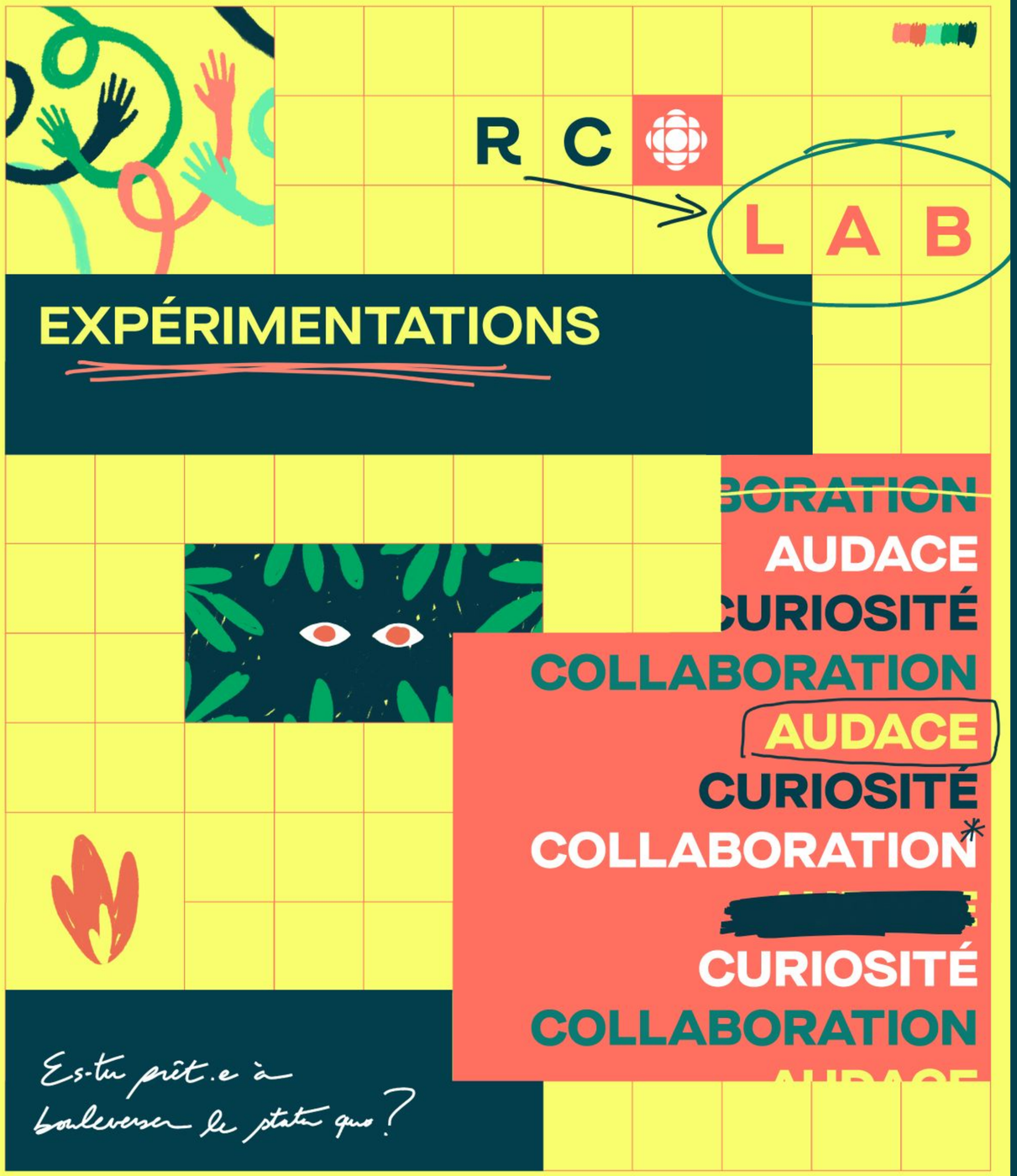


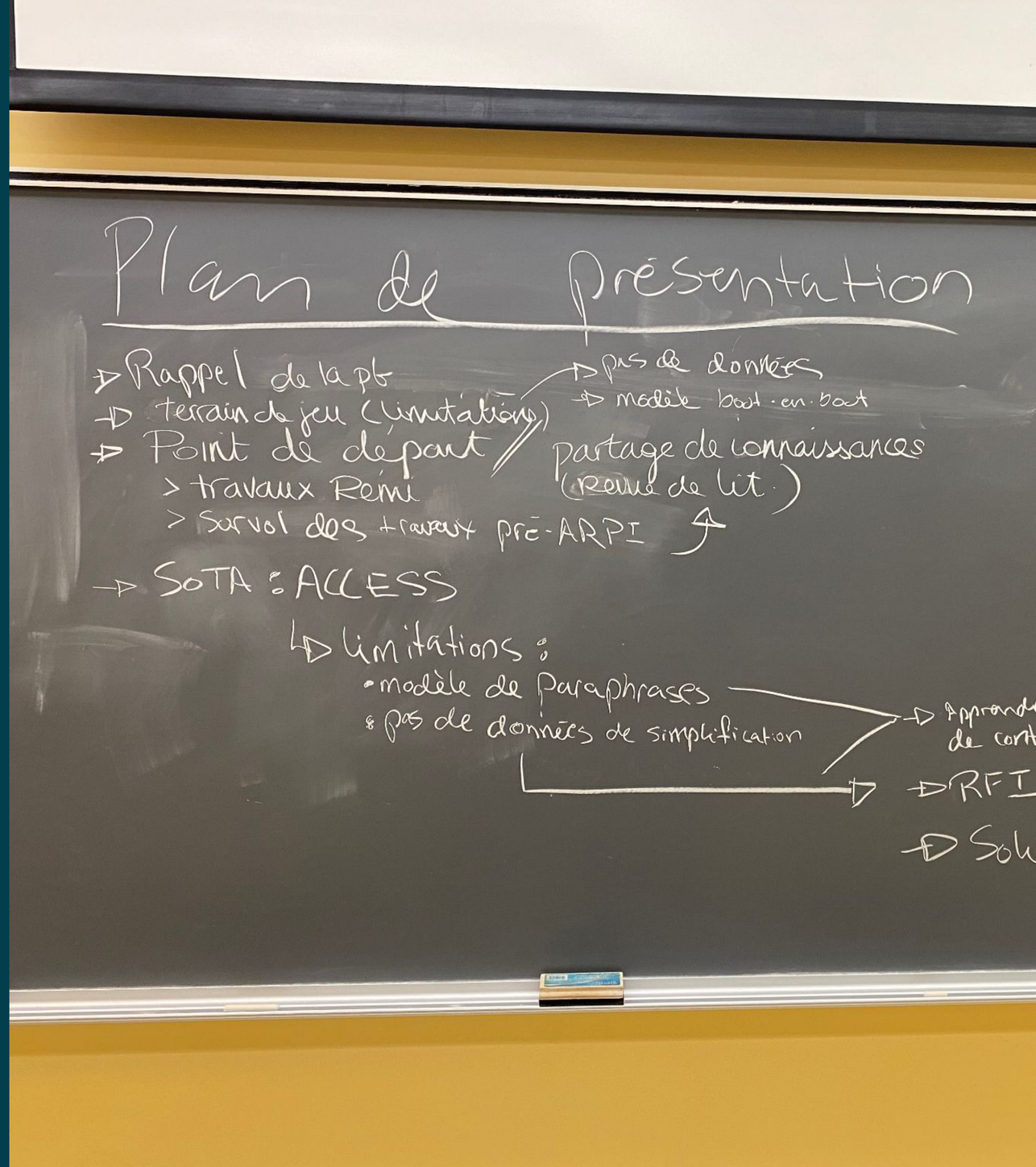
ARPI IVADO

Automatic Text Simplification

Radio-Canada
août 2022

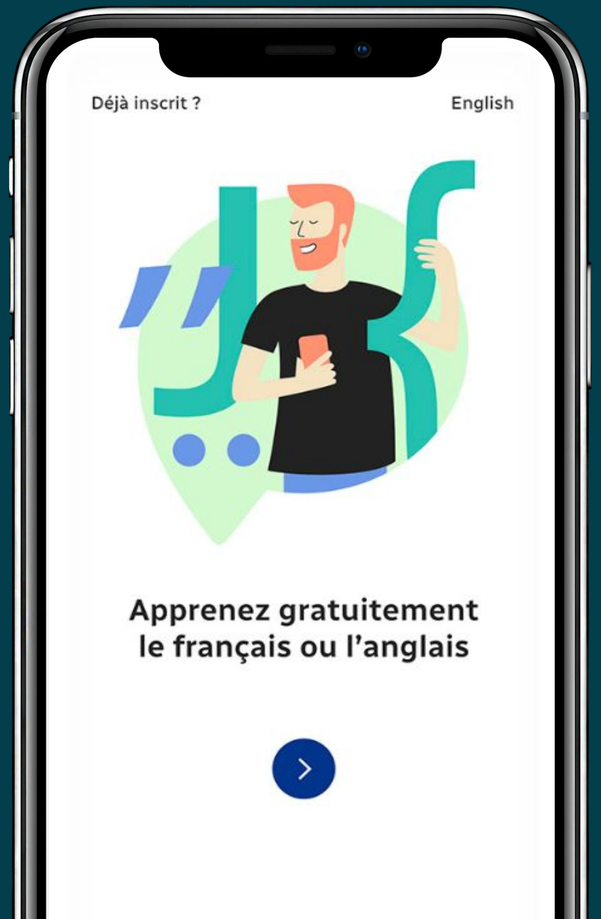
OUTLINE

1. Problem
2. Initial constraints
3. Knowledge transfer
4. Evaluation metrics
5. Datasets
6. Models
7. Other possibilities



Médias numériques

AUTOMATIC TEXT SIMPLIFICATION (ATS)



Offer a **new learning activity of reading comprehension** on the language learning platform Mauril (in French and English) using original content (articles) from CBC and Radio-Canada.

Automatic text simplification (ATS) is one of the considered solutions to help produce simpler texts (possibly) adapted to the competence level of the learner (French and English).

The team

UQÀM

UCLouvain

HEC MONTRÉAL

National Research Council Canada | Conseil national de recherches Canada

