HARNESSING TEXT GENERATION

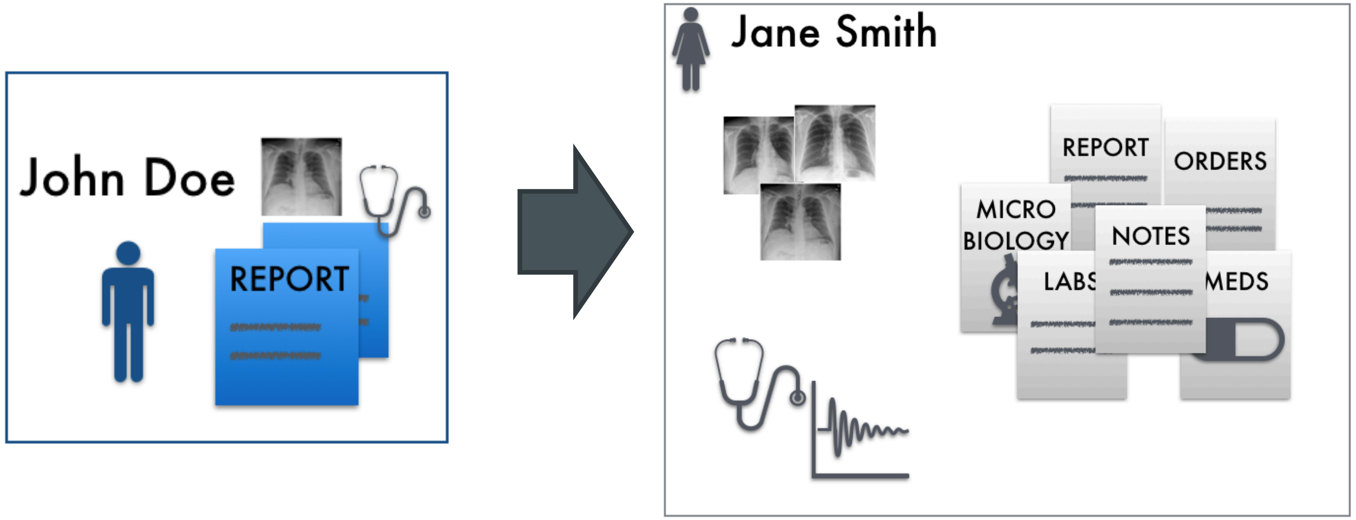
JULIA IVE

NOVEMBER 5, 2021

SYNTHETIC TEXT GENERATION

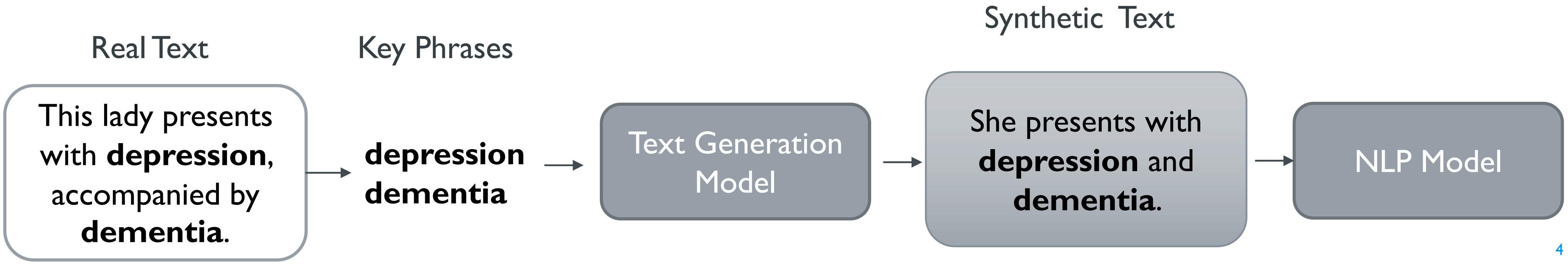
SYNTHETIC TEXT GENERATION

- Development of NLP is handicapped by data sparsity issues and access restrictions, mainly due to privacy concerns
- Textual data could be created artificially
- Key challenge: How to ensure the **validity** and **privacy** of the generated text?



SYNTHETIC TEXT GENERATION: METHODOLOGY

- Methodology to generate synthetic text for Healthcare NLP *(Ive et al. 2020, Nature Digital Medicine)*
- One of the first methodologies of the kind
- **Validity**:
  - Guide text generation with key phrases extracted from the real clinical text
  - Compare the performance of models built with synthetic data to the performance of models built with the real data
- **Privacy**:
  - Privacy safety of key phrases can be easily controlled
  - Use de-identified data for the training of the text generation models



SYNTHETIC TEXT GENERATION : METHODOLOGY

Real Text

This lady presents primarily with **depressive symptoms**, accompanied by what may be cognitive decline and **visual hallucinations**. I think it is reasonable to assume a diagnosis of **depressive episode of moderate severity** (ICD-10 F32.1). The possibility of **early dementia** needs to be borne in mind, but requires re-assessing following resolution of her **depressive symptoms**.

Meta info - patient gender and age, record type, etc.

Synthetic Text

He presents with **depressive symptoms** and frequent **visual hallucinations**. **Depressive episode of moderate severity** is highly likely. **Early dementia** can not be excluded after the resolution of his **depressive symptoms**.

SYNTHETIC TEXT GENERATION : METHODOLOGY

Real Text

Meta info - patient gender and age, record type, etc.

Synthetic Text

This lady presents primarily with **depressive symptoms**, accompanied by what may be cognitive decline and **visual hallucinations**. I think it is reasonable to assume a diagnosis of **depressive episode of moderate severity** (ICD-10 F32.1). The possibility of **early dementia** needs to be borne in mind, but requires re-assessing following resolution of her **depressive symptoms**.

He presents with **depressive symptoms** and **frequent** **visual hallucinations**. **Depressive episode of moderate severity** is highly likely. **Early dementia** can not be excluded after the resolution of his **depressive symptoms**.

