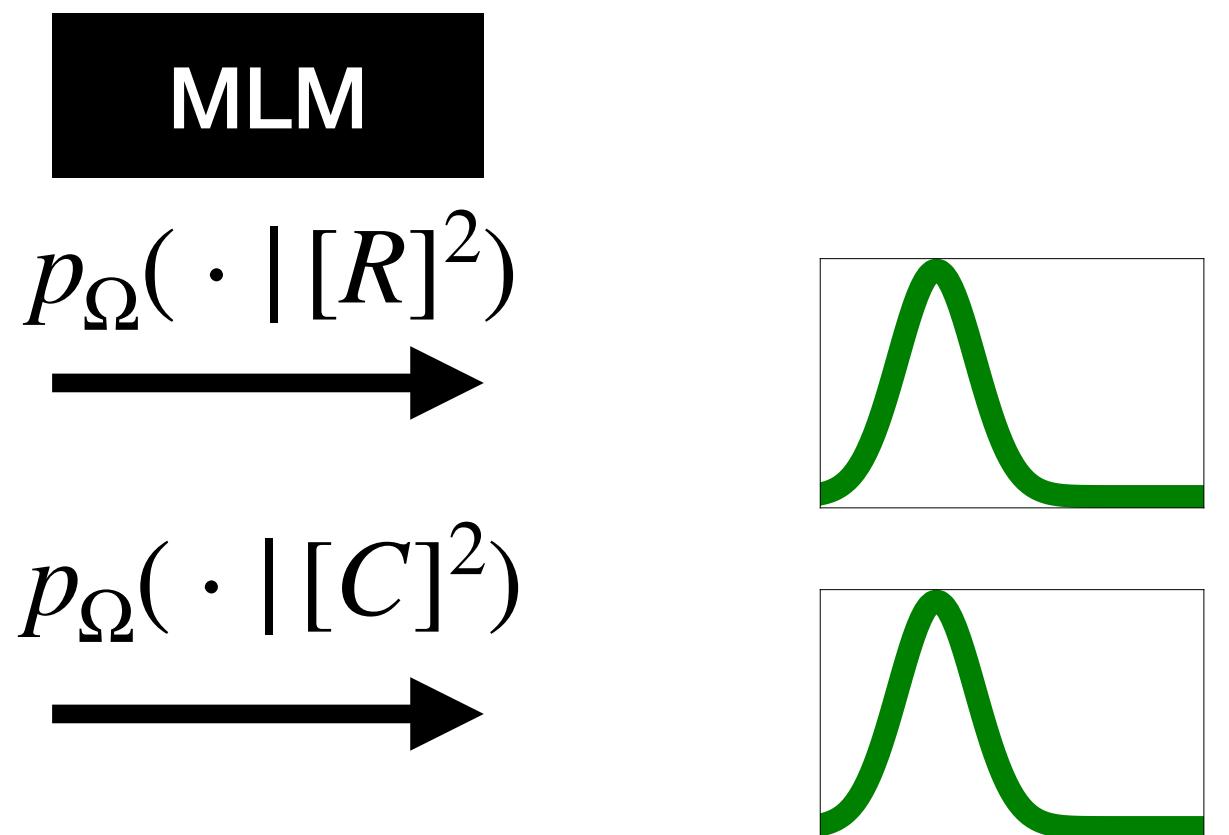
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



$$\mathcal{I}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{J} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

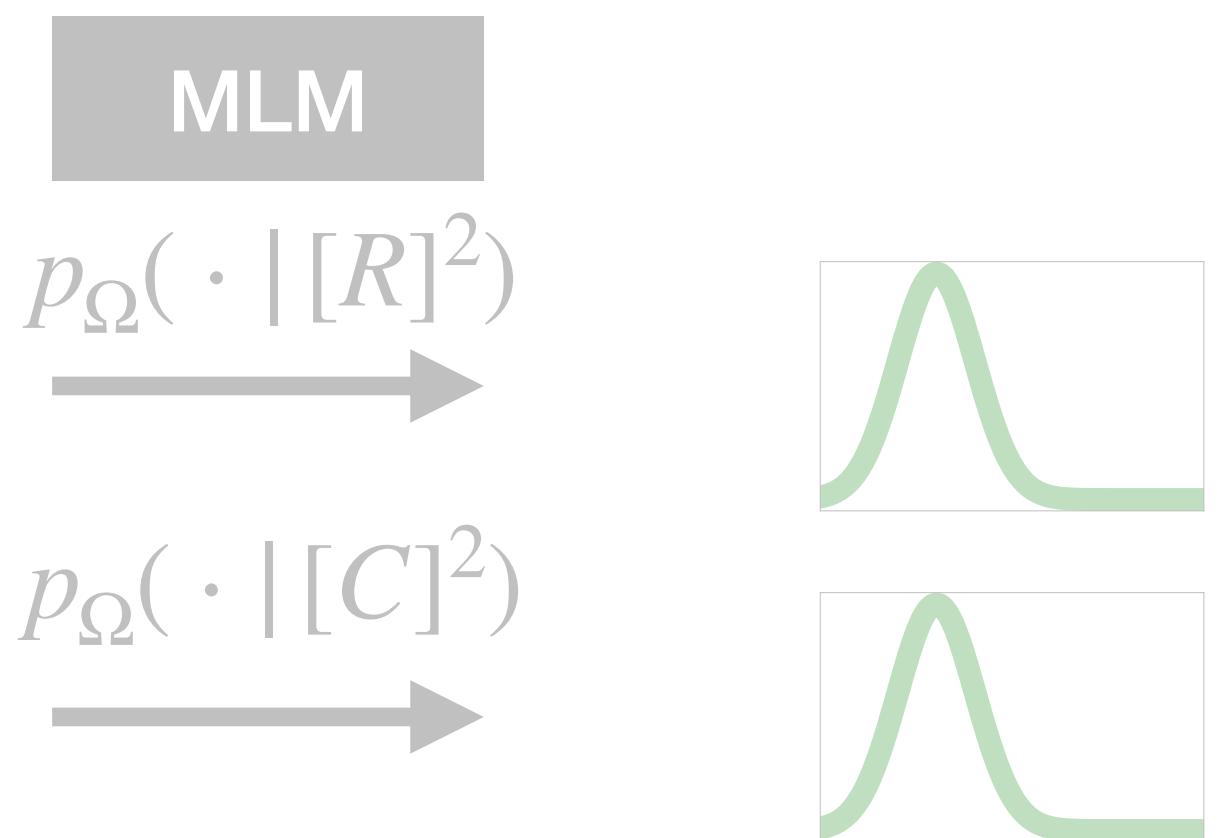
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !



$$\mathcal{J}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !

# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

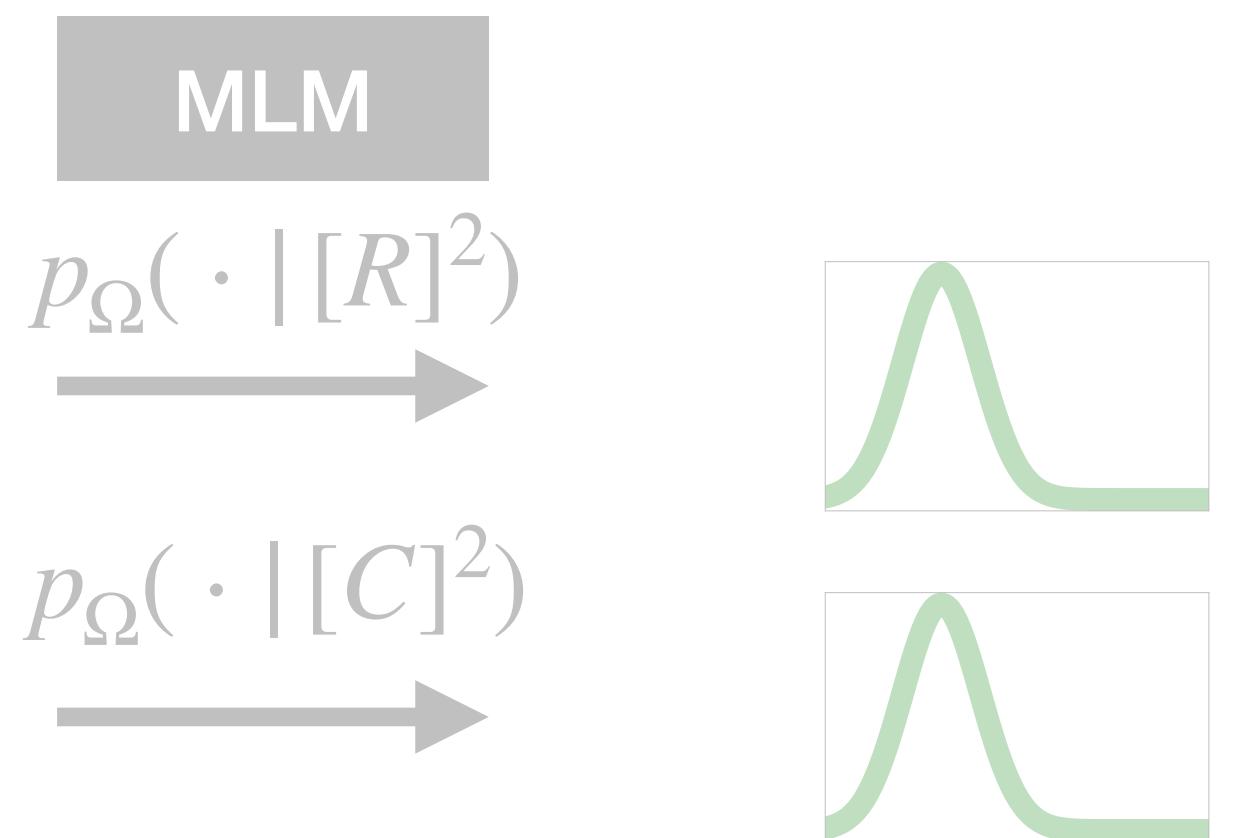
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

