Final Project

Pierre Colombo

pierre.colombo@centralesupelec.fr

MICS - CentraleSupelec

Advanced Natural Language Processing

CentraleSupélec

Project On Sequence Labelling

Conversational AI

Google Home

Apple Siri

Amazon Alexa

Definition

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

Types of Interactions

Task-oriented

What is the weather going to be like today?

It’s currently sleeting and the temperature is 2°C. Expect rain starting in the morning.

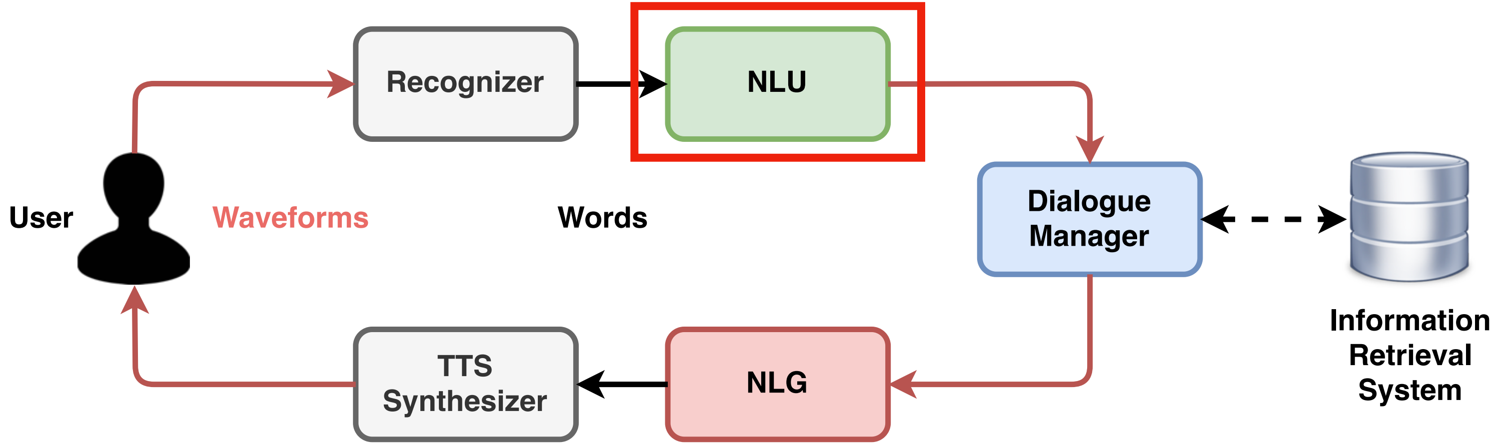
Social/Open domain

That's worrying.

What’s that?

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

Architecture of Dialogue Systems



Components of traditional dialogue systems.

We will focus on two sub-tasks of NLU:

Emotion/Sentiment (E/S) classification

Dialogue act (DA) classification

Emotions/Sentiments Labels (E/S)

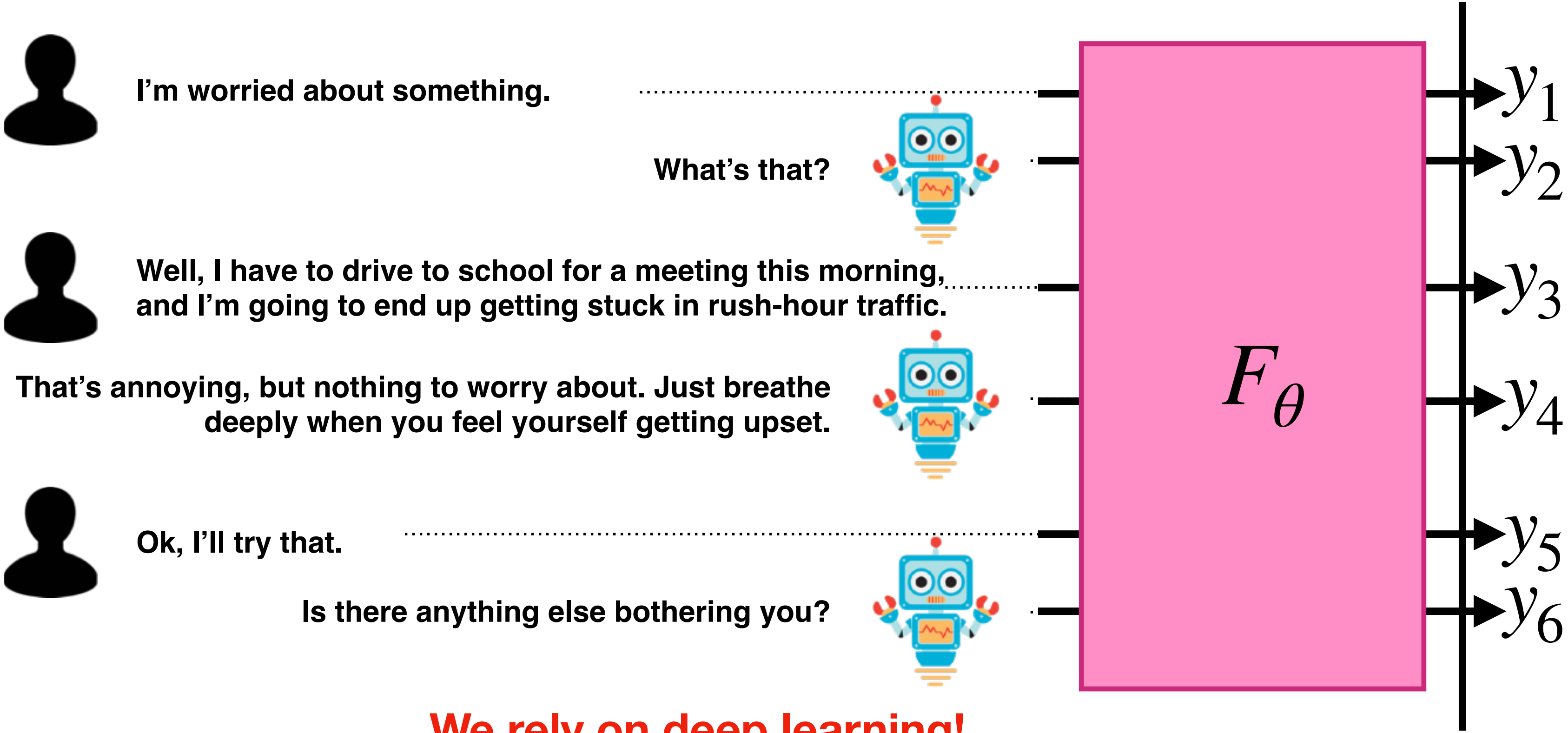
	Emotion	Sentiment
Okay, look at this one. This is my favourite.	Joy	Positive
Oh, that is so sweet !	Joy	Positive
I know ! Phoebe is gonna love dressing them in these !	Joy	Positive
Huh. Except, Phoebe's not gonna be the one that gets to dress them.	Neutral	Neutral
Because she's not gonna get to keep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

Dialog Act Labels (DA)

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

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

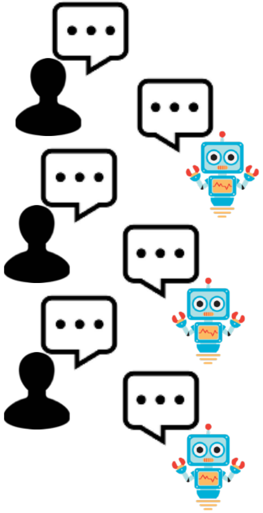
Sequence Labelling Tasks

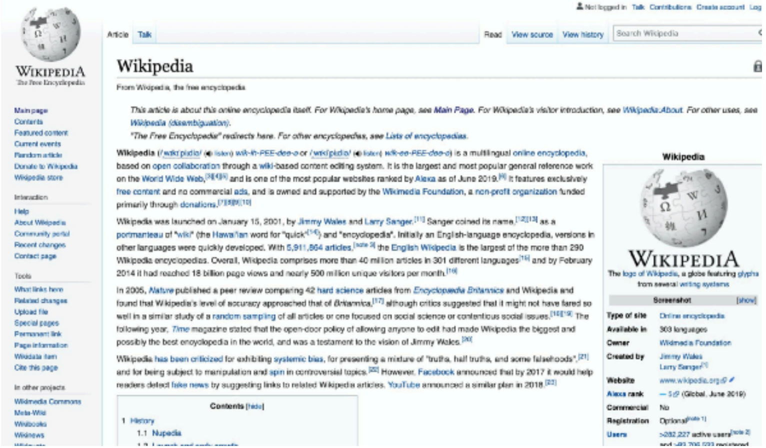


We rely on deep learning!