< 10% of rare n-grams appear in the output if not present in the input

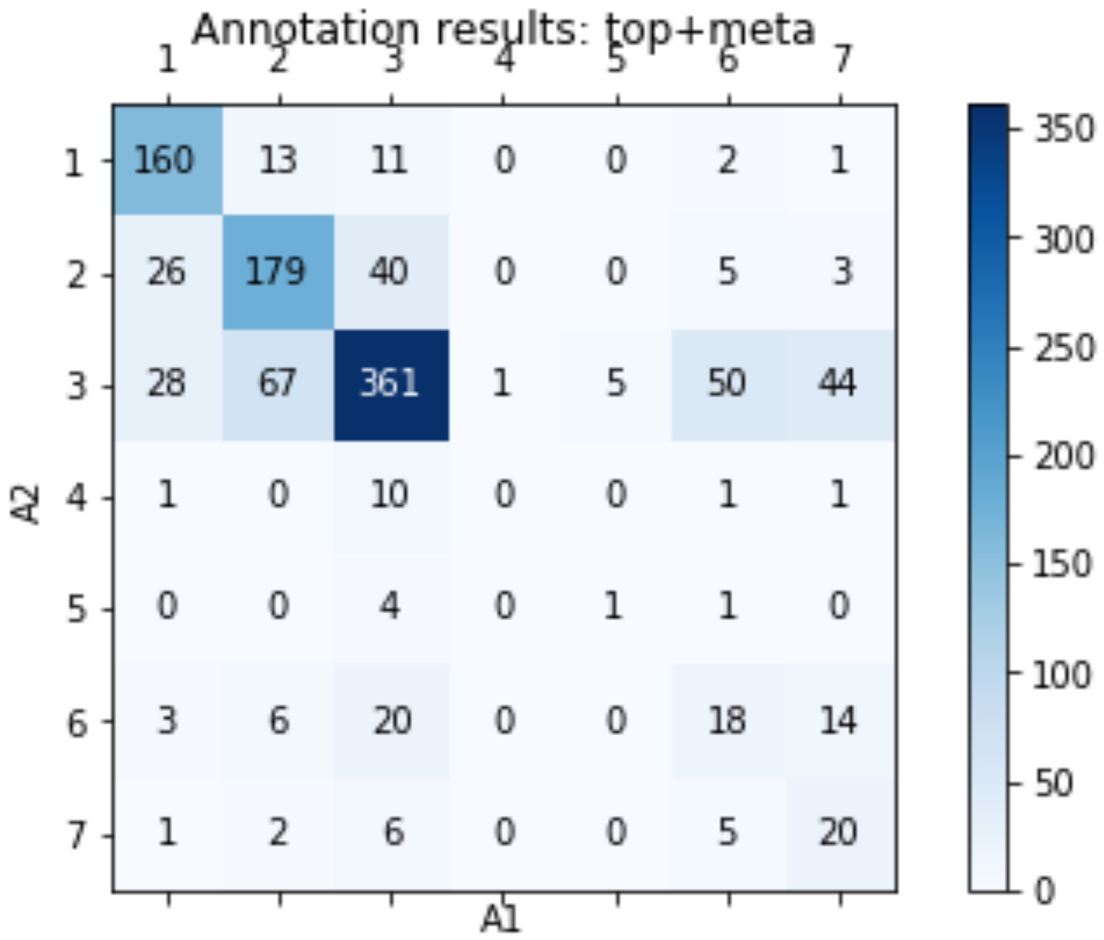
SYNTHETIC TEXT GENERATION: EVALUATION

1. Meaning fully preserved
2. Meaning preserved, details omitted
3. Meaning modified, does not contradict the diagnosis
4. Meaning modified, contradicts the diagnosis
5. Meaning modified, irrelevant
6. Makes no sense from the clinical point of view
7. Incomprehensible

- ~60 synthetic documents (1K sent) each annotated by 2 medical students
- av. accuracy 0.78, Kappa 0.54 (moderate)
- most frequent categories:
  - Meaning modified, does not contradict the diagnosis (~49%)
  - Meaning preserved, details potentially omitted (~24%)

SYNTHETIC TEXT GENERATION : EVALUATION

1. Meaning fully preserved
2. Meaning preserved, details omitted
3. Meaning modified, does not contradict the diagnosis
4. Meaning modified, contradicts the diagnosis
5. Meaning modified, irrelevant
6. Makes no sense from the clinical point of view
7. Incomprehensible



SYNTHETIC TEXT GENERATION: EXAMPLES

- Meaning modified, does not contradict the diagnosis:

REAL: That afternoon police were called after she assaulted her mother.

GEN: **This afternoon police were called by her mother.**

- Meaning preserved, details potentially omitted:

REAL: He did not have a clear understanding of why he was there or what was the problem with him.

GEN: **He has no clear understanding why he is there.**

- Incomprehensible:

REAL: This relapse of heavy drinking may have been caused by the disruption of her accommodation relocation.

GEN: **It was felt that heavy drinking may not be a accommodation relocation.**

- Makes no sense from the clinical point of view:

REAL: She acknowledged that paracetamol overdose could damage her liver.

GEN: **Paracetamol overdose could damage her shoulder.**

SYNTHETIC TEXT GENERATION: EVALUATION

- Discharge summaries from the Clinical Record Interactive Search (CRIS) database of mental health records at the Maudsley NIHR Biomedical Research Centre (30K)

ICD-10	Diagnosis	Freq., %
F20	Schizophrenia	29
F32	Major depressive disorder, single episode	21
F60	Specific personality disorders	16
F31	Bipolar affective disorder	14
F25	Schizoaffective disorder	11
F10	Mental and behavioural disorders due to use of alcohol	9

SYNTHETIC TEXT GENERATION : EVALUATION

- Main errors are due to FPs
- Artificial model lower number of FNs

	F1	
	LDA	CNN
Real	0.39	0.48
Synthetic	0.38	0.43

SYNTHETIC TEXT GENERATION: SOME CONCLUSIONS

- Basic methodology to generate synthetic health records: selecting keyphrases is particularly challenging, requires background knowledge and is target task dependent
- Release of artificial data is in general beneficial for healthcare, benefit seems higher than the risk - conclusion from the workshop with NHS governance and users
- More work is required to develop privacy-safety norms

SIMULTANEOUS MULTIMODAL MACHINE TRANSLATION

MACHINE TRANSLATION (MT)

- Task of translation from one natural language into another
- Translation process requires human background knowledge and is highly dependent on the context
- **Multimodal Machine Translation (MMT)**
  - Integration of the visual context

Woman covering her face with her hat. → MT Model → Eine Frau bedeckt ihr Gesicht mit einer Mütze.