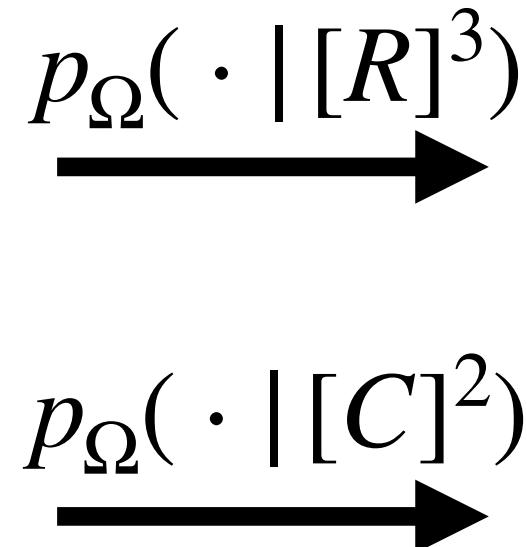


$$\mathcal{I}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

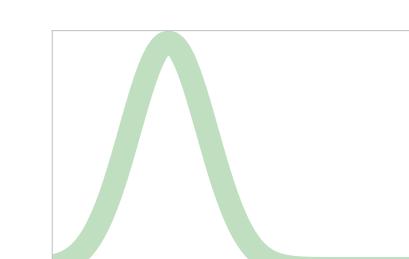
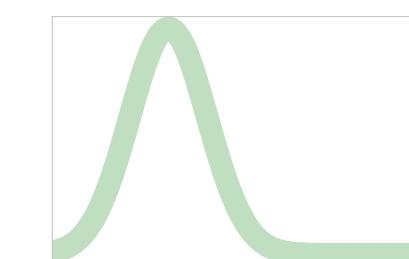
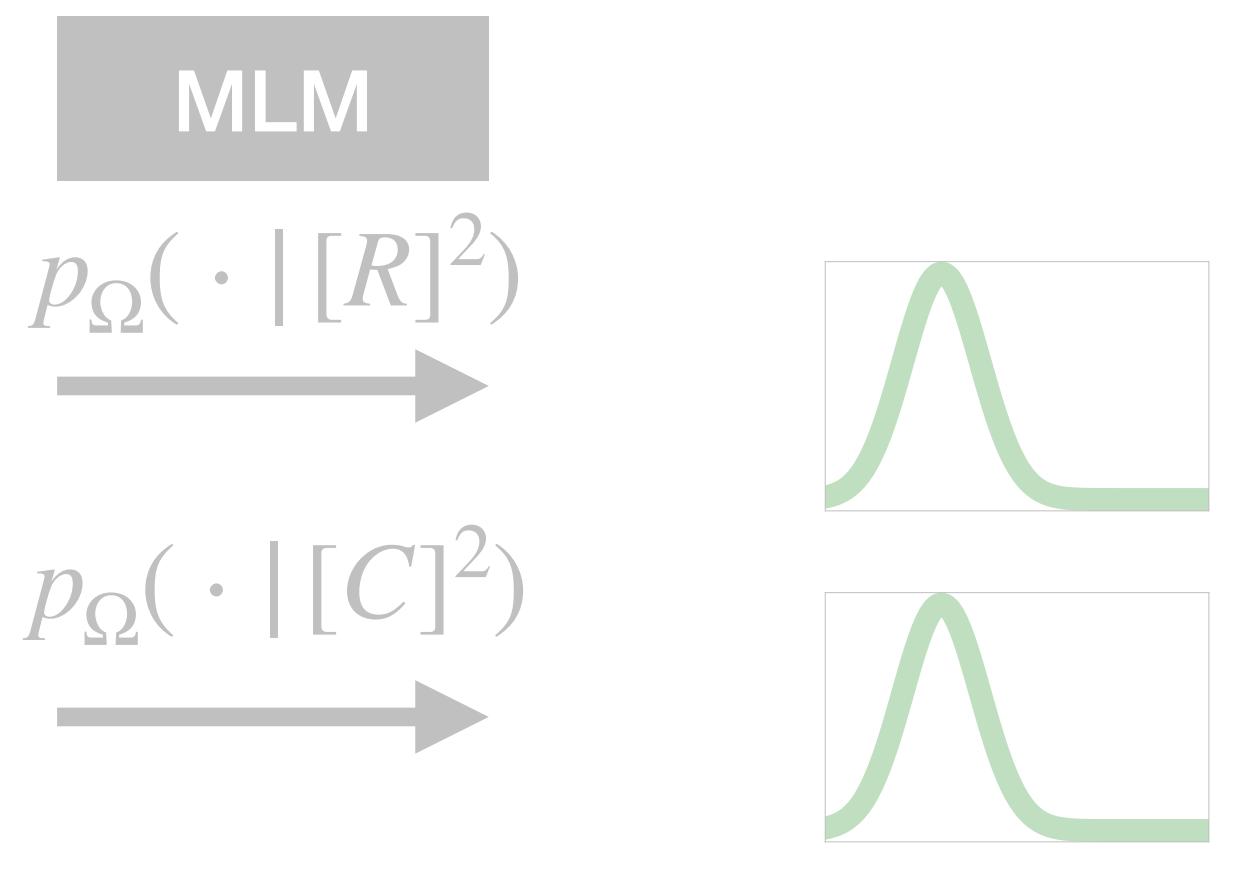
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

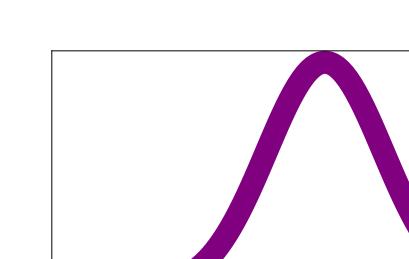
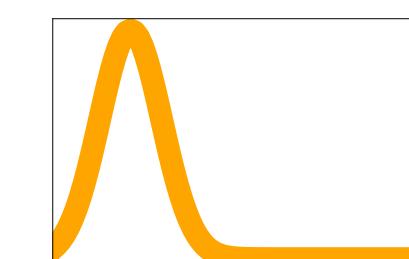
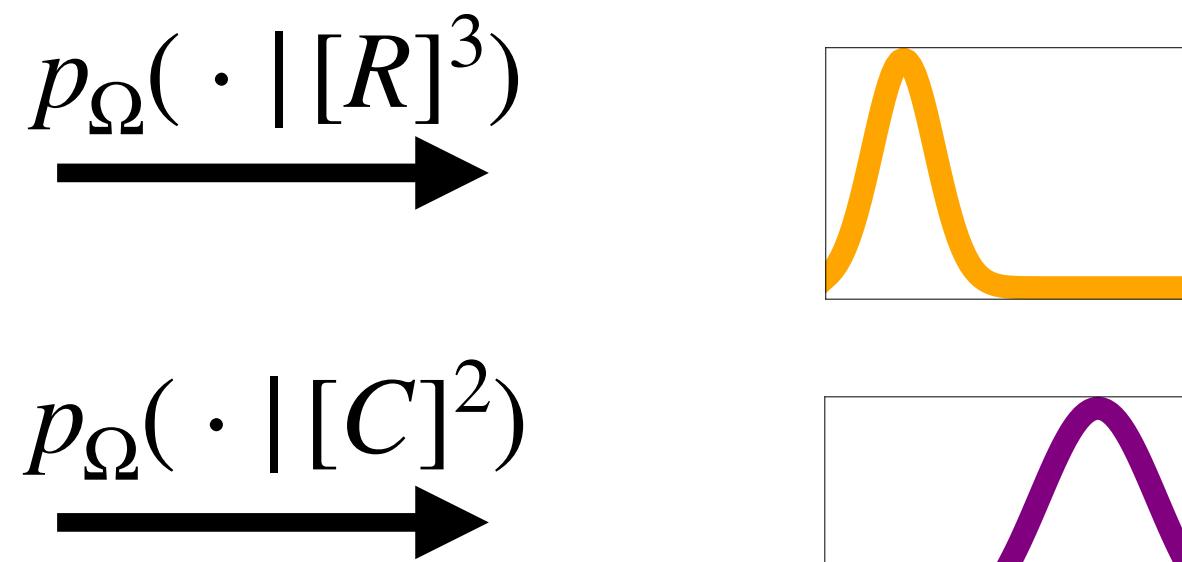


$$\mathcal{I}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !



# Intuition of InfoLM

**Goal** Compute a similarity score between R and C.

Equivalence for masked contexts

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

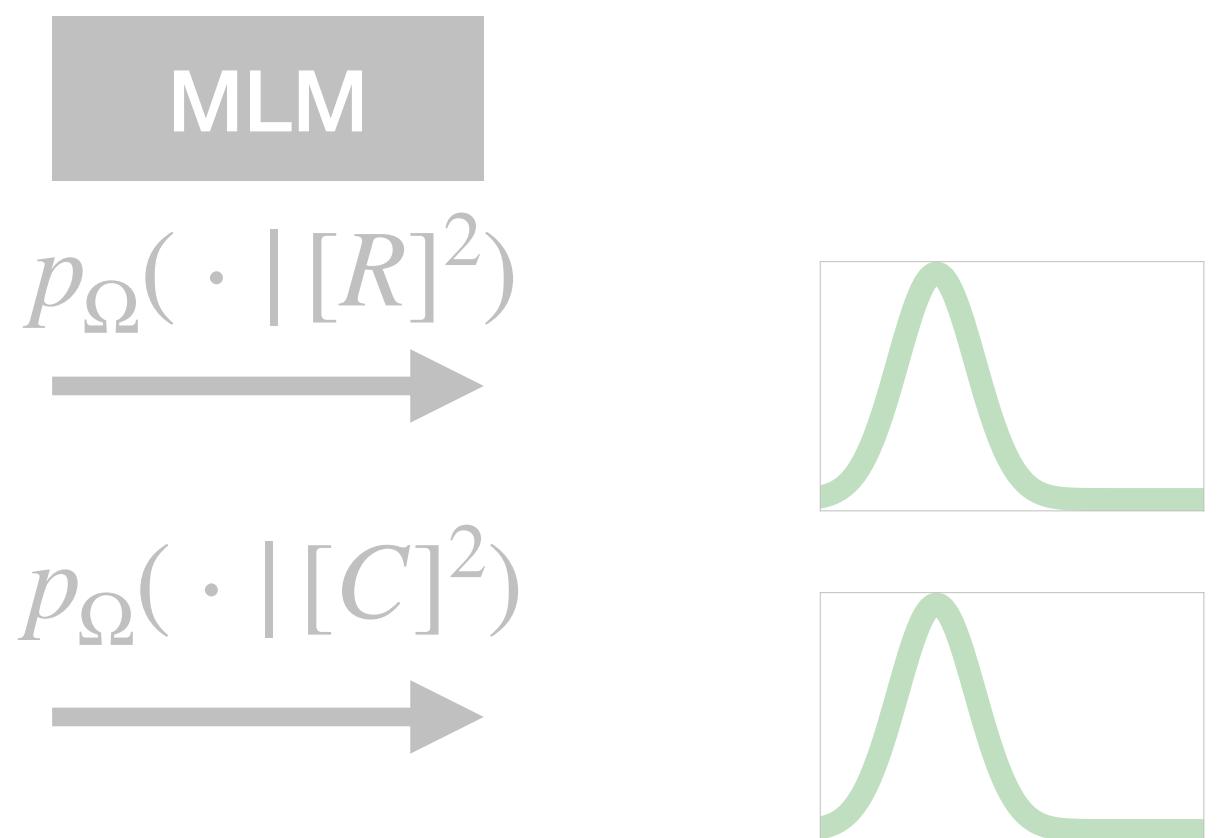
MLM predicts a distribution over  $\Omega$

$$p_{\Omega}(\cdot | [R]^i)$$

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

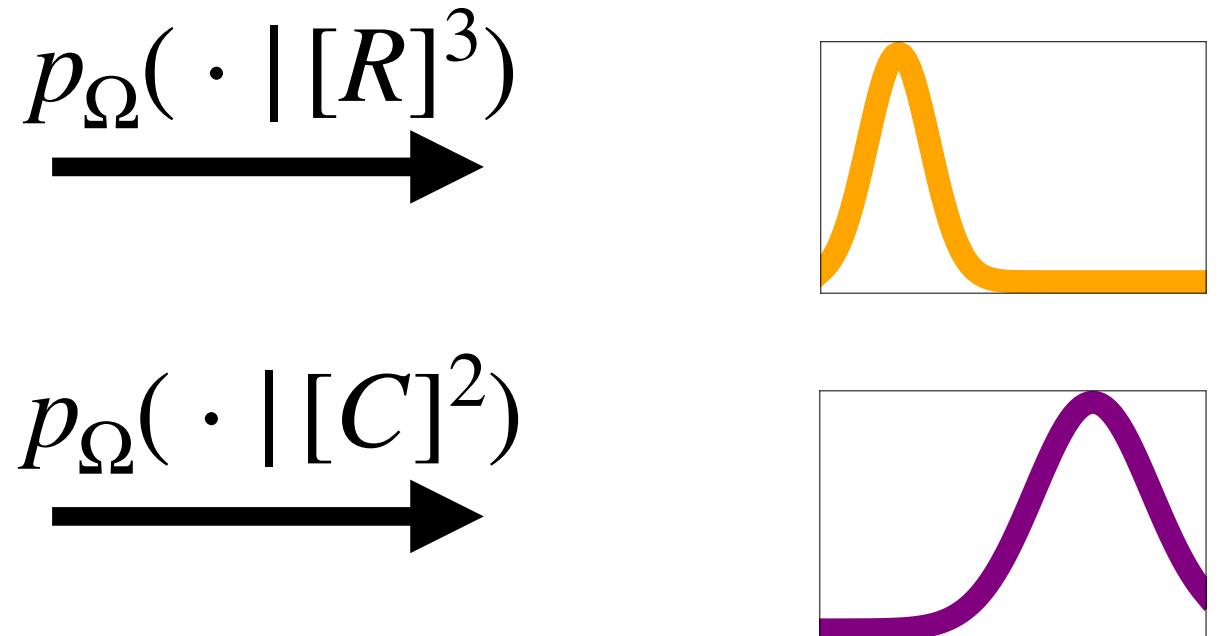


$$\mathcal{I}(p_{\Omega}(\cdot | [R]^2), p_{\Omega}(\cdot | [C]^2)) \sim 0$$

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !



$$\mathcal{I}(p_{\Omega}(\cdot | [R]^3), p_{\Omega}(\cdot | [C]^2)) \gg 0$$

# Evaluating Metric Scores

---

# Evaluating Metric Scores

## Notations

S systems  
N texts

$R_i$   
 $C_i^j$

i-th reference

i-th text candidate  
generated by j-th  
system

$h(C_i^j)$

human score

# Evaluating Metric Scores

## Notations

S systems  
N texts

$R_i$

i-th reference

$C_i^j$

i-th text candidate  
generated by j-th  
system

$h(C_i^j)$

human score

Can the metric be used to compare  
the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^S) \right]$$

$$H^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N h(C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N h(C_i^S) \right]$$

# Evaluating Metric Scores

## Notations

S systems  
N texts

$R_i$   
 $C_i^j$

i-th reference  
i-th text candidate  
generated by j-th  
system

$h(C_i^j)$   
human score

Can the metric be used to compare  
the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^S) \right]$$

$$H^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N h(C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N h(C_i^S) \right]$$

**System Aggregation**  
**Compare vector of length S**

# Evaluating Metric Scores

## Notations

S systems  
N texts

$R_i$

i-th reference

$C_i^j$

i-th text candidate  
generated by j-th  
system

$h(C_i^j)$

human score

Can the metric be used to compare  
the performance of two systems?