Data Specificities & Constraints

1) Transcripts of spoken **dialogues**





Data Specificities & Constraints

1) Transcripts of spoken dialogues

- i think, i think the stand they have, or, or the way the command respect, i, i support that.
- i think that is a, a positive thing for them after, um, uh, thousands of years,
- they have to, uh, they ha,
- i think they in,
- when they be, became a country they more than, or, more or less decided they were n't going to take it anymore
- and, uh.

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

Switchboard Corpus: 200k labelled utterances → Supervised Learning

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

Large unlabelled corpus → Self Supervised Learning + Transfert Learning

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

Monolingual dialogue

Um, what did you do this weekend?

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

Uh-Huh.

What do you have planned for your yard?

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

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

Code switching

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

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

Uh-Huh.

Qu'as tu prévu pour ton yard?

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

Data Specificities & Constraints

1) Transcripts of spoken conversations

2) Amount of labelled data

3) Language used in the data

4) Spoken conversations are multimodal

Human Communication is multimodal.

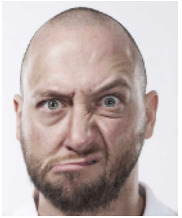
« What you say »

« How you say it »

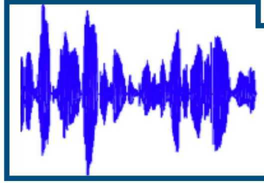
Verbal

I hate this game!

Visual



Audio



Formalisation of Sequence Labelling

Conversation C

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

u_1 I'm worried about something.

What's that? u_2

u_3 Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.

That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset. u_4

u_5 Ok, I'll try that.

Is there anything else bothering you? u_6

Formalisation of Sequence Labelling

Utterance

$$u_i = (w_1^i, w_2^i, \ldots, w_n^i)$$

u_1 I'm worried about something.

$w_1^1 w_2^1$ w_3^1 w_4^1 w_5^1 w_6^1

What's that? u_2

u_3 Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.

That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset. u_4

u_5 Ok, I'll try that.

Is there anything else bothering you? u_6

Hierarchy is an important feature in dialogue!

Independent Utterance Level Prediction

u_1 I'm worried about something.

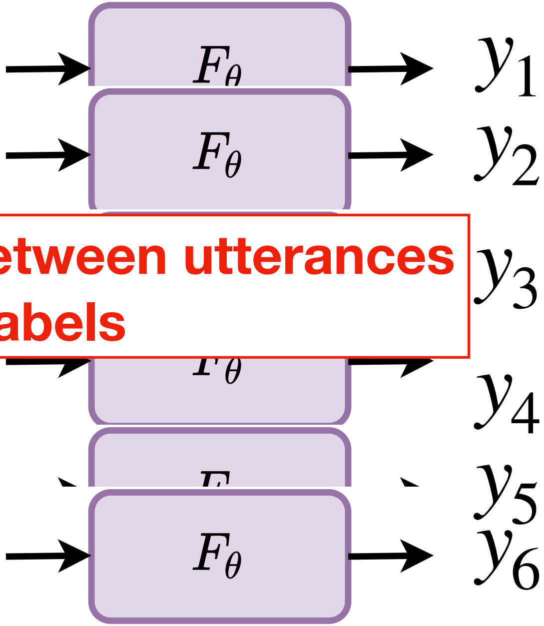
u_2 What's that?

u_3 Well, I have to dr
and I'm going to

u_4 That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.

u_5 Ok, I'll try that.

u_6 Is there anything else bothering you?



Ignore the dependencies between utterances and between labels

Modelling Utterances Dependencies

Context

Context

u_1 I'm worried about something.

u_2 What's that?

u_3 Well, I have to dr
and I'm going to

No explicit conditioning on previous labels predictions

u_4 That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset. $\rightarrow F_\theta \rightarrow y_4$

u_5 Ok, I'll try that. $\rightarrow F_\theta \rightarrow y_5$

u_6 Is there anything else bothering you?

Modelling utterances and labels dependencies

u_1 I’m worried about something.

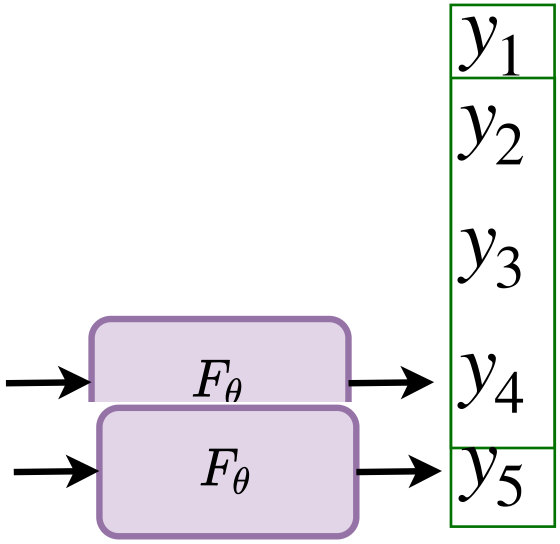
u_2 What’s that?

u_3 Well, I have to drive to school for a meeting this morning, and I’m going to end up getting stuck in rush-hour traffic.

u_4 That’s annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.

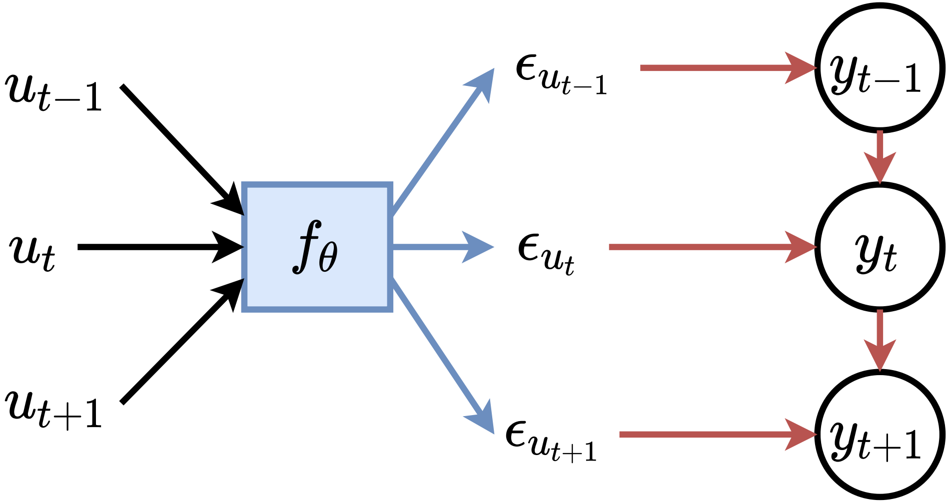
u_5 Ok, I’ll try that.

u_6 Is there anything else bothering you?



Structured Prediction with CRF

$$\boxed{u_1, \ldots, u_T \xrightarrow{F_\theta} y_1, \ldots, y_T}$$



Need for global dependencies

$$u_1, \ldots, u_n \xrightarrow{F_\theta} y_1, \ldots, y_n$$

Is there anyone who doesn't know Nancy?

Do you- Do you know Nancy?

Me?