MULTIMODAL MACHINE TRANSLATION (MMT)

Woman covering her face with her **hat**. → MT Model → Eine Frau bedeckt ihr Gesicht mit einer **Mütze**.

https://www.freepik.com/free-photo/smiling-lady-mittens-scarf-hat-eyes_3397688.htm#page=1&query=Winter%20hat&position=18

MULTIMODAL MACHINE TRANSLATION (MMT)

Woman covering her face with her **hat**.

MT Model

Eine Frau bedeckt ihr Gesicht mit einem **Hut**.

RL FOR MULTIMODAL SIMT

- Task of translating a continuous source stream
- **Motivation**: Visual input is helpful when textual context is incomplete
- Find the balance between how much context is needed to generate the translation reliably (**quality**), and how long the listener has to wait (**latency**)
- Learn a reinforcement learning policy to emit:
  - **READ** from source actions
  - **WRITE** translation word actions

RL FOR MULTIMODAL SIMT

- Deterministic policies *(Ma et al., 2019; Caglayan et al., 2020*) are simple and effective for similar language pairs
- However, they have poor generalisation ability in practise, especially for more distant language pairs
  - When significant restructuring is required while translating, e.g.:
    - **EN**: Yesterday I have been to London
    - **DE**: Gestern bin ich in London gewesen
      (Yesterday have I to London been)

RL FOR MULTIMODAL SIMT: EVALUATION

- Translation quality: BLEU *(Papineni et al., 2002)*
- Latency: Average Lag (AVL) *(Ma et al., 2019)*

$$\mathrm{AL}(X,Y) = \frac{1}{\tau}\sum_{t=1}^{\tau} g(t) - \frac{t-1}{\gamma} \qquad (\gamma = \frac{|Y|}{|X|})$$

# of tokens the **writer** lags behind the **reader**, as a function of the **# of input** tokens read.

- Latency: Average Proportion (AVP)
  - Number of source tokens required to commit a translation *(Cho and Esipova, 2016)*

RL FOR MULTIMODAL SIMT

REINFORCE algorithm with baseline *(Gu et al., 2017)*

- Agent with two actions: READ / WRITE
- MT model is the static environment (not updated during training)
  - Unidirectional GRU encoder/decoder with attention

- Simple policy gradient algorithm
- Baseline network addresses the reward variance: we estimate the gradients by subtracting the its rewards from the current rewards

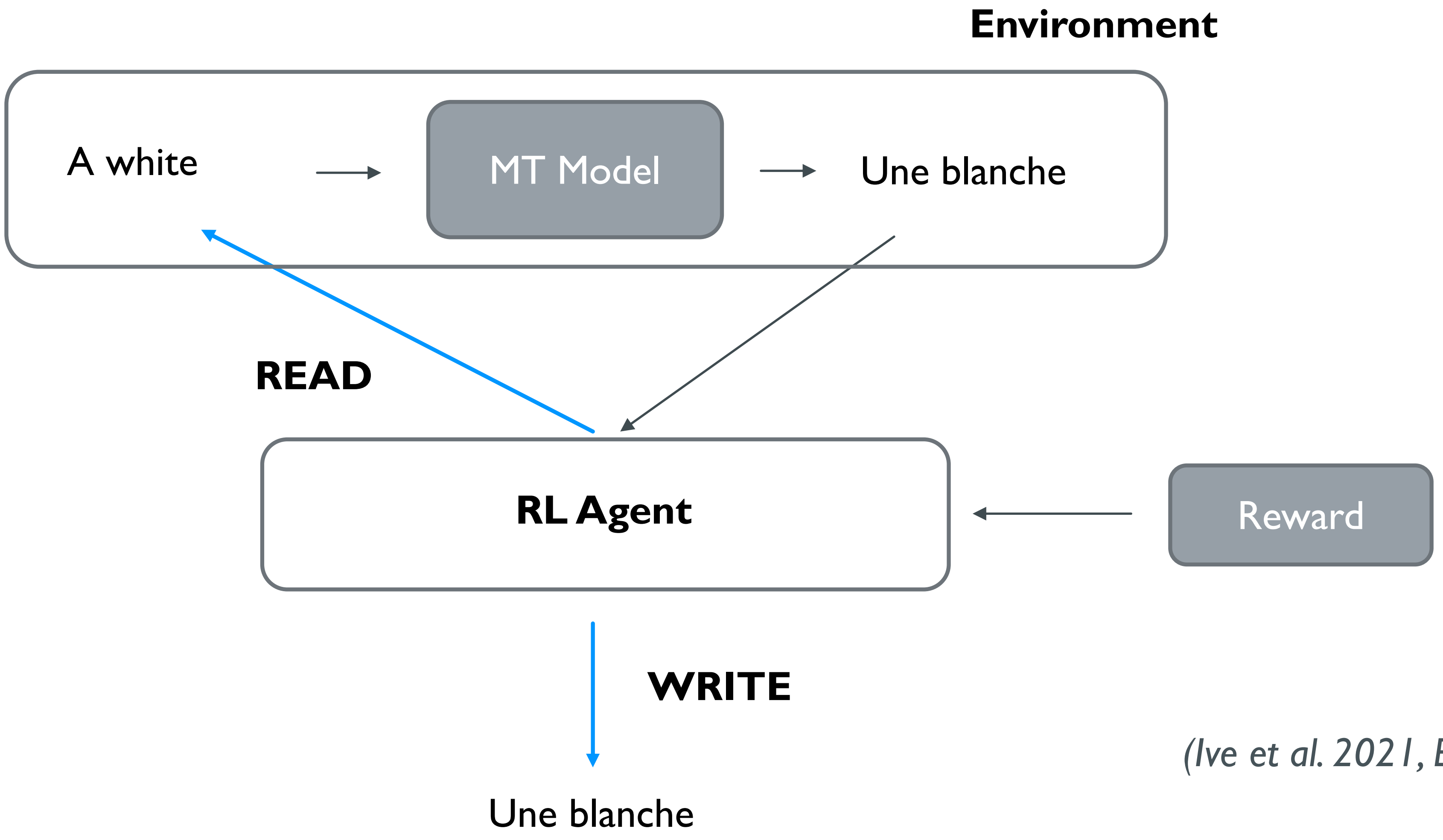
RL FOR MULTIMODAL SIMT : TEXT ONLY RESULTS