Can the metric be used as a loss or reward  
of a system?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^S) \right]$$

$$H^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N h(C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N h(C_i^S) \right]$$

$$K_{text} = \frac{1}{N} \sum_{i=1}^N K(M_i^{text}, H_i^{text})$$

$$H_i^{text} = [h(C_i^1), \dots, h(C_i^S)]$$

$$M_i^{text} = [m(R_i, C_i^1), \dots, m(R_i, C_i^S)]$$

System Aggregation  
Compare vector of length S

# Evaluating Metric Scores

## Notations

S systems  
N texts

$R_i$

i-th reference

$C_i^j$

i-th text candidate  
generated by j-th  
system

$h(C_i^j)$

human score

Can the metric be used to compare  
the performance of two systems?

Can the metric be used as a loss or reward  
of a system?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N m(R_i, C_i^S) \right]$$

$$H^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N h(C_i^1), \dots, \frac{1}{N} \sum_{i=1}^N h(C_i^S) \right]$$

$$K_{text} = \frac{1}{N} \sum_{i=1}^N K(M_i^{text}, H_i^{text})$$

$$H_i^{text} = [h(C_i^1), \dots, h(C_i^S)]$$

$$M_i^{text} = [m(R_i, C_i^1), \dots, m(R_i, C_i^S)]$$

**System Aggregation**  
**Compare vector of length S**

**Text Aggregation**  
**Averaged correlation**

# Experimental Setting

---

## Experimental Setting

---

### Data2text Generation

- Results on **WebNLG 2020**

Gardent et al. 2017

- **Correctness / Data Coverage / Relevance**  
**Fluency / Text Structure**

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

- Results on English only

## Experimental Setting

---

### Data2text Generation

- Results on **WebNLG 2020**

Gardent et al. 2017

- Correctness / Data Coverage / Relevance  
Fluency / Text Structure

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

- Results on English only

### Summary Generation

- Results on **SummEval**

Nallapati et al. 2016)

Bhandari et al. (2020)

- Correlation with **pyramid score**

Nenkova and Passonneau 2004

- Results on English only

# Experimental Setting

## Data2text Generation

- Results on **WebNLG 2020**

Gardent et al. 2017

- **Correctness / Data Coverage / Relevance**  
**Fluency / Text Structure**

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

- Results on English only

## Summary Generation

- Results on **SummEval**

Nallapati et al. 2016)

Bhandari et al. (2020)

- Correlation with **pyramid score**

Nenkova and Passonneau 2004

- Results on English only

# Results

---

# Results

---

## Task

(John\\_Blaha birthDate 1942\\_08\\_26)  
(John\\_Blaha birthPlace San\\_Antonio)  
(John\\_E\\_Blaha job Pilot)



John Blaha, born in San  
Antonio on 1942-08-26,  
worked as a pilot

# Results

---

## Task

(John\\_Blaha birthDate 1942\\_08\\_26)  
 (John\\_Blaha birthPlace San\\_Antonio)  
 (John\\_E\\_Blaha job Pilot)



John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	51.7	38.3	100.0	100.0	100.0	77.0	61.4	46.6	99.5	99.7	98.3
Relev	89.7	99.1	75.0	96.0	93.8	81.6	77.0	61.4	46.6	100.0	100.0	100.0	77.2	61.1	45.0
TextS	80.8	80.1	60.0	71.6	51.4	36.6	99.5	99.7	98.3	77.2	61.1	45.0	100.0	100.0	100.0
$\mathcal{D}_{AB}$	88.8	<u>89.3</u>	<u>76.6</u>	<u>81.8</u>	<u>82.6</u>	<u>70.0</u>	86.6	92.0	76.6	<u>89.8</u>	<u>87.9</u>	<u>73.3</u>	86.6	91.4	75.0
$\mathcal{D}_\alpha$	88.8	<u>89.3</u>	<u>76.6</u>	<u>81.8</u>	<u>82.6</u>	<u>70.0</u>	86.6	92.0	76.6	<u>89.8</u>	<u>87.9</u>	<u>73.3</u>	86.6	91.4	75.0
$\mathcal{D}_\beta$	81.4	50.0	71.6	48.4	79.7	65.0	44.8	84.7	76.6	49.3	72.3	60.0	48.0	83.8	75.0
$\mathcal{L}_1$	75.2	33.8	61.6	32.4	53.8	40.0	22.7	83.5	73.3	32.2	57.9	45.0	25.6	83.2	71.6
$\mathcal{R}$	<u>89.7</u>	86.0	75.0	78.7	70.5	51.6	<u>93.3</u>	<u>95.7</u>	<u>85.3</u>	87.6	84.4	70.0	<u>92.4</u>	<u>93.8</u>	<u>81.6</u>
JS	79.4	81.1	70.0	69.3	75.5	60.0	89.4	91.4	75.0	81.7	70.5	60.0	91.9	91.1	73.3
BertS	<u>85.5</u>	83.4	<u>73.3</u>	74.7	<u>68.2</u>	53.3	<u>92.3</u>	<u>95.5</u>	<u>85.0</u>	<u>83.3</u>	<u>79.4</u>	<u>65.0</u>	91.9	<u>95.0</u>	<u>83.3</u>
MoverS	84.1	<u>84.1</u>	<u>73.3</u>	<u>78.7</u>	66.2	<u>53.3</u>	91.2	92.1	78.3	82.1	77.4	65.0	90.1	91.4	76.3
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	<u>89.4</u>	90.5	78.3	63.0	65.2	51.6	88.5	89.1	76.6
R-1	80.6	65.0	65.0	61.1	<u>59.6</u>	48.3	76.5	76.3	60.3	64.3	<u>69.2</u>	56.7	75.9	77.5	58.3
METEOR	<u>86.5</u>	<u>66.3</u>	<u>70.0</u>	<u>77.3</u>	50.2	46.6	86.7	90.5	78.3	<u>82.1</u>	<u>65.2</u>	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	<u>93.5</u>	<u>80.0</u>	75.0	70.2	<u>77.6</u>	89.5	91.1	78.6

# Results

Task

(John\\_Blaha birthDate 1942\\_08\\_26)  
 (John\\_Blaha birthPlace San\\_Antonio)  
 (John\\_E\\_Blaha job Pilot)

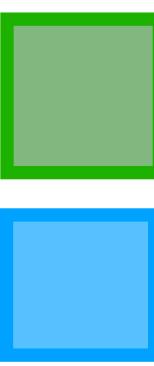


John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$	r	$\rho$	$\tau$
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	51.7	38.3	100.0	100.0	100.0	77.0	61.4	46.6	99.5	99.7	98.3
Relev	89.7	99.1	75.0	96.0	93.8	81.6	77.0	61.4	46.6	100.0	100.0	100.0	77.2	61.1	45.0
TextS	80.8	80.1	60.0	71.6	51.4	36.6	99.5	99.7	98.3	77.2	61.1	45.0	100.0	100.0	100.0
$\mathcal{D}_{AB}$	88.8	<u>89.3</u>	<u>76.6</u>	<u>81.8</u>	<u>82.6</u>	<u>70.0</u>	86.6	92.0	76.6	<u>89.8</u>	<u>87.9</u>	<u>73.3</u>	86.6	91.4	75.0
$\mathcal{D}_\alpha$	88.8	<u>89.3</u>	<u>76.6</u>	<u>81.8</u>	<u>82.6</u>	<u>70.0</u>	86.6	92.0	76.6	<u>89.8</u>	<u>87.9</u>	<u>73.3</u>	86.6	91.4	75.0
$\mathcal{D}_\beta$	81.4	50.0	71.6	48.4	79.7	65.0	44.8	84.7	76.6	49.3	72.3	60.0	48.0	83.8	75.0
$\mathcal{L}_1$	75.2	33.8	61.6	32.4	53.8	40.0	22.7	83.5	73.3	32.2	57.9	45.0	25.6	83.2	71.6
$\mathcal{R}$	<u>89.7</u>	86.0	75.0	78.7	70.5	51.6	<u>93.3</u>	<u>95.7</u>	<u>85.3</u>	87.6	84.4	70.0	<u>92.4</u>	<u>93.8</u>	<u>81.6</u>
JS	79.4	81.1	70.0	69.3	75.5	60.0	89.4	91.4	75.0	81.7	70.5	60.0	91.9	91.1	73.3
BertS	<u>85.5</u>	83.4	<u>73.3</u>	74.7	<u>68.2</u>	53.3	<u>92.3</u>	<u>95.5</u>	<u>85.0</u>	<u>83.3</u>	<u>79.4</u>	<u>65.0</u>	<u>91.9</u>	<u>95.0</u>	<u>83.3</u>
MoverS	<u>84.1</u>	<u>84.1</u>	<u>73.3</u>	<u>78.7</u>	66.2	<u>53.3</u>	91.2	92.1	78.3	82.1	77.4	65.0	90.1	91.4	76.3
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	<u>89.4</u>	90.5	78.3	63.0	65.2	51.6	88.5	89.1	76.6
R-1	80.6	65.0	65.0	61.1	<u>59.6</u>	48.3	76.5	76.3	60.3	64.3	<u>69.2</u>	56.7	75.9	77.5	58.3
METEOR	<u>86.5</u>	<u>66.3</u>	<u>70.0</u>	<u>77.3</u>	50.2	46.6	86.7	90.5	78.3	<u>82.1</u>	<u>65.2</u>	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	<u>93.5</u>	80.0	75.0	70.2	<u>77.6</u>	89.5	91.1	78.6