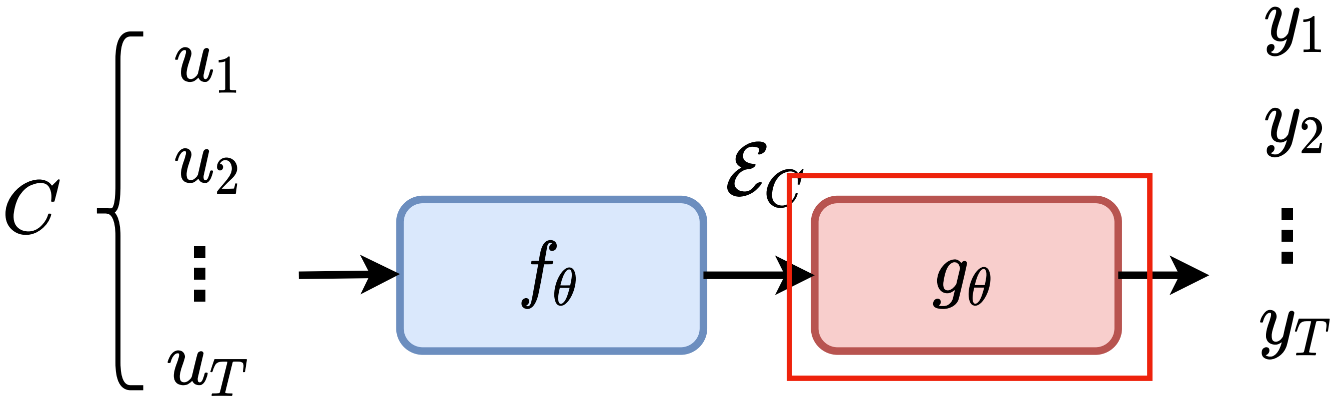
Mm-hmm

I know Nancy

Neural Machine Translation (NMT) the problem of global dependencies has been addressed using seq2seq models

Seq2Seq approach

$$\boxed{u_1, \ldots, u_T \xrightarrow{F_\theta} y_1, \ldots, y_T}$$



Goal: Encode the conversation into a single hidden vector which is then decoded to produce an output sequence

Monolingual Evaluation

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
$\texttt{DyDA}_\texttt{a}$	11k	1k	1k	102k	4	DA	25.5k
MT*	121	22	25	36k	12	DA	3k
Oasis*	508	64	64	15k	42	DA	357
$\texttt{DyDA}_\texttt{e}$	11k	1k	1k	102k	7	E	2.2k
$\texttt{MELD}^\star_\texttt{s}$	934	104	280	13k	3	S	4.3k
$\texttt{MELD}^\star_\texttt{e}$	934	104	280	13k	7	S	1.8k
IEMO	108	12	31	10k	6	E	1.7k
SEM	62	7	10	5,6k	3	S	1.9k

Dialog Acts

Emotions & Sentiments

Multilingual Evaluation

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

Project On Text Similarity

Evaluating Natural Language Generation (NLG)

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Generated Sentence

Human

S_2 J'apprécie courir dehors !

Groundtruth Sentence

Automatic NLG evaluation is made complex by language variability

Contrarily to other tasks such as Classification there is no « obvious » measure of performance

On automatic evaluation in NLG

Automatic Evaluation is used as a surrogate of human judgement

1. Cheap
2. Fast
3. Reproductible
4. Easy to use

Automatic NLG evaluation is needed at various development stages

1. Compare different systems.

Benchmark !!!

2. Debug NLG systems without annotators. (Karpinska et al. 2021)

3. Improve learning of systems by deriving new losses.

Let's formalize the problem of automatic evaluation

s_1: The weather is cold today.

s_2: It is freezing today

→ 0.8 Similar

s_1: I like those cats.

s_2: It is freezing today

→ 0.1 Dissimilar

We want to build a metric m

$$m : \mathcal{S} \times \mathcal{S} \rightarrow [0,1]$$

$$(S_1, S_2) \rightarrow m(S_1, S_2)$$

Success Criterion:

When do we know that m is good?

→ **Correlation with human scores**

(Koehn 2009; Specia, Raj, and Turchi 2010; Chatzikoumi 2020)

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

Embedding Based

Intuition

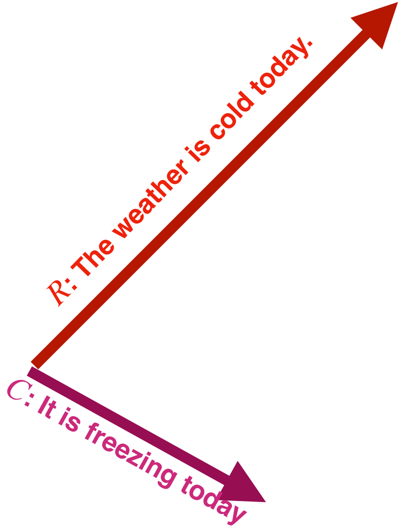
R: The weather is cold today.

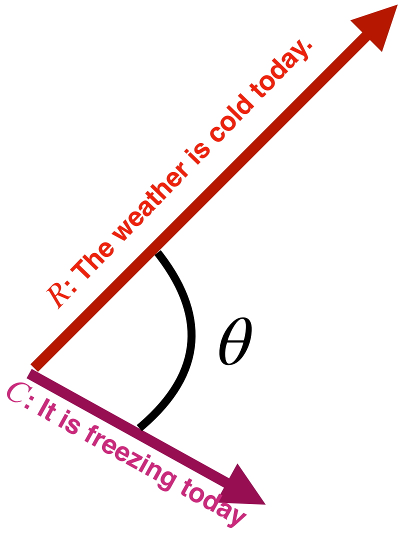
C: It is freezing today

0.8

1. Choose your embedding

2. Choose a similarity function





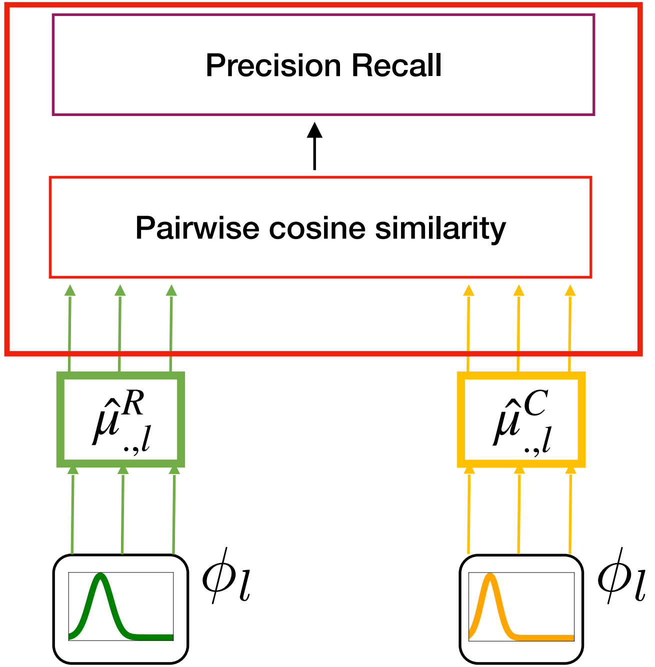
Advantage

1. Deal with paraphrases
2. Include "semantic"

Limitation

1. Not interpretable

BertScore



R: The weather is cold today

C: It is freezing this morning

This is a distance between two empirical distributions !