- Translation quality and latency competitive with deterministic (wait-k) policies
  (Caglayan et al., 2020)

Standard (not simultaneous) NMT

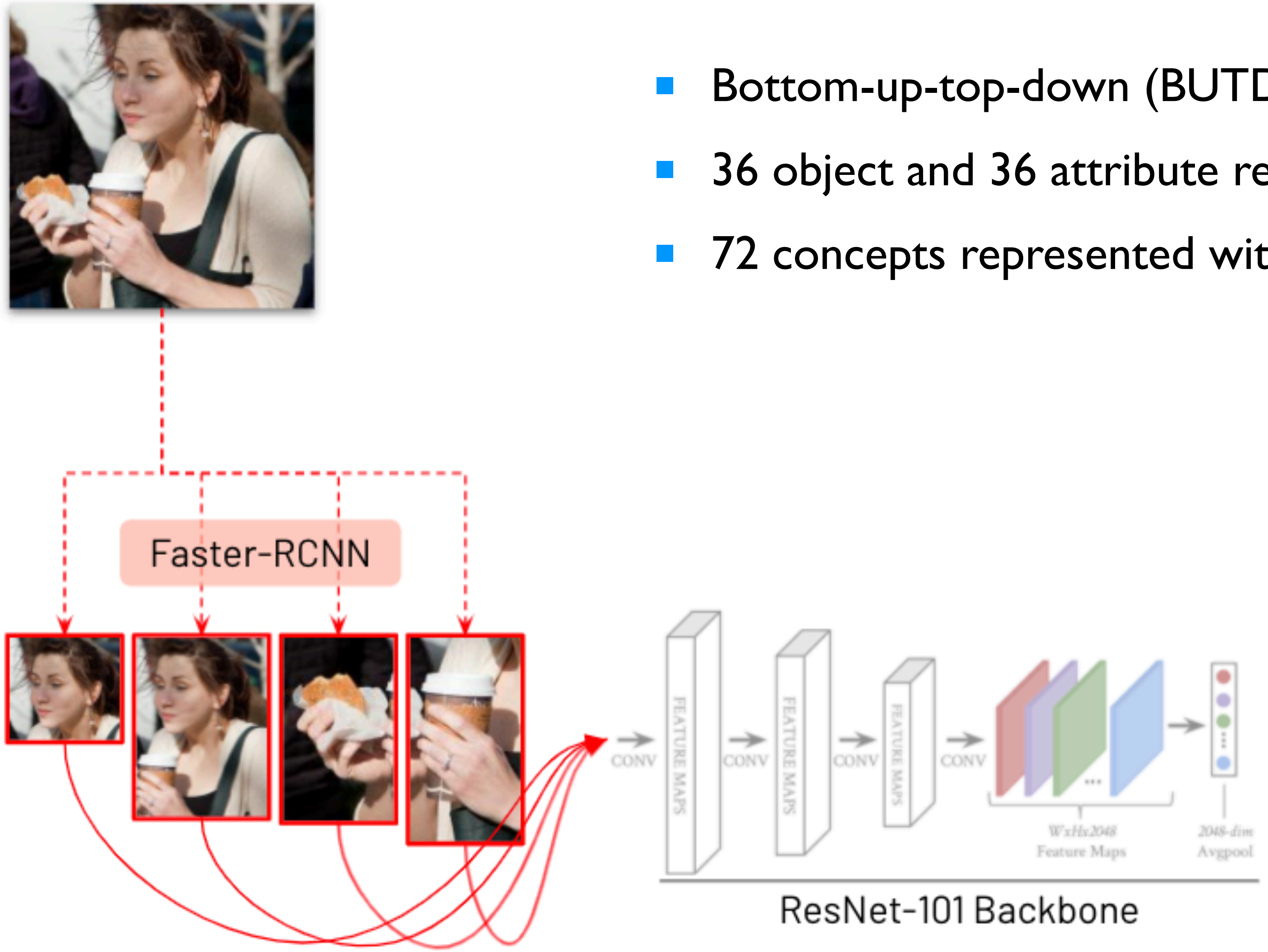
		test 2016			test 2017		
		BLEU↑	AVL↓	AVP↓	BLEU↑	AVL↓	AVP↓
EN → FR	Consecutive	58.0	13.1	1.0	50.6	11.1	1.0
	Wait-2	48.1	2.6	0.7	42.9	2.6	0.7
	Wait-3	54.0	3.5	0.7	48.6	3.5	0.7
	RL	50.8	3.3	0.7	44.3	3.0	0.7
EN → DE	Consecutive	35.5	13.1	1.0	27.7	11.1	1.0
	Wait-2	28.3	2.2	0.6	22.5	2.2	0.7
	Wait-3	32.6	3.0	0.7	25.4	3.0	0.7
	RL	31.0	2.7	0.7	23.0	2.6	0.7