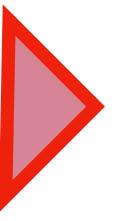
# Multimodal sentiment analysis

---



Fusion Block

Embedding Block

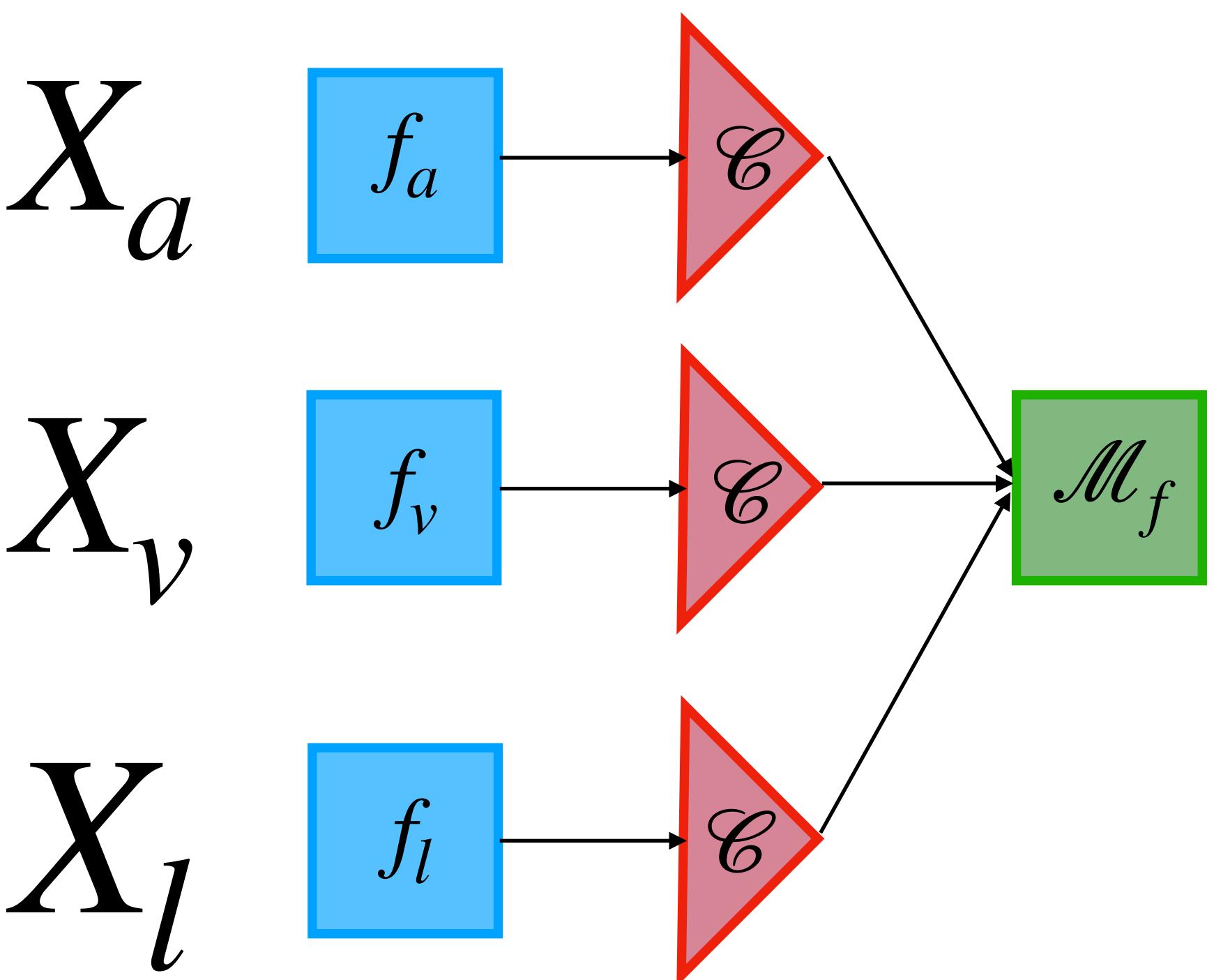


Predictor block

# Multimodal sentiment analysis



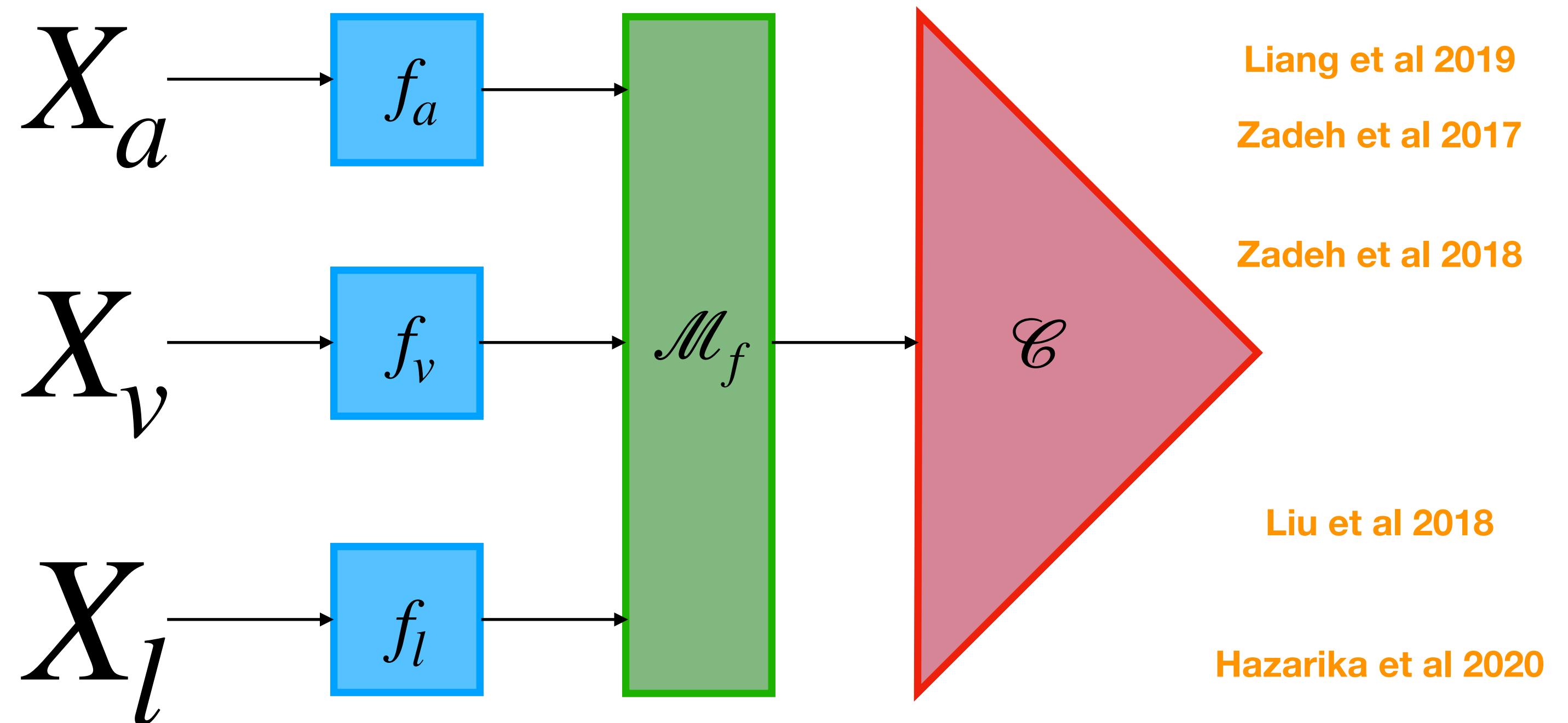
## Late Fusion



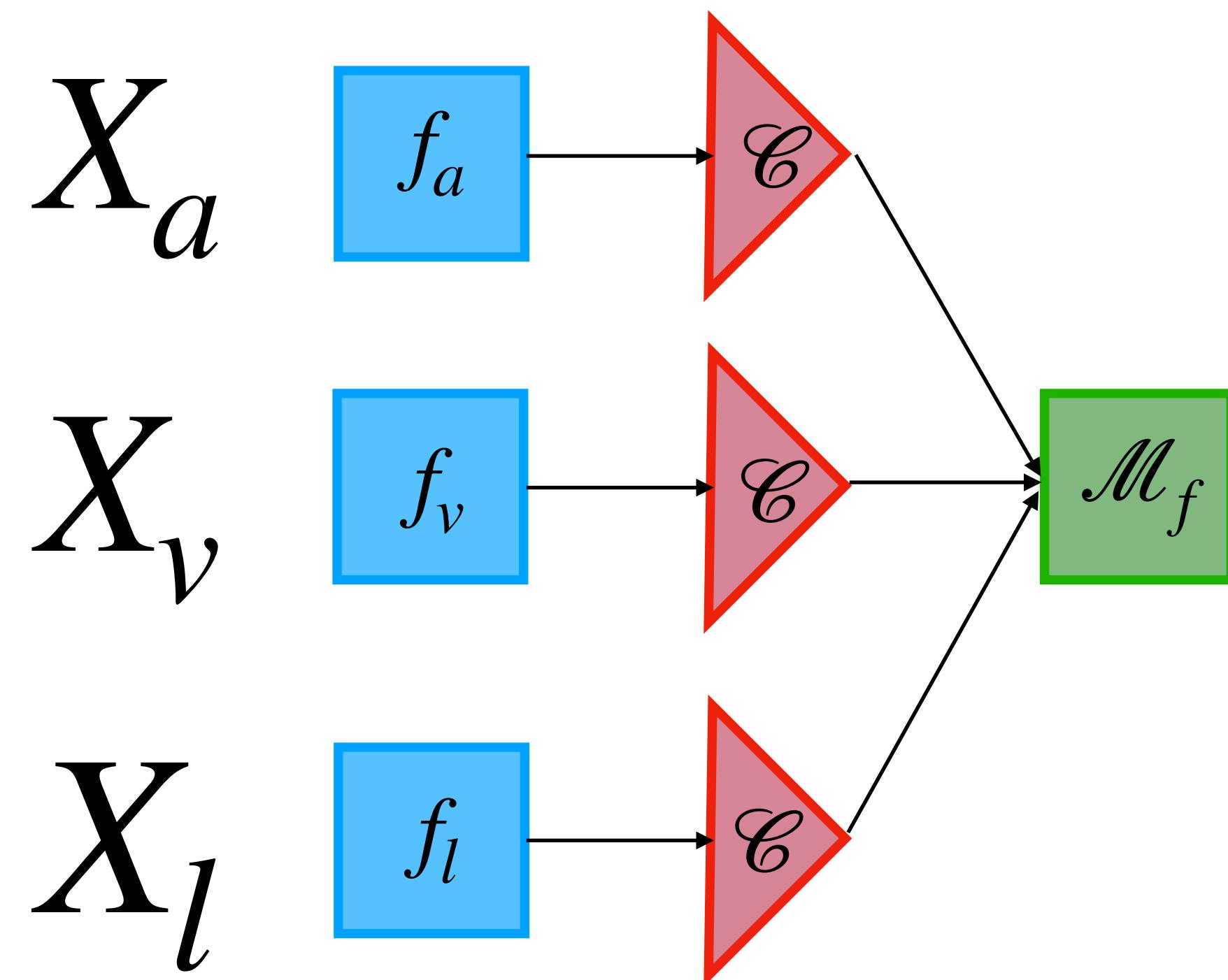
# Multimodal sentiment analysis



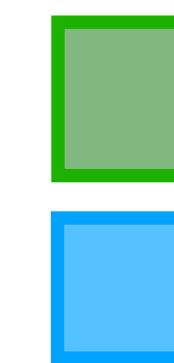
## Early Fusion



## Late Fusion

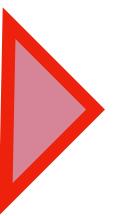


# Multimodal sentiment analysis



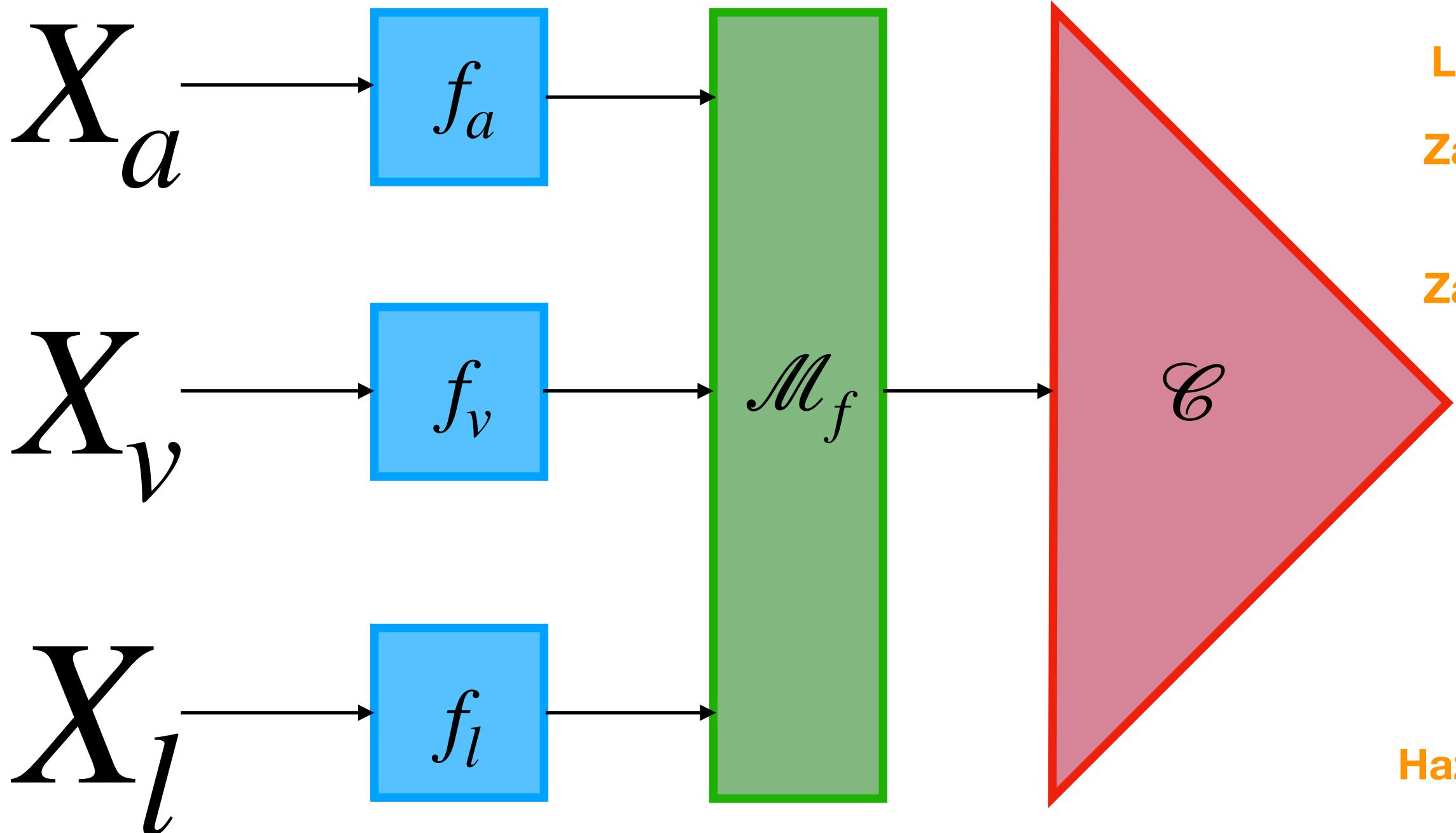
Fusion Block

Embedding Block

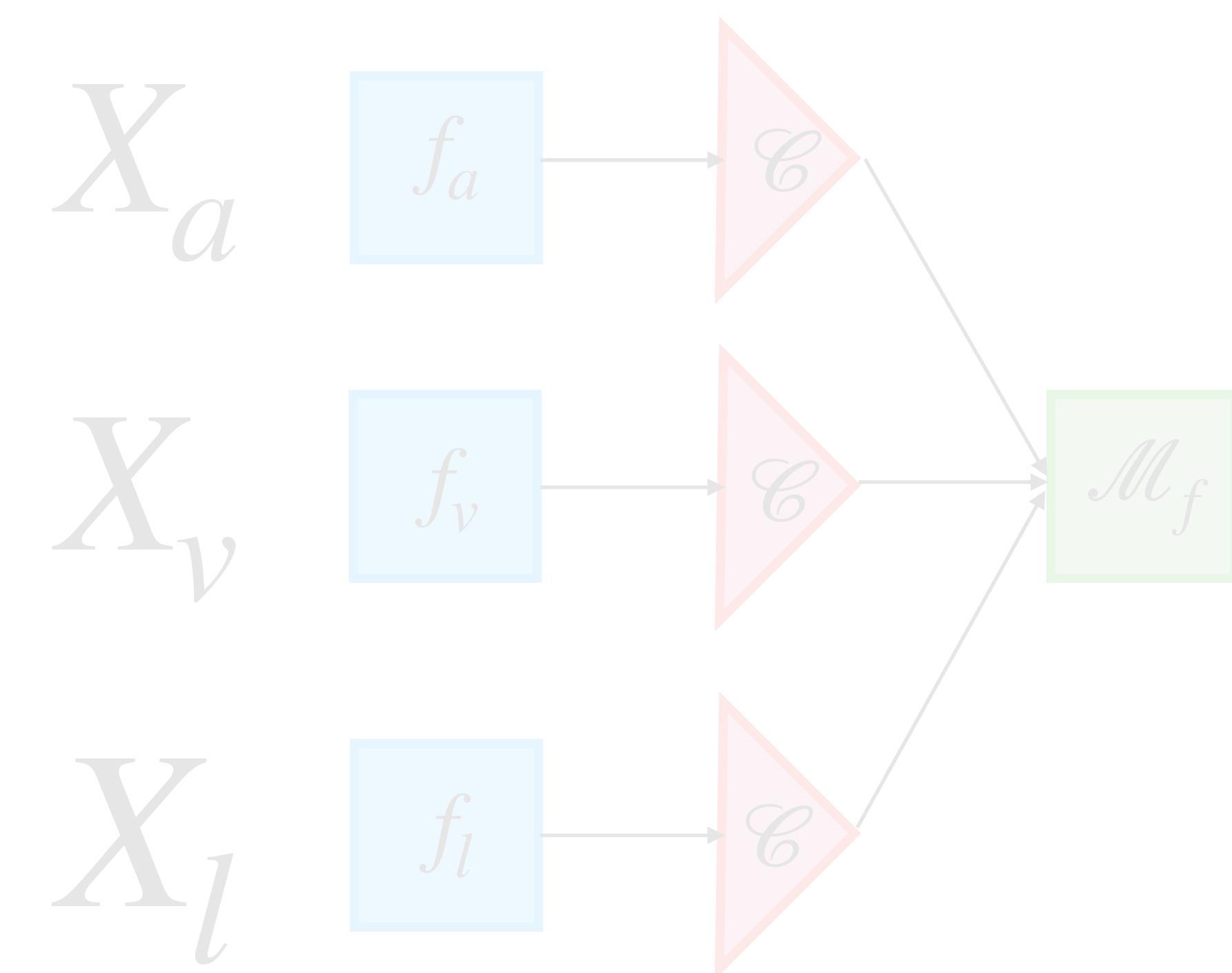


Predictor block

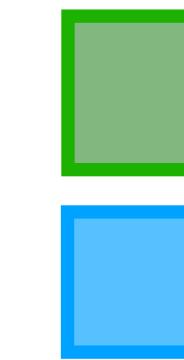
## Early Fusion



## Late Fusion



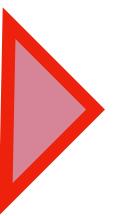
# Multimodal sentiment analysis



Fusion Block

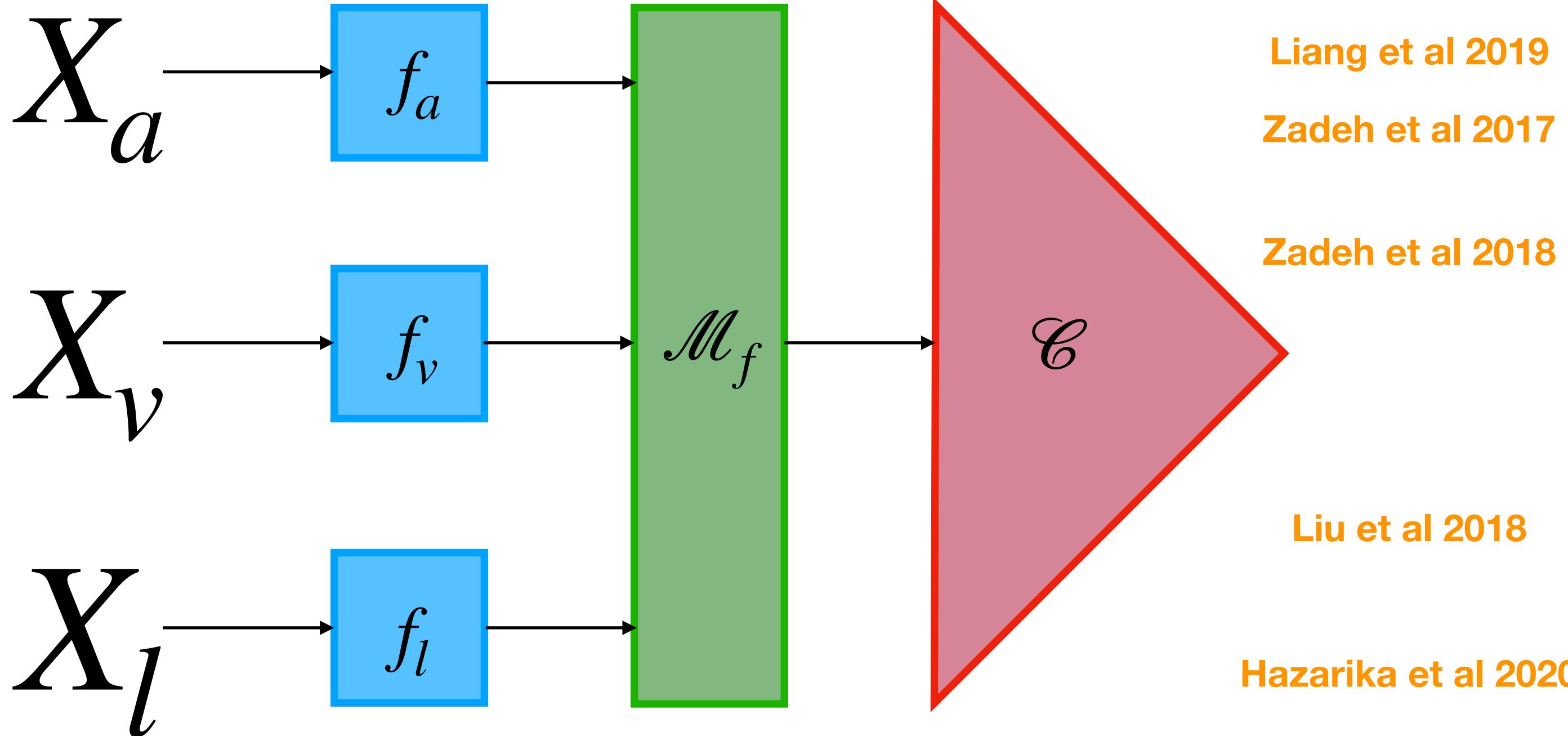


Embedding Block

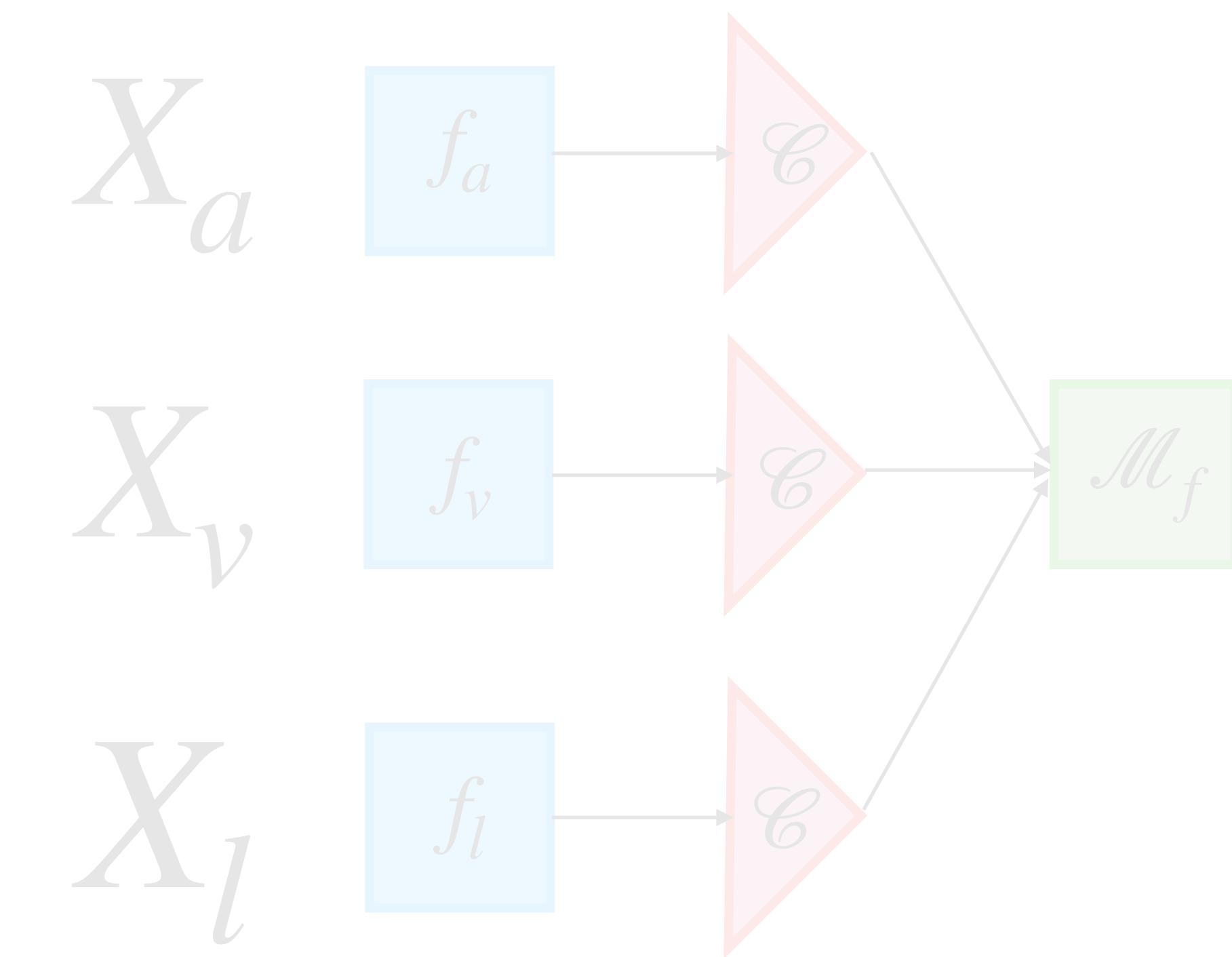


Predictor block

## Early Fusion



## Late Fusion



Fusion in previous work mainly rely on complex neural networks  
Few previous works on improving fusion using loss function!

## Limitation of InfoMax

---

## Limitation of InfoMax

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

## Limitation of InfoMax

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

Middle/High size labelled corpora

**Switchboard Dialog Act (100k + utterances)** Godfrey et al., 1992

**MRDA (110k + utterances)** Shriberg et al., 2004

**Dialy Dialog Act (100k + utterances)** Li et al., 2017

## Limitation of InfoMax

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

Middle/High size labelled corpora

**Switchboard Dialog Act (100k + utterances)** Godfrey et al., 1992

**MRDA (110k + utterances)** Shriberg et al., 2004

**Dialy Dialog Act (100k + utterances)** Li et al., 2017

Not practical but deep learning is data hungry!

We will rely on pretraining

# Estimation of MI

---

## Estimation of MI

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