Advantage

1. **Deal with paraphrases**
2. **Include "semantic"**

Limitations

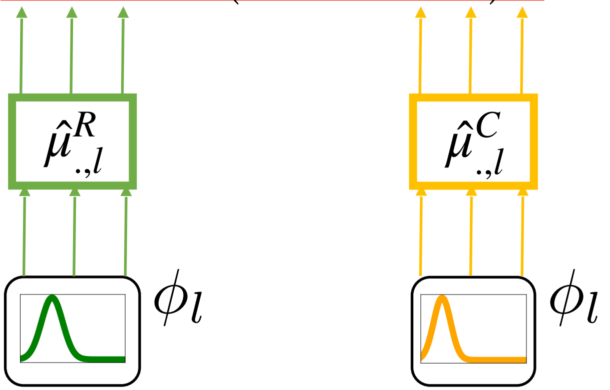
1. **Use only one layer**
2. **Use arbitrary sequence of operation**

Still not interpretable

DepthScore

This is a distance between two empirical distributions !

$$DR_{p,\varepsilon}(\mu,\nu) = \left(\int_{\varepsilon}^{1} d_{\mathscr{H}}\left(D_{\mu}^{\alpha}, D_{\nu}^{\alpha}\right)^{p} d\alpha \right)^{\frac{1}{p}}$$



R: The weather is cold today

C: It is freezing this morning

Advantage

1. **Deal with paraphrases**
2. **Include "semantic"**

Limitations

1. **Use only one layer**

Still not interpretable

Embedding Based Metric With Neural Networks

Intuition

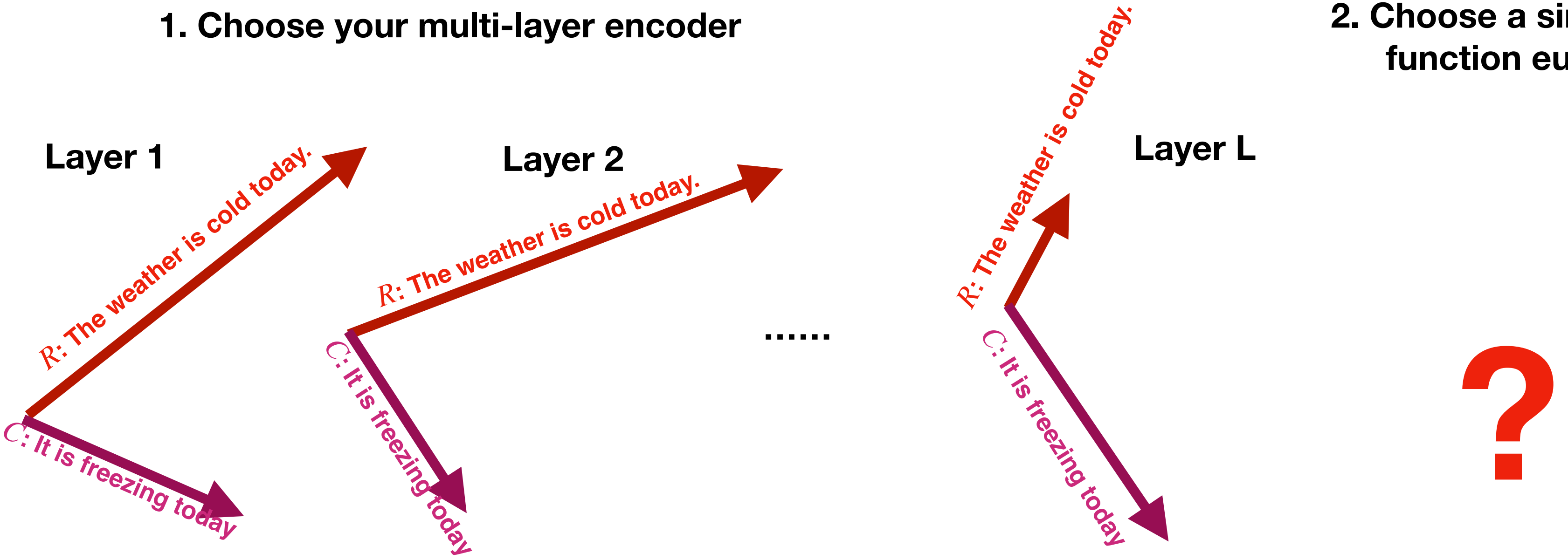
R: The weather is cold today.

C: It is freezing today

0.8

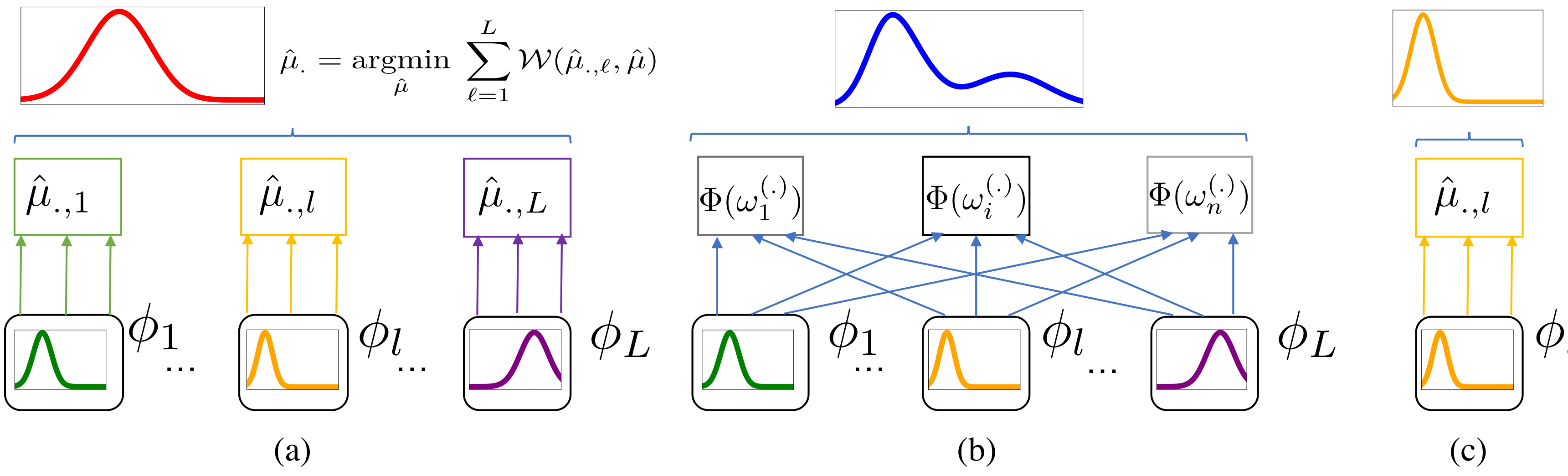
1. Choose your multi-layer encoder

2. Choose a similarity function euh??



BaryScore vs BertScore vs MoverScore

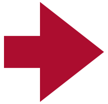
Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. (Oral) EMNLP 2021



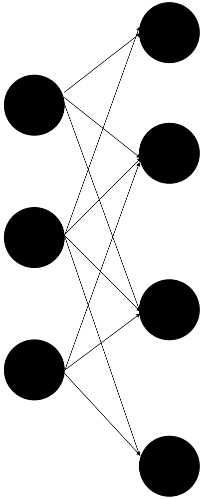
BaryScore MoverScore BertScore

Statistical Measures of Similarity

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!

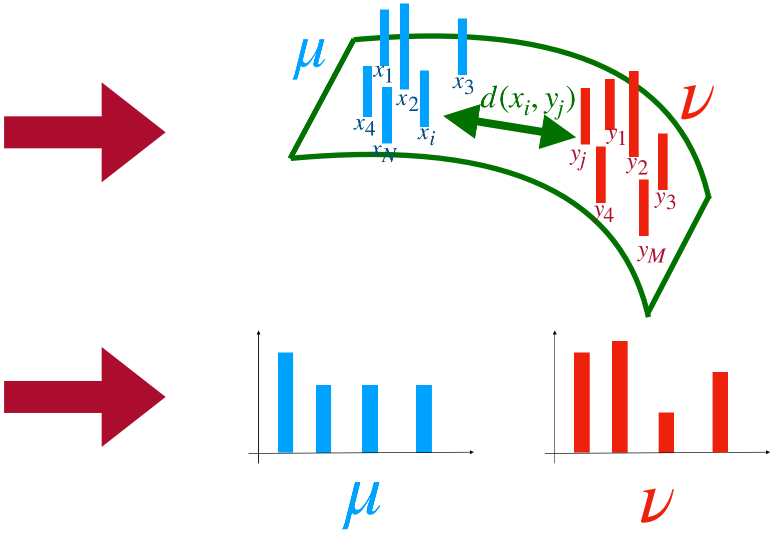


Input Text



Neural Network

High dimensional data



Soft Probabilities

InfoLM

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

Assumptions for InfoLM

Goal Compute a similarity score between R and C.