RL FOR MULTIMODAL SIMT



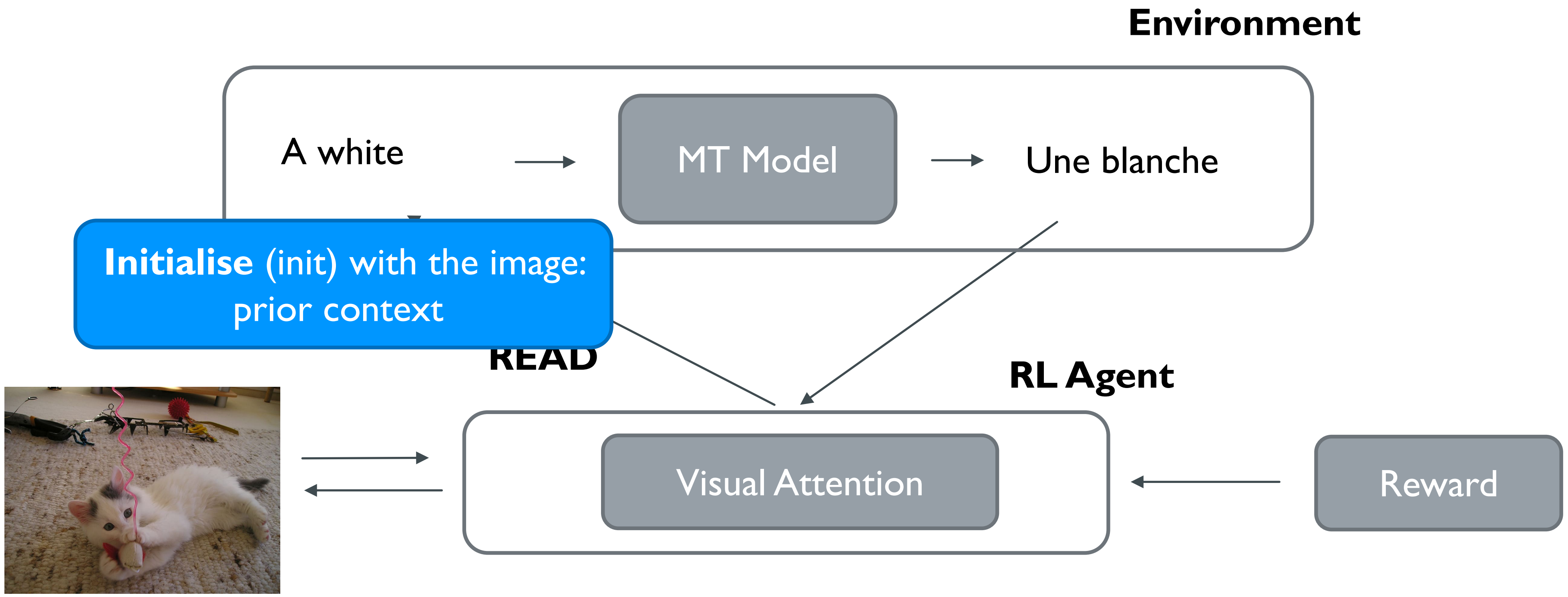
(Ive et al. 2021, EACL)

RL FOR MULTIMODAL SIMT: VISUAL FEATURES



- Bottom-up-top-down (BUTD) *(Anderson et al., 2016)*
- 36 object and 36 attribute region proposals
- 72 concepts represented with 100-dim word embeddings