## Estimation of MI

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

Middle/High size labelled corpora

**Switchboard Dialog Act (100k + utterances)** Godfrey et al., 1992

**MRDA (110k + utterances)** Shriberg et al., 2004

**Dialy Dialog Act (100k + utterances)** Li et al., 2017

## Estimation of MI

---

Current state-of-the art considers deep learning models

Contextual recurrent encoder + Multi Layer Perceptron Bothe et al. 2018

Contextual recurrent encoder + CRF/Recurrent decoder Chen et al., 2018

Middle/High size labelled corpora

Switchboard Dialog Act (100k + utterances) Godfrey et al., 1992

MRDA (110k + utterances) Shriberg et al., 2004

Dialy Dialog Act (100k + utterances) Li et al., 2017

Not practical but deep learning is data hungry!

We will rely on pretraining

## Goal

**Compute a similarity score  
between R and C.**

# InfoLM

---

Goal

Compute a similarity score  
between R and C.

R: It is cold today.

C: It is freezing today

---

## Algorithm 1: InfoLM

---

```
1: INPUT Candidate text  $\mathbf{y}_i^s$  of length L, Reference text  $\mathbf{x}_i$   
   of length M, measure of information  $\mathcal{I}$   
2:  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}) = 0, 0$   
3: for  $k \in [1, L]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s)$   
4:    $p_{\Omega|T}(\cdot|\mathbf{y}_i^s) = p_{\Omega|T}(\cdot|\mathbf{y}_i^s) + \gamma_k \times p_{\Omega|T}(\cdot|[\mathbf{y}_i^s]^k)$   
5: end for  
6: for  $j \in [1, M]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{x}_i)$   
7:    $p_{\Omega|T}(\cdot|\mathbf{x}_i) = p_{\Omega|T}(\cdot|\mathbf{x}_i) + \bar{\gamma}_k \times p_{\Omega|T}(\cdot|[\mathbf{x}_i]^j)$   
8: end for  
9: OUTPUT  $\mathcal{I}[p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}_i)]$ 
```

---

InfoLM

# InfoLM

Goal

Compute a similarity score between R and C.

R: It is cold today.

C: It is freezing today

## Algorithm 1: InfoLM

```