Tools Use a **pretrained MLM**

MLM predicts a distribution over Ω

$p_\Omega(\cdot \mid [R]^i)$

Use a **measure of information**

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

Name	Notation	Domain	Expression
α-divergence (Csiszár 1967)	$\mathcal{D}_\alpha$	$\alpha \notin \{0,1\}$	$\frac{1}{\alpha(\alpha-1)}(1-\sum q_i^{1-\alpha} p_i^\alpha)$
γ divergence (Fujisawa and Eguchi 2008)	$\mathcal{D}_\gamma^\beta$	$\beta \notin \{0,-1\}$	$\frac{1}{\beta(\beta+1)} \log \sum p_i^{\beta+1} + \frac{1}{\beta+1} \log \sum q_i^{\beta+1} - \frac{1}{\beta} \log \sum p_i q_i^\beta$
AB Divergence (Cichocki, Cruces, and Amari 2011)	$\mathcal{D}_{sAB}^{\alpha,\beta}$	$(\alpha,\beta) \in (\mathbb{R}^*)^2$, $\beta+\alpha \neq 0$	$\frac{1}{\beta(\beta+\alpha)} \log \sum p_i^{\beta+\alpha} + \frac{1}{\beta+\alpha} \log \sum q_i^{\beta+\alpha} - \frac{1}{\beta} \log \sum p_i^\alpha q_i^\beta$
$\mathcal{L}_1$ distance	$\mathcal{L}_1$		$\sum \lvert p_i - q_i \rvert$
$\mathcal{L}_2$ distance	$\mathcal{L}_2$		$\sqrt{\sum (p_i - q_i)^2}$
$\mathcal{L}_\infty$ distance	$\mathcal{L}_\infty$		$\max_i \lvert p_i - q_i \rvert$
Fisher-Rao distance	R		$\frac{2}{\pi} \arccos \sum \sqrt{p_i \times q_i}$

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 Ω

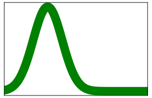
$p_\Omega(\cdot \mid [R]^i)$

MLM

Similar context

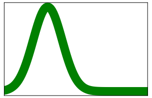
R: It is [MASK] today.

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



C: It is f [MASK] his morning !

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

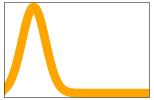


$$\mathcal{I}\left(p_\Omega(\cdot \mid [R]^2), p_\Omega(\cdot \mid [C]^2)\right) \sim 0$$

Dissimilar context

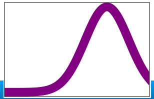
R: It is cold t [MASK]

$\xrightarrow{p_\Omega(\cdot \mid [R]^3)}$



C: It is f [MASK] his morning !

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



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

Context Aggregation

Goal Compute a similarity score between R and C.

How to aggregate contexts? → Weighted Sum!

Reference

[MASK] s cold today.

...

It is [MASK] oday.

...

It is cold today. [MASK]

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

$$InfoLM(R, C) \triangleq \mathcal{I}\left(P, Q\right)$$

Candidate

[MASK] is freezing this morning !

...

It is [MASK] his morning !

...

It is freezing this morning ! [MASK]

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

Setting

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance Fluency / Text Structure
- Results on English only

Gardent et al. 2017

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

Summary Generation

- Results on SummEval
- Correlation with pyramid score
- Results on English only

Nallapati et al. 2016)

Bhandari et al. (2020)

Nenkova and Passonneau 2004

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

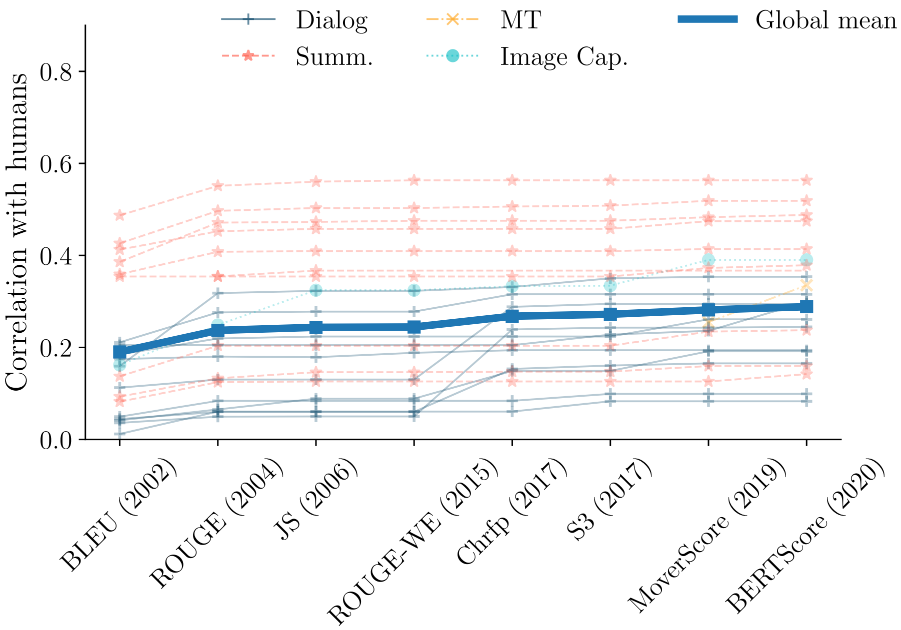
John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

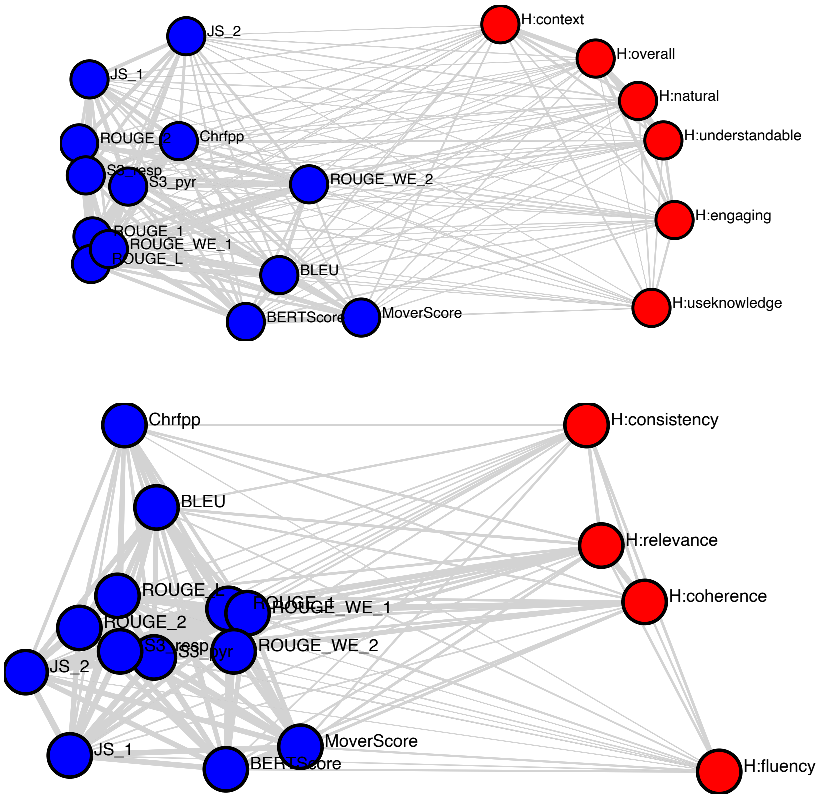
Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	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	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	86.6	91.4	75.0
$\mathcal{D}_{\alpha}$	88.8	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	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
$\mathcal{R}$	**89.7**	86.0	75.0	78.7	70.5	51.6	**93.3**	**95.7**	**85.3**	87.6	84.4	70.0	**92.4**	93.8	81.6
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	85.5	83.4	73.3	74.7	68.2	53.3	92.3	95.5	85.0	83.3	79.4	65.0	91.9	**95.0**	**83.3**
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	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	89.4	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	59.6	48.3	76.5	76.3	60.3	64.3	69.2	56.7	75.9	77.5	58.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	86.7	90.5	78.3	82.1	65.2	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	93.5	80.0	75.0	70.2	**77.6**	89.5	91.1	78.6

Parameter Free

Are we really making progress?