RL FOR MULTIMODAL SIMT: AGENT-SIDE INTEGRATION

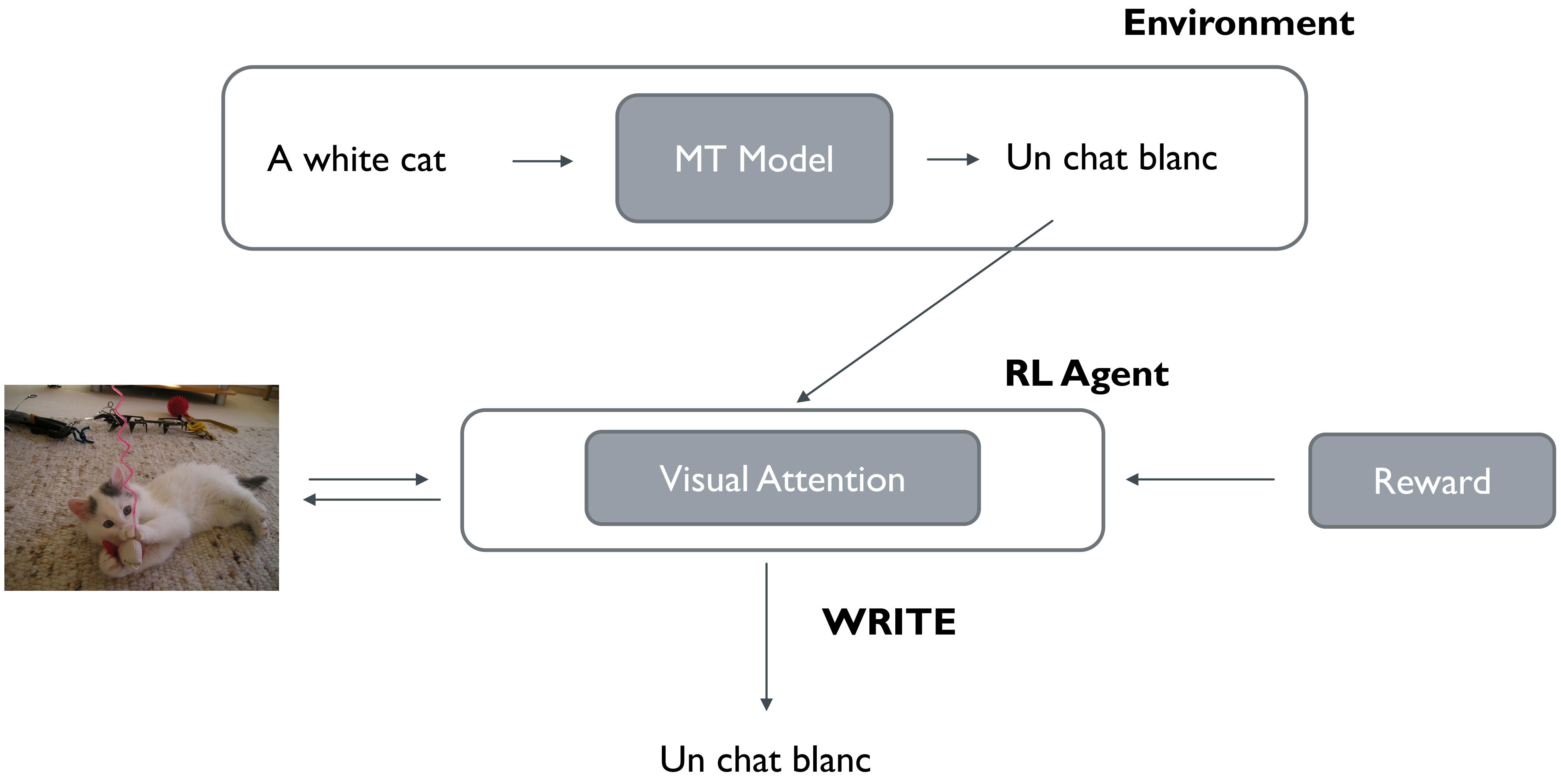


Initialise (init) with the image: prior context

Attention (att) over the visual feature vector

(Ive et al. 2021, EACL)

RL FOR MULTIMODAL SIMT: AGENT-SIDE INTEGRATION



RL FOR MULTIMODAL SIMT: ENV-SIDE INTEGRATION

Decoder Attention (env)
over the visual feature vector

Environment

A white cat → MT Model → Un chat blanc

RL Agent

Visual Attention ← Reward

WRITE

Un chat blanc

RL FOR MULTIMODAL SIMT: RESULTS

- Multi30k dataset of captions and their human translations
- Initialisation setups tend to improve the latency
- Attention setups tend to improve the quality

		test 2016			test 2017		
		BLEU↑	AVL↓	AVP↓	BLEU↑	AVL↓	AVP↓
EN → FR	Consecutive	58.0	13.1	1.0	50.6	11.1	1.0
	Wait-2	48.1	2.6	0.7	42.9	2.6	0.7
	Wait-3	54.0	3.5	0.7	48.6	3.5	0.7
	RL	50.8	3.3	0.7	44.3	3.0	0.7
	+att	53.0*	4.0	0.7	46.5*	3.7	0.8
	+init	49.6	2.8	0.7	43.3	2.6	0.7
	+init-att	52.6*	3.8	0.7	46.3*	3.6	0.7
	+env	54.0*	3.3	0.7	47.2*	3.1	0.7
	+env-init-att	54.0*	3.9	0.7	47.7*	3.8	0.8
EN → DE	Consecutive	35.5	13.1	1.0	27.7	11.1	1.0
	Wait-2	28.3	2.2	0.6	22.5	2.2	0.7
	Wait-3	32.6	3.0	0.7	25.4	3.0	0.7
	RL	31.0	2.7	0.7	23.0	2.6	0.7
	+att	33.3*	3.3	0.7	24.7*	3.0	0.7
	+init	29.7	2.8	0.7	21.3	2.4	0.7
	+init-att	34.1*	3.3	0.7	25.3*	3.1	0.7
	+env	30.0	2.5	0.6	21.7	2.2	0.6
	+env-init-att	31.4	3.0	0.7	24.0*	2.9	0.7

RL FOR MULTIMODAL SIMT: RESULTS

- Combination of initialisation, attention and environment (env-init-att) helps to find a middle ground

		test 2016			test 2017		
		BLEU↑	AVL↓	AVP↓	BLEU↑	AVL↓	AVP↓
EN → FR	Consecutive	58.0	13.1	1.0	50.6	11.1	1.0
	Wait-2	48.1	2.6	0.7	42.9	2.6	0.7
	Wait-3	54.0	3.5	0.7	48.6	3.5	0.7
	RL	50.8	3.3	0.7	44.3	3.0	0.7
	+att	53.0*	4.0	0.7	46.5*	3.7	0.8
	+init	49.6	2.8	0.7	43.3	2.6	0.7
	+init-att	52.6*	3.8	0.7	46.3*	3.6	0.7
	+env	54.0*	3.3	0.7	47.2*	3.1	0.7
	+env-init-att	54.0*	3.9	0.7	47.7*	3.8	0.8
EN → DE	Consecutive	35.5	13.1	1.0	27.7	11.1	1.0
	Wait-2	28.3	2.2	0.6	22.5	2.2	0.7
	Wait-3	32.6	3.0	0.7	25.4	3.0	0.7
	RL	31.0	2.7	0.7	23.0	2.6	0.7
	+att	33.3*	3.3	0.7	24.7*	3.0	0.7
	+init	29.7	2.8	0.7	21.3	2.4	0.7
	+init-att	34.1*	3.3	0.7	25.3*	3.1	0.7
	+env	30.0	2.5	0.6	21.7	2.2	0.6
	+env-init-att	31.4	3.0	0.7	24.0*	2.9	0.7

RL FOR MULTIMODAL SIMT

- RL-based approach allows for flexible exploration of multimodal information (agent, environment or both)
- Multimodal information accounts for performance improvement (both quality and latency)
- Initialisation setups reduce the latency and stimulate more WRITE actions
- Agent tends to exploit different kinds of image information than the environment (especially in more challenging cases: EN-DE)

REINFORCEMENT LEARNING FOR MACHINE TRANSLATION

RL FOR MT

- **Training**: maximum likelihood estimation for the corpus of N sentences:
  - Current y_t depends on the previous **gold** $y_{<t}$ and the source sentence x

$$\mathcal{L}_{\text{MLE}} = \sum_{i=1}^{N} \sum_{t=1}^{T} p(y_t^i | y_1^i, \ldots, y_{t-1}^i, x^i)$$