1: INPUT Candidate text  $\mathbf{y}_i^s$  of length L, Reference text  $\mathbf{x}_i$  of length M, measure of information  $\mathcal{I}$ 
2:  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}) = 0, 0$ 
3: for  $k \in [1, L]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s)$ 
4:    $p_{\Omega|T}(\cdot|\mathbf{y}_i^s) = p_{\Omega|T}(\cdot|\mathbf{y}_i^s) + \boxed{\gamma_k} \times p_{\Omega|T}(\cdot|[\mathbf{y}_i^s]^k)$ 
5: end for
6: for  $j \in [1, M]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{x}_i)$ 
7:    $p_{\Omega|T}(\cdot|\mathbf{x}_i) = p_{\Omega|T}(\cdot|\mathbf{x}_i) + \boxed{\bar{\gamma}_k} \times p_{\Omega|T}(\cdot|[\mathbf{x}_i]^j)$ 
8: end for
9: OUTPUT  $\mathcal{I}[p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}_i)]$ 
```

Masked Language Model

InfoLM

Aggregated context R

Aggregated context C

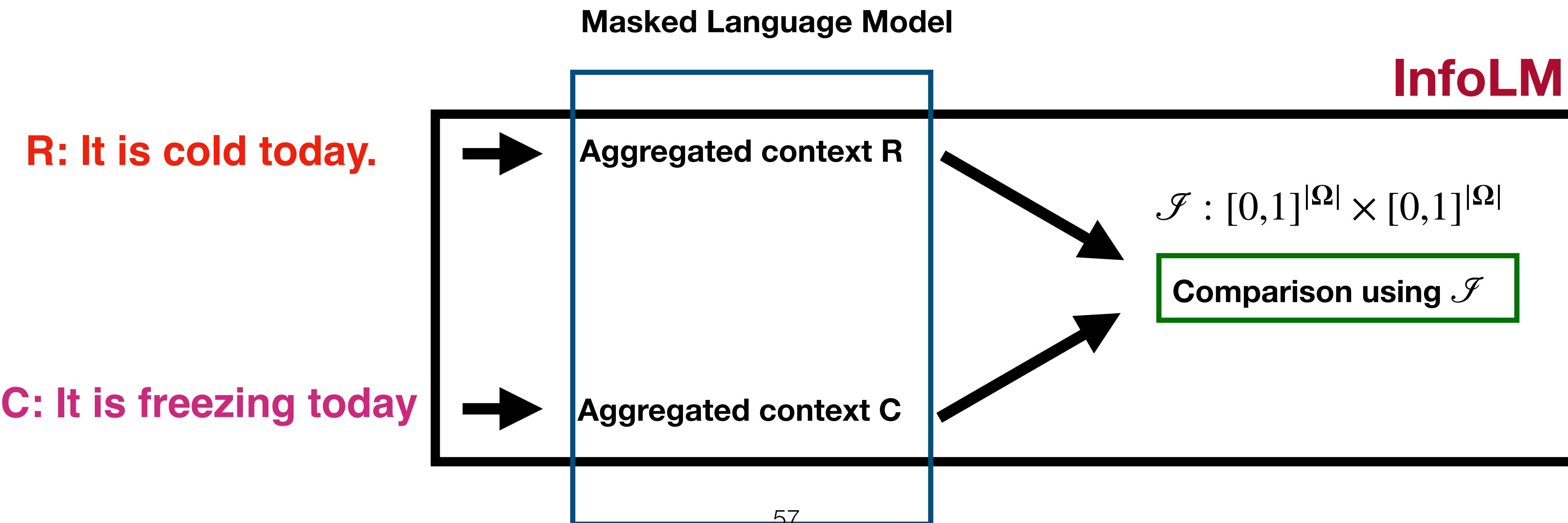
# InfoLM

Goal

Compute a similarity score between R and C.

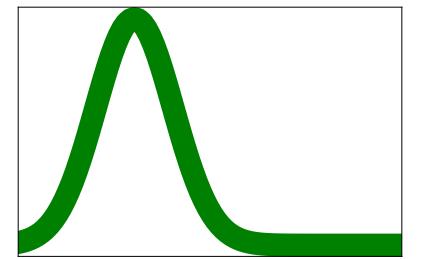
## Algorithm 1: InfoLM

```
1: INPUT Candidate text  $\mathbf{y}_i^s$  of length L, Reference text  $\mathbf{x}_i$  of length M, measure of information  $\mathcal{I}$ 
2:  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}) = 0, 0$ 
3: for  $k \in [1, L]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{y}_i^s)$ 
4:    $p_{\Omega|T}(\cdot|\mathbf{y}_i^s) = p_{\Omega|T}(\cdot|\mathbf{y}_i^s) + \boxed{\gamma_k} \times p_{\Omega|T}(\cdot|[\mathbf{y}_i^s]^k)$ 
5: end for
6: for  $j \in [1, M]$  do ▷ Compute  $p_{\Omega|T}(\cdot|\mathbf{x}_i)$ 
7:    $p_{\Omega|T}(\cdot|\mathbf{x}_i) = p_{\Omega|T}(\cdot|\mathbf{x}_i) + \boxed{\bar{\gamma}_k} \times p_{\Omega|T}(\cdot|[\mathbf{x}_i]^j)$ 
8: end for
9: OUTPUT  $\mathcal{I}[p_{\Omega|T}(\cdot|\mathbf{y}_i^s), p_{\Omega|T}(\cdot|\mathbf{x}_i)]$ 
```



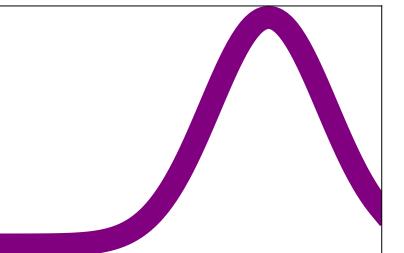
# InfoLM context aggregation

R: [MASK]is cold today.



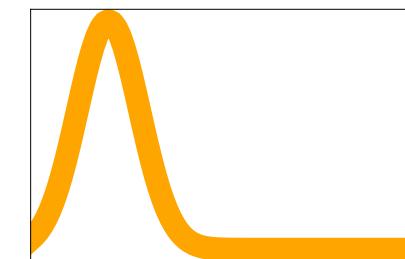
....