P. Colombo, M. Peyrard, N. Noiry, R. West, P. Piantanida. The Glass Ceiling of Automatic Evaluation in Natural Language Generation





Takeaways:

C. Chung, P. Colombo, F Suchanek , C. Clavel. *Of Human Criteria and Automatic Metrics: A Survey and Benchmark of the Evaluation of Story Generation.* (oral) COLING 2022

Prompt: When you die the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.

Human: 3,000 years have I been fighting. Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place I came from before this. [...]

Generated #1: First of all, not everyone was entitled to be an exterminator. But the ones that were – maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people – had all the opportunity to work for the damn plant killer, and it's hard work. [...]

Generated #2: It was hell. Not exactly a place of torture. There were no guards in prison and you couldn't just walk through it, either, because you would get killed regardless. hell was a young man, and he was lying on his floor. He was unconscious. [...]

Story	RELEVANCE	COHERENCE	EMPATHY	SURPRISE	ENGAGEMENT	COMPLEXITY
Human	5	5	1	3	4	1
	2	2	3	2	2	3
	4	4	3	2	4	4
Generated #1	2	4	3	1	1	1
	2	2	2	1	2	2
	2	3	2	3	3	3
Generated #2	5	5	3	3	3	2
	3	2	3	2	2	3
	3	4	3	4	4	3

Metric	Human	Generated #1	Generated #2
BLEU	100	0.95	1.18
ROUGE-1	1.00	0.24	0.33
chrF	100	31.6	38.9
BERTScore	1.00	0.50	0.52
MoverScore	1.00	0.51	0.51
BaryScore	0.00	0.92	0.92
S3	1.39	0.07	0.15
BARTScore	−0.98	−3.97	−4.03
SUPERT	0.94	0.37	0.36

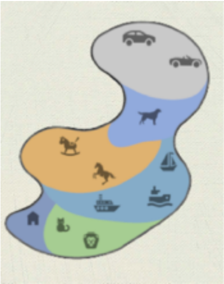
Human evaluation

Automatic evaluation

How to aggregate the scores and decide which system is the best?

Project On Anomaly Detection

DNN



Training $\longrightarrow$ Inference

where:

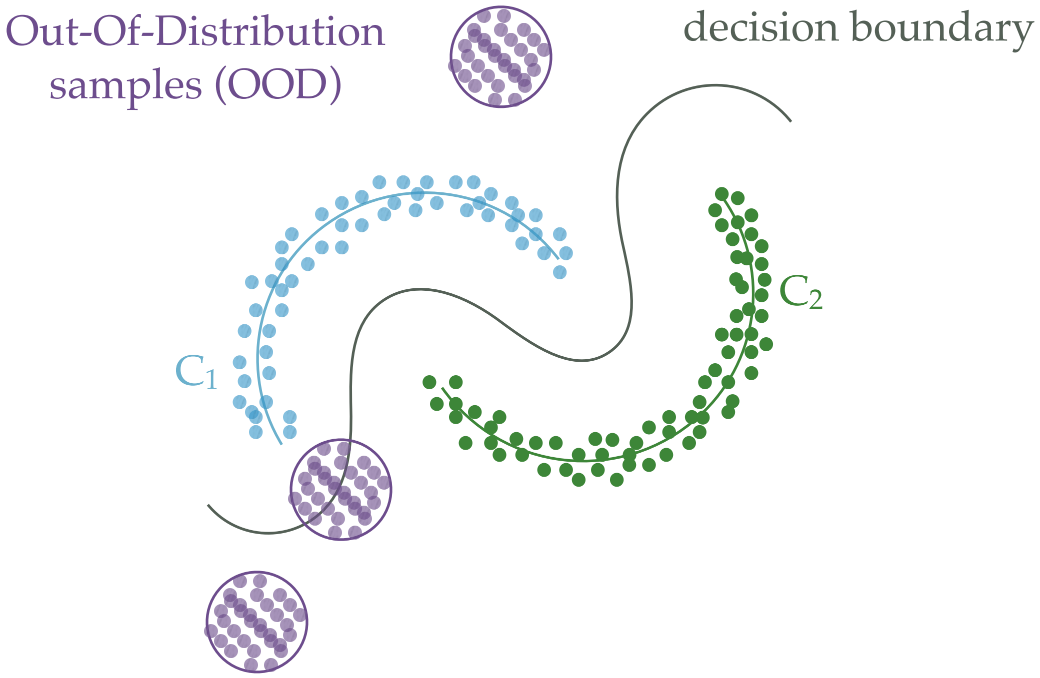
- $\mathbf{x}_i$ is a training text
- y_i its target
- $\boldsymbol{\theta}$ are the parameters of the DNN
- $f^L_{\boldsymbol{\theta}}(\cdot)$ are the logits of the DNN
- $\mathbf{x}$ is a testing image
- $\hat{y}$ its predicted label according to the DNN.

$$\mathbf{x}_i \longrightarrow \text{DNN} \xrightarrow{f^L_{\boldsymbol{\theta}}(\mathbf{x}_i)} \text{Loss} \longrightarrow \boldsymbol{\theta}, \qquad y_i \longrightarrow \text{Loss}$$

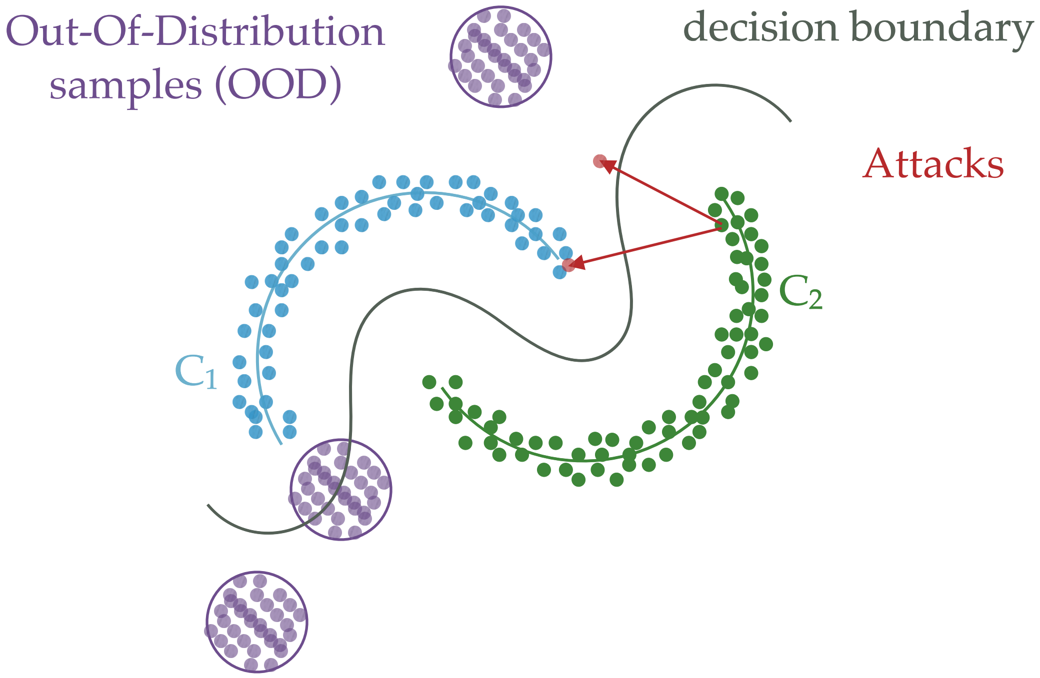
$$\mathbf{x} \longrightarrow \text{DNN} \xrightarrow{f^L_{\boldsymbol{\theta}}(\mathbf{x})} \text{argmax} \longrightarrow \hat{y}$$

Securing Neural Networks

Trusting the environment



Malicious Attacks



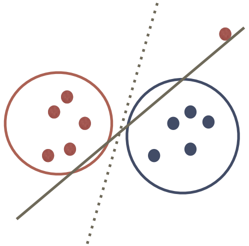
Why Attacks?