- **Inference**: e.g., greedy search:
  - Current $\hat{y}_t$ depends on the previously **generated** $\hat{y}_{<t}$ and the source sentence x

$$\hat{y}_t^i = argmax\, p(y | \hat{y}_1^i, \ldots, \hat{y}_{t-1}^i, x^i)$$

RL FOR MT

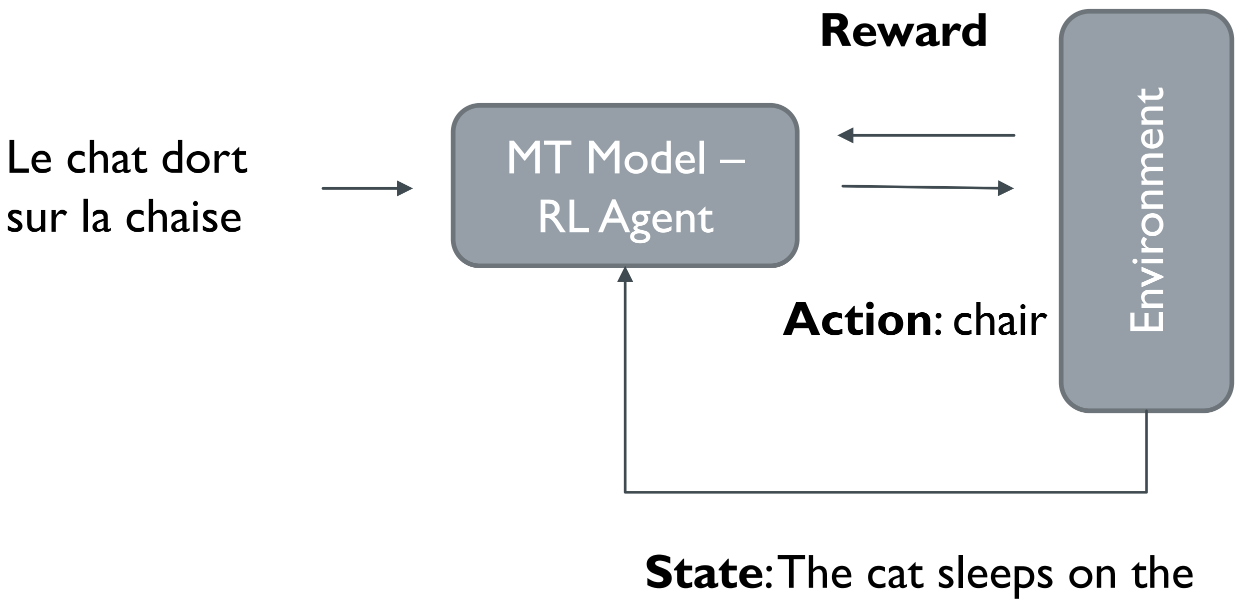
- Difference in data distribution (“exposure bias” problem)
- Difference in training and inference objectives
  - **Word-level** training supervision and **sequence-level** inference
  - Non-differentiable evaluation metrics are used at inference time:
    - They generally compare string similarity between the system output and reference outputs, e.g.: BLEU *(Papineni, 2002)* measures the n-gram precision

RL FOR MT

- Machine Translation metrics (e.g., BLEU) allow better control over quality and **Reinforcement Learning (RL)** allows to incorporate them into the training procedure
- Those rewards can though cause frequency bias to most common translations and decrease the quality of translation of ambiguous words
- Proposal of a dynamic unsupervised reward function to optimise the search space exploitation for entropy-regularised Actor-Critic Architectures *(Ive et al. 2021b, EACL)*
  - Improved generalization
  - Improved translation of ambiguous words

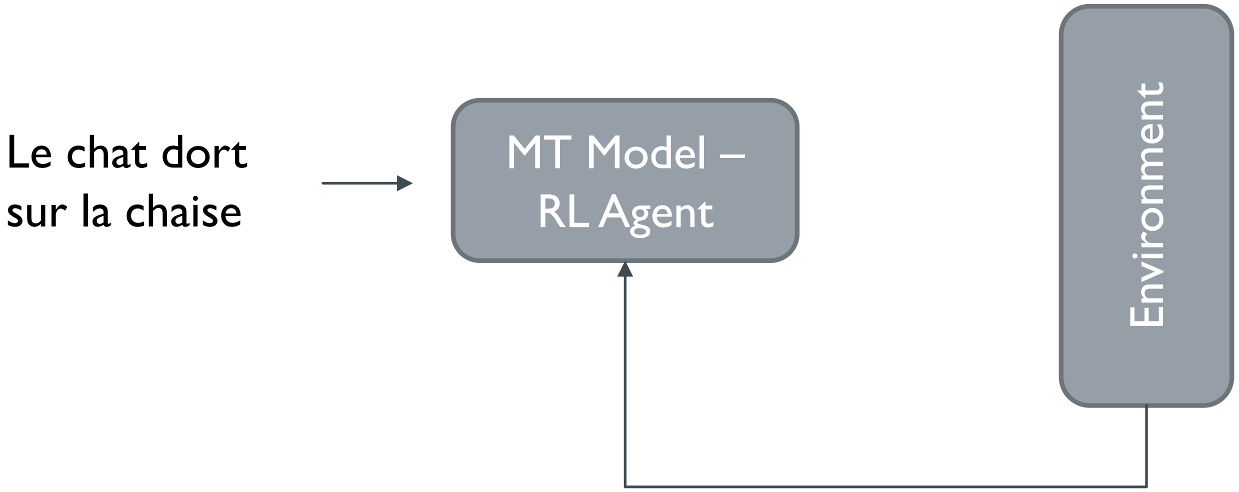
RL FOR MT

- Learn from experience
- The objective of the RL training is to maximise the expected reward