R: It is [MASK] today.



....

R: It is cold today [MASK]



# InfoLM context aggregation

**Goal** Compute a similarity score between R and C.

From equivalent context to text similarity

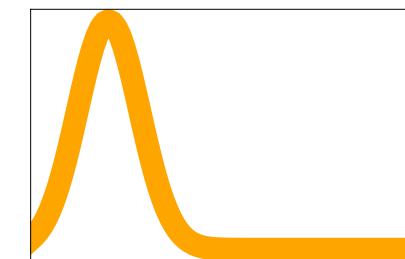
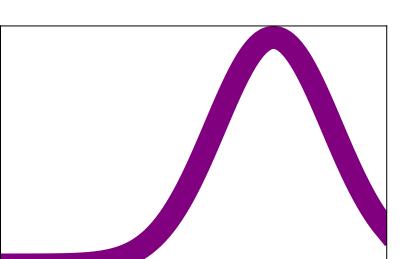
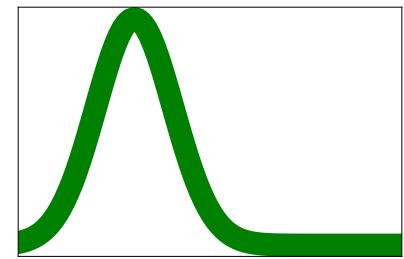
R: [MASK]is cold today.

....

R: It is [MASK] today.

....

R: It is cold today [MASK]



# InfoLM context aggregation

**Goal** Compute a similarity score between R and C.

From equivalent context to text similarity

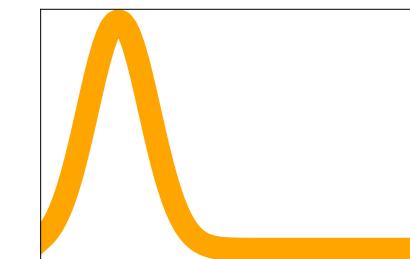
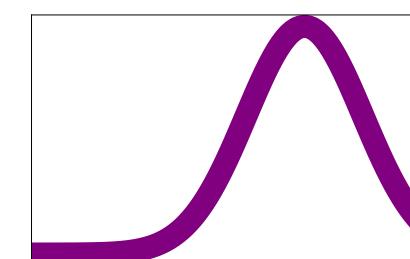
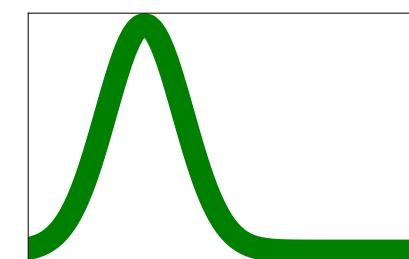
R: [MASK]is cold today.

....

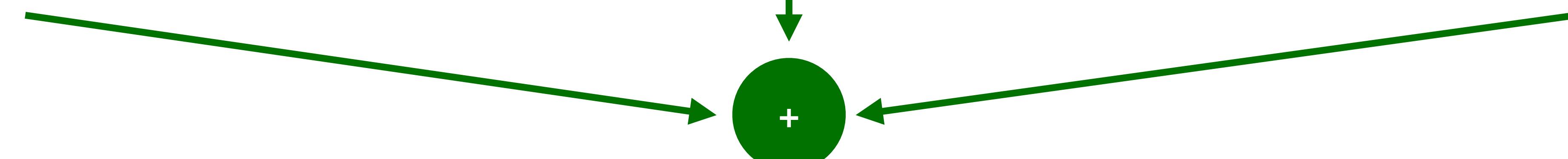
R: It is [MASK] today.

....

R: It is cold today [MASK]



Aggregation



# InfoLM context aggregation

**Goal** Compute a similarity score between R and C.

From equivalent context to text similarity

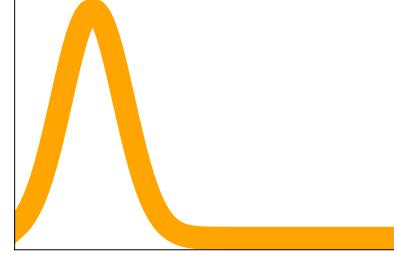
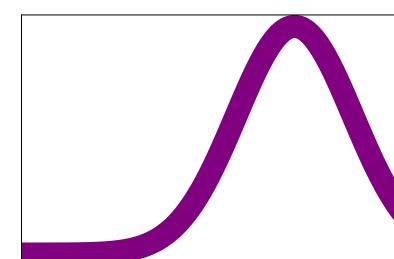
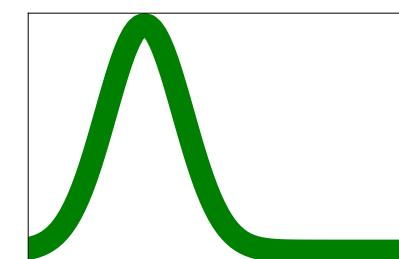
R: [MASK]is cold today.

....

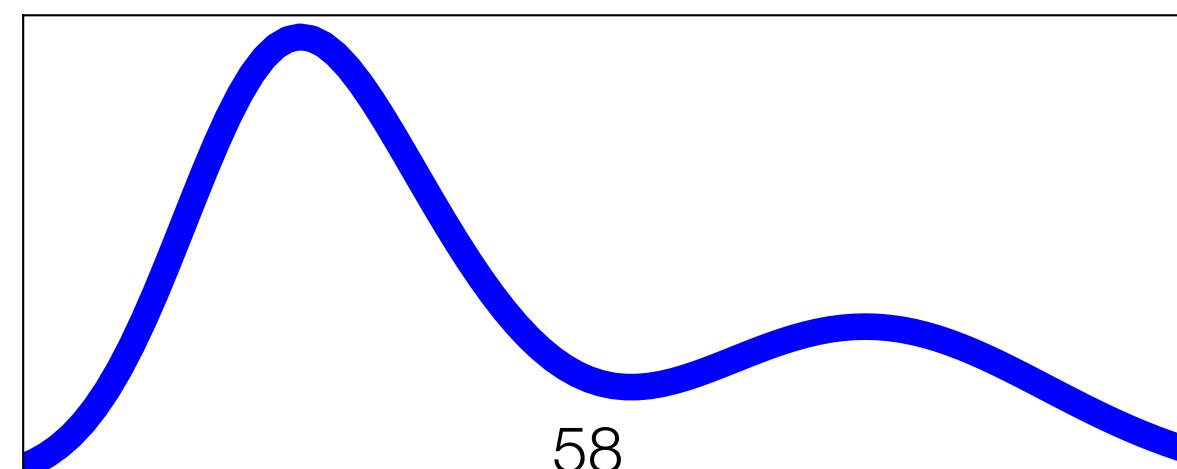
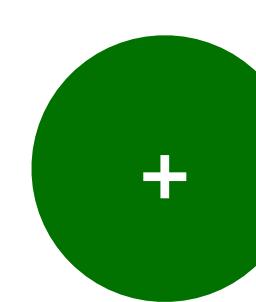
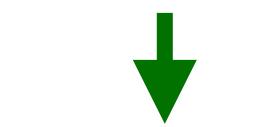
R: It is [MASK] today.

....

R: It is cold today [MASK]



Aggregation



Aggregated context

# Robustness to modality drop

---

# **Robustness to modality drop**

---

## **1. Most of the information carried by Language**

- Drop Language

# Robustness to modality drop

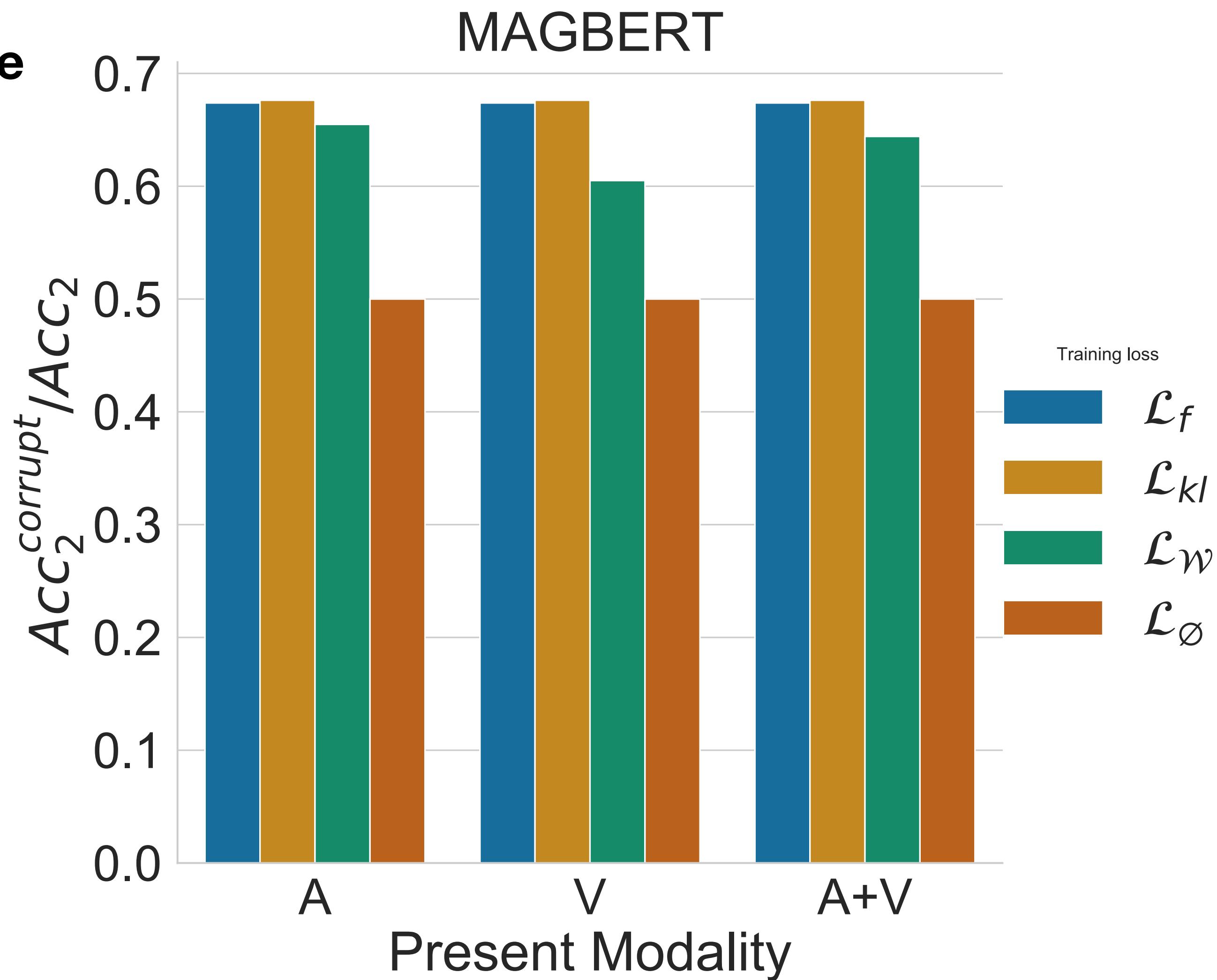
## 1. Most of the information carried by Language

- Drop Language

How maximising the MDM affect robustness?