- Can be seen as a worst-case scenario
- Essential in critical systems
- Gives insights into DNNs behaviour

Different type of attacks?

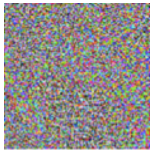
Training → Inference

Data Poisoning



Adversarial Attacks[5]

+0.007



=

"Panda" 57.7% confidence

"nematode" 8.2% confidence

"Gibbon" 99.3% confidence

Adversarial problem[6]

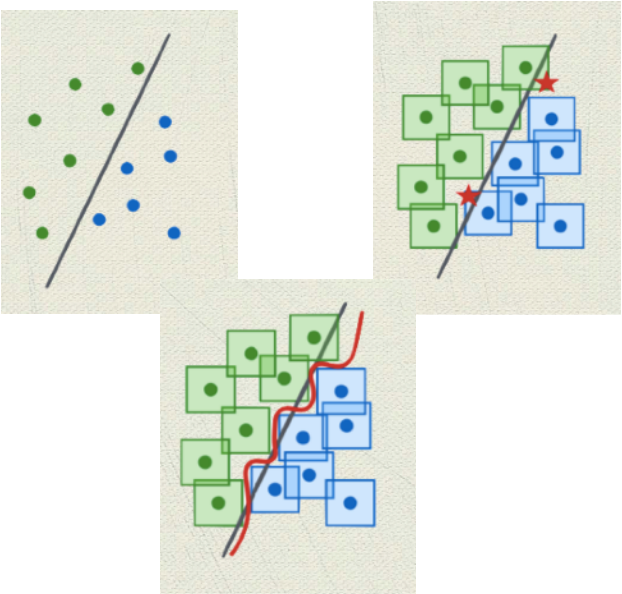
$$\mathbf{x}' = \arg\min_{\mathbf{x}'\in[0,1]^d} \|\mathbf{x}' - \mathbf{x}\|_p \text{ s.t. } \arg\max f_{\boldsymbol{\theta}}^L(\mathbf{x}') = t,$$

where:

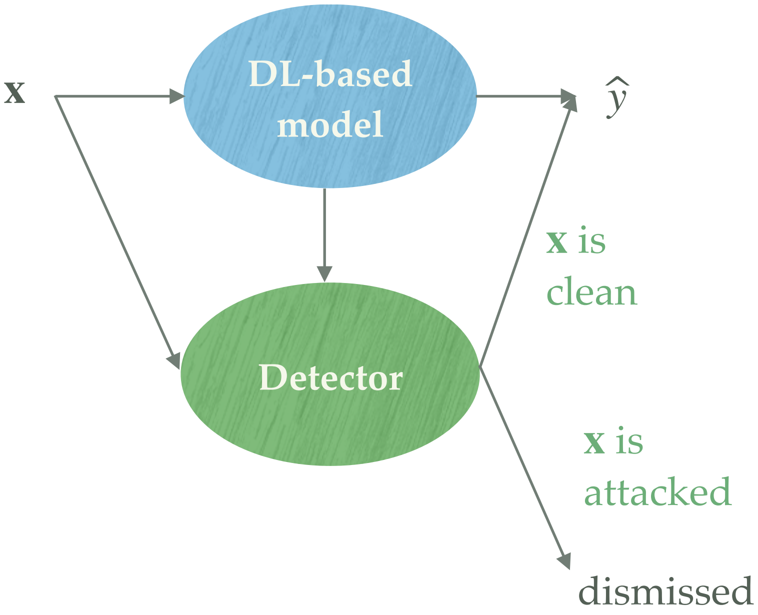
- $(\mathbf{x}, y)$ clean sample
- $\mathbf{x}'$ adversarial sample
- $f_{\boldsymbol{\theta}}^L$ deep classifier parametrized by $\boldsymbol{\theta} \in \Theta$
- t either targeted class or any class other than y

Different type of defense

Robustness[7]

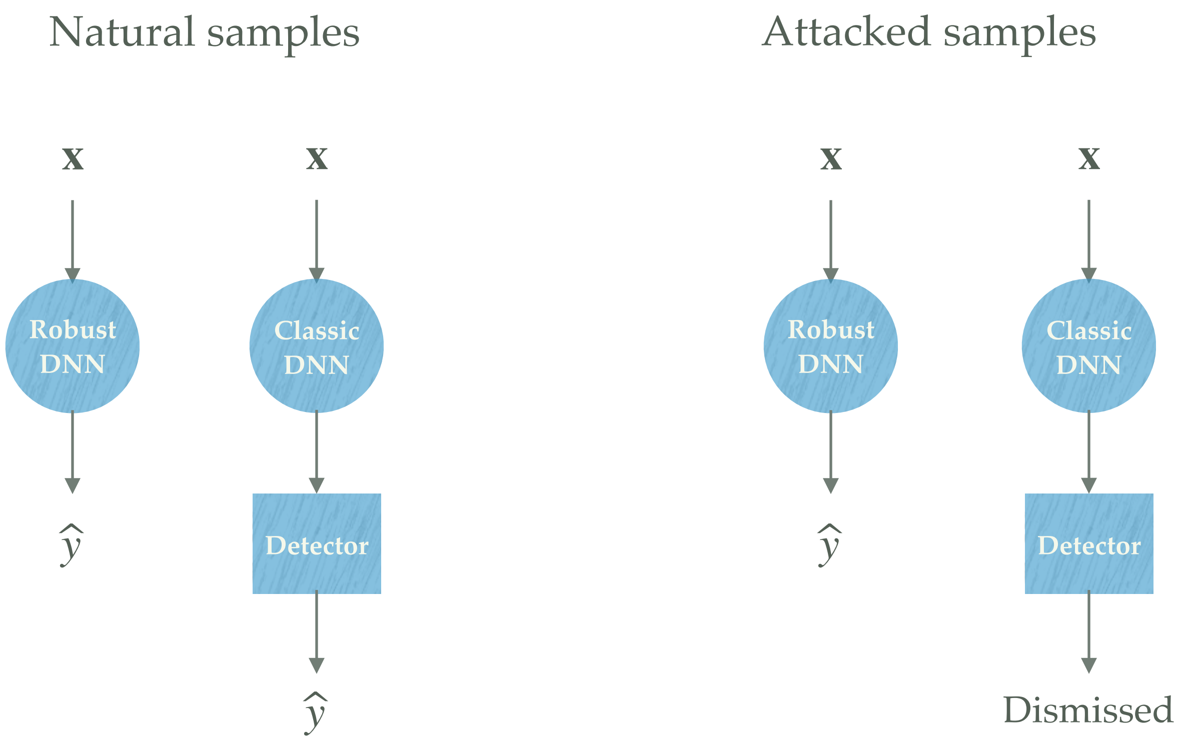


Detection



[7] Alexander Madry, et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*, ICLR 2018.