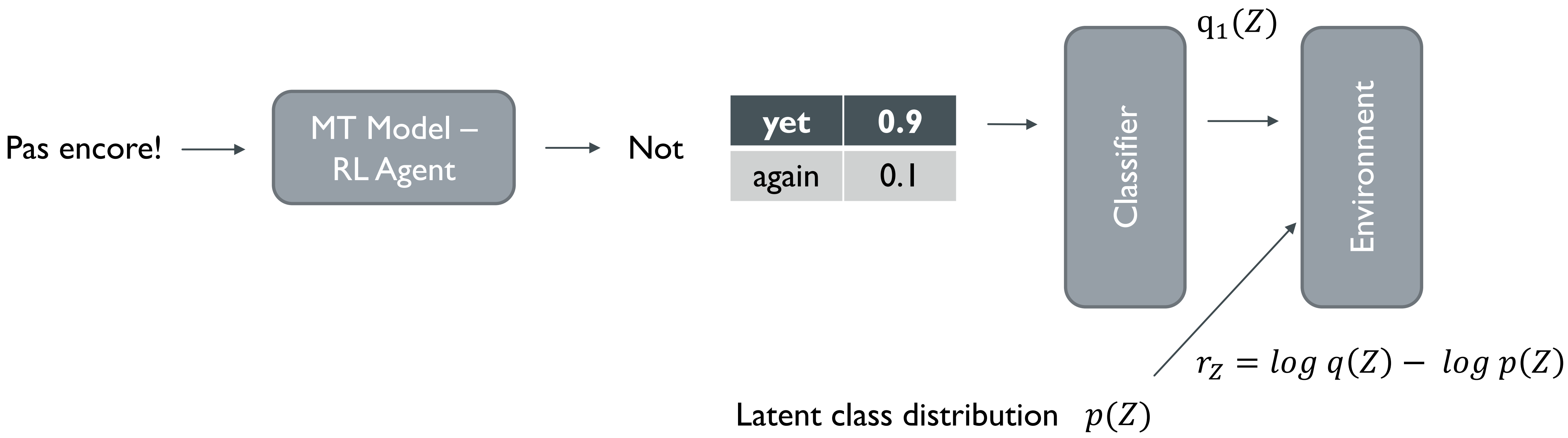
RL FOR MT

- Learn from experience
- The objective of the RL training is to maximise the expected reward

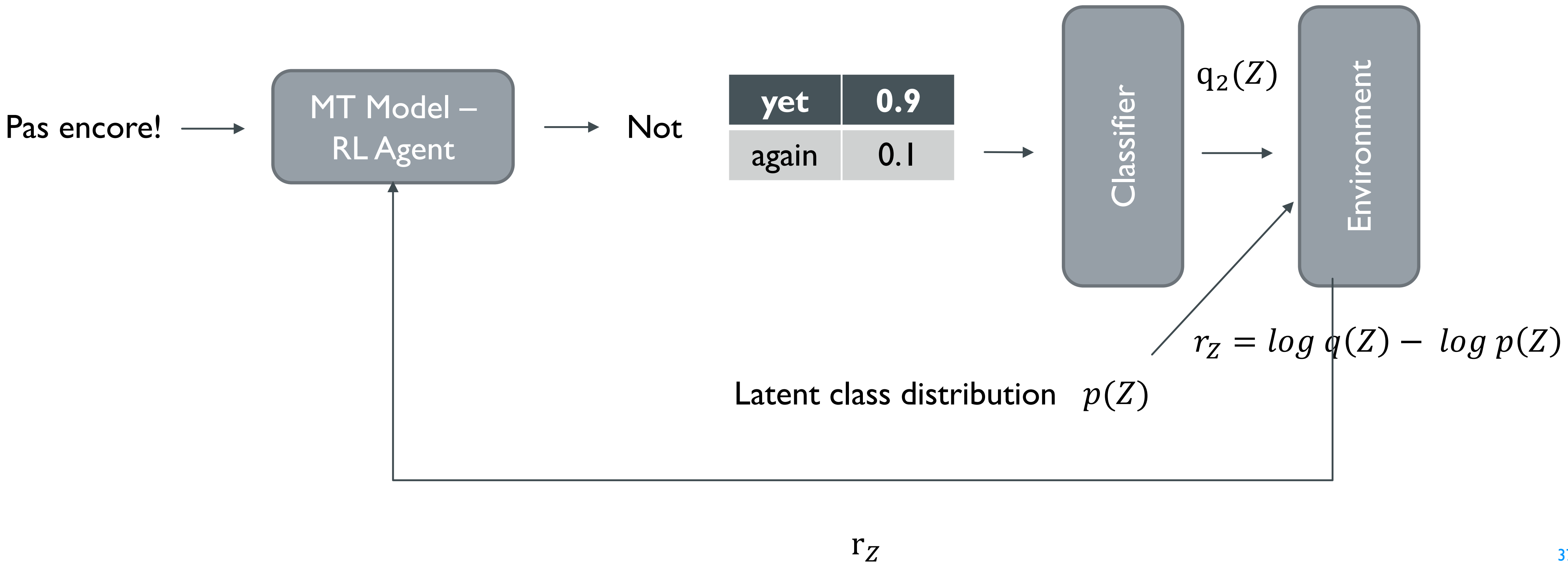


State: The cat sleeps on the chair

RL FOR MT: UNSUPERVISED REWARD



RL FOR MT: UNSUPERVISED REWARD



RL FOR MT: UNSUPERVISED REWARD

Pas encore! → MT Model – RL Agent → Not

yet	0.3
again	0.4

RL FOR MT: RESULTS -Lexical Translation Accuracy

- Lexical Translation (MLT) test set to benchmark lexical choice *(Lala et al., 2019)*

		MLT 2016↑	MLT 2017↑
EN-FR	MLE	81.60	79.65
	BLEU reward	81.94	79.76
	Unsupervised reward	82.75	80.62
EN-DE	MLE	65.34	70.91
	BLEU reward	64.74	71.93
	Unsupervised reward	65.54	73.41

RL FOR MT: EXAMPLE- English-French

Source	The teen jumps the **hill** with his bicycle
Reference	Ado saute sur la **colline** ‘hill' avec son vélo .
MLE	Adolescent saute sur la **pente** ‘slope' avec son vélo
BLEU reward	Adolescent saute la **pente** ‘slope' avec son vélo
Unsupervised reward	Adolescent saute la **colline** ‘hill' avec son vélo

RL FOR MT

- Unsupervised reward contributes to search space exploration and re-balancing
- It is beneficial when we have to choose between possible translations for an ambiguous word
- BLEU reward is more reliable when the goal is to produce one single possible translation