### Experiment

1. Train using 3 modalities
2. At inference we use only Audio or Video



# Robustness to modality drop

## 1. Most of the information carried by Language

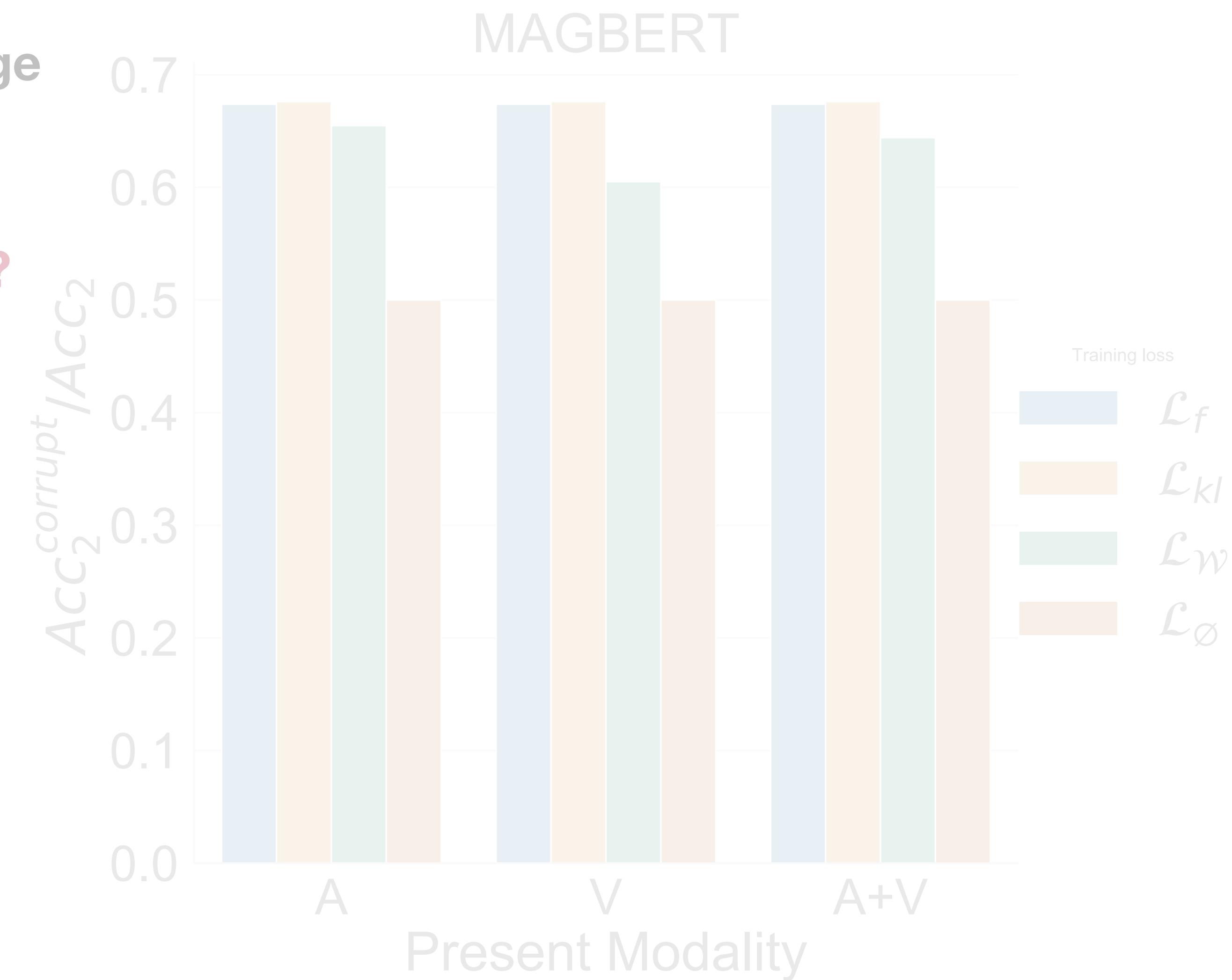
- Drop Language

How maximising the MDM affect robustness?

### Experiment

1. Train using 3 modalities
2. At inference we use only Audio or Video

Our MDM loss improve robustness



# 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 ?**

**I stole. And I robbed.**



**Euhhhh !**

Question



**That is quite the rap sheet, Prison Mike.**



**And I never got caught neither!**



**Well, you are in prison...**

Statement  
opinion

Statement

Appreciation

# Importance of Sequence labelling

# Importance of Sequence labelling

## Why are sequence labelling tasks useful?

**Speaker modelling**

He et al. 2021

**Dialog State Tracking**

Perez et al. 2017

**Avoid generic response problem**

Yi et al. 2019

# Importance of Sequence labelling

## Why are sequence labelling tasks useful?

Speaker modelling

He et al. 2021

Dialog State Tracking

Perez et al. 2017

Avoid generic response problem

Yi et al. 2019



## Generic response problem

What'd you do, Prison Mike ?

I don't know



Did'n't you rob someone?

I don't know



# Importance of Sequence labelling

## Generic response problem

### Why are sequence labelling tasks useful?

Speaker modelling

He et al. 2021



Dialog State Tracking

Perez et al. 2017

Avoid generic response problem

Yi et al. 2019

### Types of label considered:

Dialog Act

DIT++ / DAMSL / DiAML

Bunt et al 2010

Core and Allen, 1997



Emotion and Sentiment

Polarity (+/-)

6 emotions (+neutrals)

Ekman · 1999

What'd you do, Prison Mike ?

I don't know



Didnt' you rob someone?

I don't know



## **Limitation of current systems**

---

## Limitation of current systems

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

## Limitation of current systems

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

Middle/High size labelled corpora

**Switchboard Dialog Act (100k + utterances)** Godfrey et al., 1992

**MRDA (110k + utterances)** Shriberg et al., 2004

**Dialy Dialog Act (100k + utterances)** Li et al., 2017

## Limitation of current systems

---

Current state-of-the art considers deep learning models

**Contextual recurrent encoder + Multi Layer Perceptron** Bothe et al. 2018

**Contextual recurrent encoder + CRF/Recurrent decoder** Chen et al., 2018

Middle/High size labelled corpora

**Switchboard Dialog Act (100k + utterances)** Godfrey et al., 1992

**MRDA (110k + utterances)** Shriberg et al., 2004

**Dialy Dialog Act (100k + utterances)** Li et al., 2017

Not practical but deep learning is data hungry!

We will rely on pretraining

# SILICONE

---

**Research only consider middle/high size corpora**

# SILICONE

---

**Research only consider middle/high size corpora**

Corpus	<i>Train</i>	<i>Val</i>	<i>Test</i>	Utt.	<i>Labels</i>	Task	Utt./ <i>Labels</i>
SwDA*	1k	100	11	200k	42	DA	4.8k
MRDA*	56	6	12	110k	5	DA	2.6k
DyDA <sub>a</sub>	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 <sub>e</sub>	11k	1k	1k	102k	7	E	2.2k
MELD <sub>S</sub> *	934	104	280	13k	3	S	4.3k
MELD <sub>E</sub> *	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 <sub>a</sub>	11k	1k	1k	102k	4	DA	25.5k
MT*	121	22	25	36k	12	DA	3k
Oasis*	508	64	64	15k	42	DA	357
<hr/>							
DyDA <sub>e</sub>	11k	1k	1k	102k	7	E	2.2k
MELD <sub>S</sub> *	934	104	280	13k	3	S	4.3k
MELD <sub>e</sub> *	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

Emotions  
&  
Sentiments



# **Text Style Transfert**

---

# Text Style Transfert

---

$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	Input	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

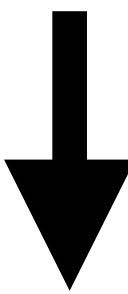
# Text Style Transfert

Transferring style is easier with disentangled representations

$\lambda$	Model	Sentence
	Input	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	Input	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

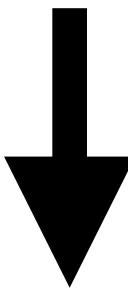
Transferring style **is easier with disentangled representations**



$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
10	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

Transferring style **is easier with disentangled representations**

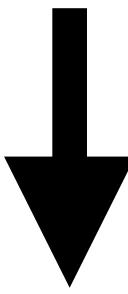


**There is no free lunch!**

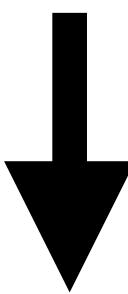
$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

Transferring style **is easier with disentangled representations**



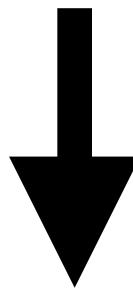
**There is no free lunch!**



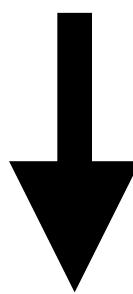
$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

Transferring style **is easier with disentangled representations**



**There is no free lunch!**

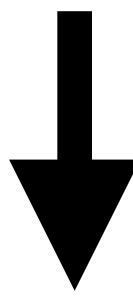


**Disentangling also removes important information about the content.**

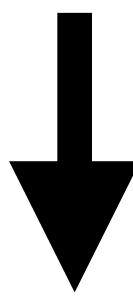
$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

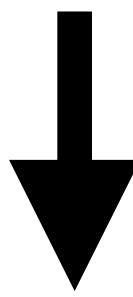
Transferring style **is easier with disentangled representations**



**There is no free lunch!**



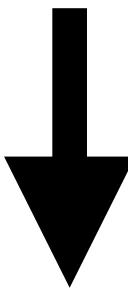
**Disentangling also removes important information about the content.**



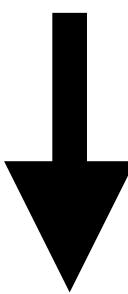
$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Text Style Transfert

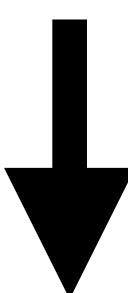
Transferring style **is easier** with disentangled representations



There is no **free lunch!**



Disentangling also **removes** important information about the **content.**



Our loss **behave better** than **CLUB** and **ADV**

$\lambda$	Model	Sentence
	<b>Input</b>	<b>It's freshly made, very soft and flavorful.</b>
0.1	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
1	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
	KL	it's very fresh, and very flavorful and flavor.
5	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
10	vCLUB-S	i hate it.
	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	<b>Input</b>	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.
	$D_{\alpha=1.8}$	it's not worth the \$ NUM.

# Research Questions

## NLU

### Spoken Language vs Written Language

How to adapt the MI maximization principle to the hierarchy of conversations and to build generic representation for transcripts that takes into account the specifics of dialogue?

What are the consequences of introducing hierarchy? How can this inductive bias be further leveraged to improve the learning phase?



### Multimodal Data

Does it make sense to apply the MI maximization principle to learn representations of multi-modal conversations?

If so how can we adapt it to multimodal data?

What new properties are learnt by the representations when using the MI maximization principle?

## NLG

### Controlled Generation

### Evaluation of Generation

# Research Questions

## NLU

### Spoken Language vs Written Language

**How to adapt the MI maximization principle to the hierarchy of conversations and to build generic representation for transcripts that takes into account the specifics of dialogue?**

**What are the consequences of introducing hierarchy? How can this inductive bias be further leveraged to improve the learning phase?**



### Multimodal Data

**Does it make sense to apply the MI maximization principle to learn representations of multi-modal conversations?**

**If so how can we adapt it to multimodal data?**

**What new properties are learnt by the representations when using the MI maximization principle?**

## NLG

### Controlled Generation

**What conditions can we introduce to learn disentangled representations to remove attribute information from the latent space?**

**How do these conditions affect the learned representations?**

**What is the trade-off that exists between the disentanglement and the quality of the representations?**

### Evaluation of Generation

# Research Questions

## NLU

### Spoken Language vs Written Language

How to adapt the MI maximization principle to the hierarchy of conversations and to build generic representation for transcripts that takes into account the specifics of dialogue?

What are the consequences of introducing hierarchy? How can this inductive bias be further leveraged to improve the learning phase?



### Multimodal Data

Does it make sense to apply the MI maximization principle to learn representations of multi-modal conversations?

If so how can we adapt it to multimodal data?

What new properties are learnt by the representations when using the MI maximization principle?

## NLG

### Controlled Generation

What conditions can we introduce to learn disentangled representations to remove attribute information from the latent space?

How do these conditions affect the learned representations?

What is the trade-off that exists between the disentanglement and the quality of the representations?

### Evaluation of Generation

Can we use the measure of information to propose a new metric that automatically evaluates text generation?

Are the measures of information flexible enough to correlate well with different task-specific criteria?