Abstention

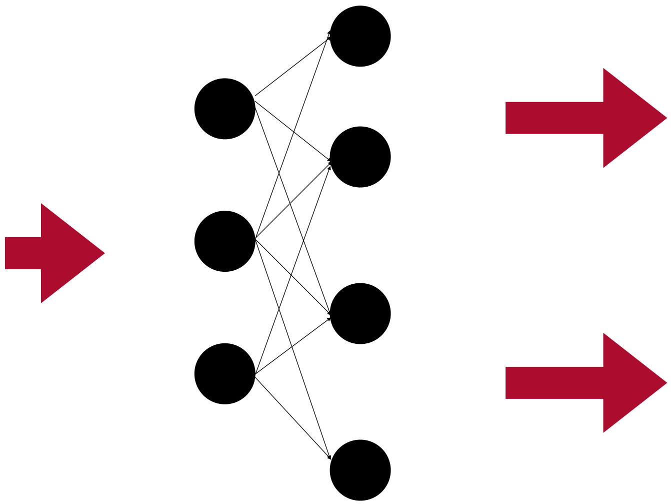


Natural samples
Attacked samples
x
x
x
x
Robust DNN
Classic DNN
Robust DNN
Classic DNN
$\hat{y}$
Detector
$\hat{y}$
Detector
$\hat{y}$
Dismissed

Neural Network

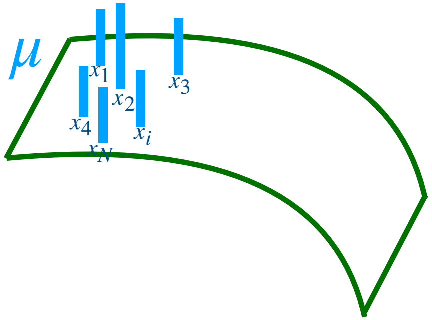
Hello, Chicago.
If there is anyone out
there who still doubts that
America is a place where
all things are possible,
who still wonders if the
dream of our founders is
alive in our time, [….].
Yes we can!

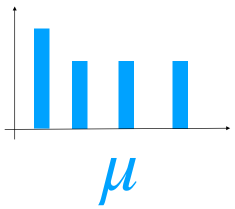
Input Text



Neural Network

High dimensional data

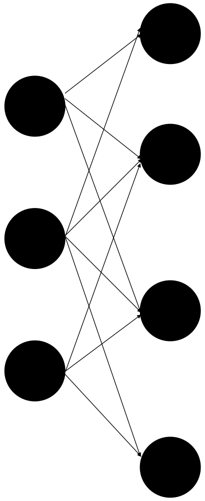


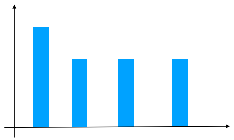


Soft Probabilities

Softmax Based Scores

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!





μ

MSP

$$max\ p(\cdot\,|\,x)$$

Doctor

$$\sum_i p^2(i\,|\,x)$$

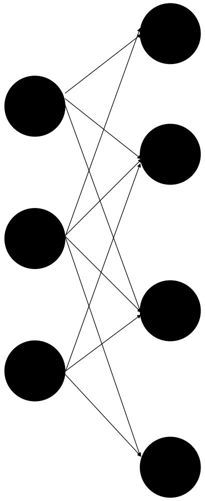
Input Text

Neural Network

Neural Network

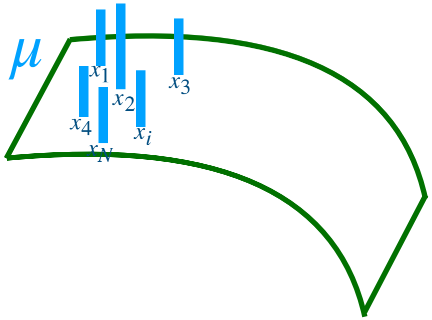
Hello, Chicago.
If there is anyone out
there who still doubts that
America is a place where
all things are possible,
who still wonders if the
dream of our founders is
alive in our time, [….].
Yes we can!

Input Text



Neural Network

High dimensional data

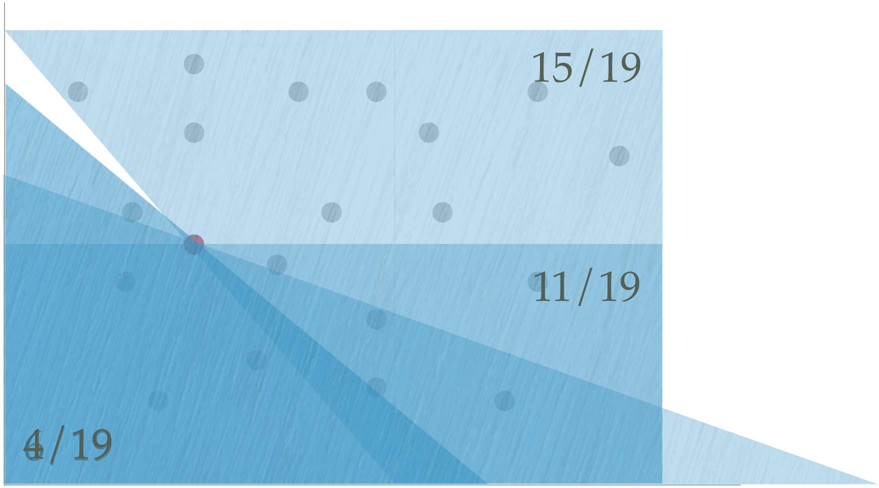


Cosine
Depth
Mahalanobis

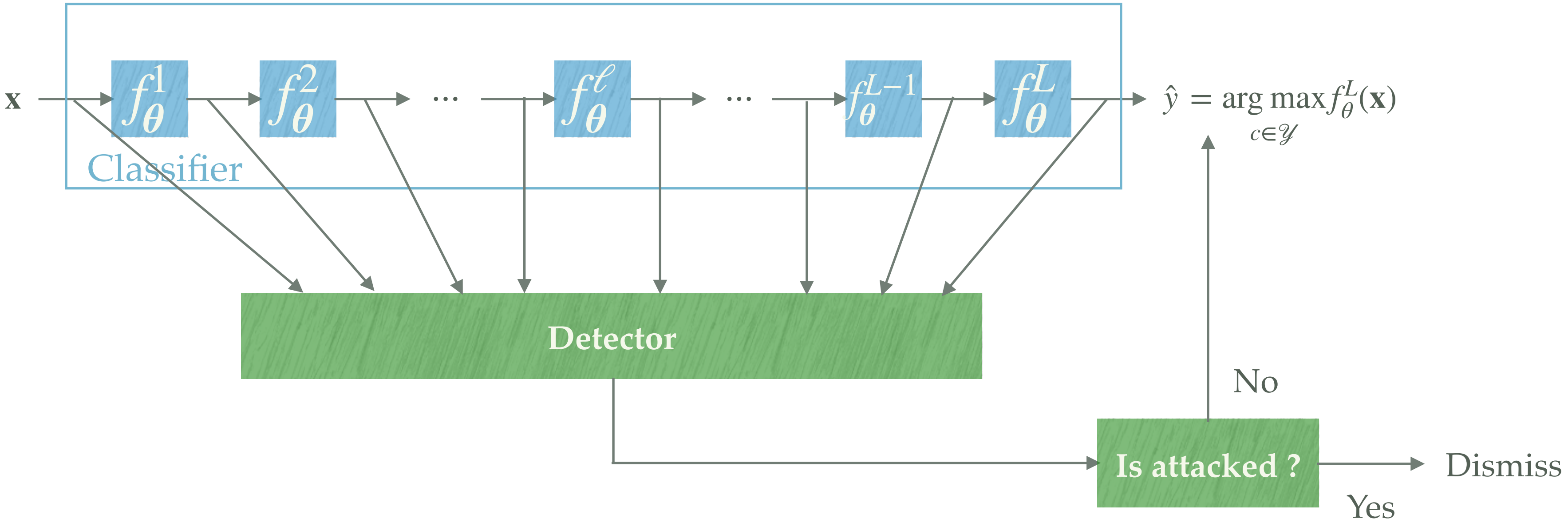
Data Depth

$$D : \mathbb{R}^d \times \mathscr{P}(\mathbb{R}^d) \longrightarrow [0,1]$$
$$(\mathbf{z}, p_{\mathbf{z}}) \longmapsto D(\mathbf{z}, p_{\mathbf{z}})$$

Halfspace depth[13]: $D_H(\mathbf{z}, p_{\mathbf{z}}) = \inf_{H \in \mathscr{H}(\mathbf{z})} p_{\mathbf{z}}(H)$



Deep Classifier - Multilayer models



Deep Classifier - Multilayer models