INITIAL CONSTRAINTS

No training datasets in French

High quality requirements

End-to-end model

KNOWLEDGE TRANSFER

CNRC

- Overview of SOTA and important questions
  - What is the exact need? Who is the simplification for?
  - How can we create the required data and resources for NLP/ML models?
  - How much time/effort and what sort of human expertise is needed?
  - What is the ground truth for “simplified” text? How do we ensure that?
  - Once a NLP based solution is ready, how does one do *extrinsic evaluation* with users/readers beyond the data driven evaluation?
- Automatic translation perspectives

UQAM

- Analysis of available evaluation metrics, models and datasets
- Construction of a new corpus of simple French text (not aligned to a complex equivalent)
  - Scraping the journal in simple French of RFI
  - Scraping the belgian journal of simple French L’Essentiel

KNOWLEDGE TRANSFER

University of Louvain

- State-of-the-art approach: MUSS (Martin et al., 2022)

- Evaluation of ATS systems remains an open question:.
  - Grammaticality and meaning preservation : BLEU
  - SARI somewhat uninformative
  - FKGL has been shown to be inadequate
  - Manual evaluation is needed, but no standard exists

- Lexical simplification is more common than syntactic simplification

- In ASSET, common test dataset: 50% of simplified sentences are the result of 3 transformations of fewer from the origin.

- Machine translation of WikiLarge dataset can provide competitive results.

EVALUATION METRICS

Automatic evaluation performed with the library EASSE

• Machine translation metrics:	
○ BLEU	P, G
○ METEOR	P, G
• Simplification metrics::	
○ SARI	P?, G?, S?
○ SAMSA	S
• Quality estimation::	
○ Sum-QE & Simple-QE	P, G, S
• Surface-level metrics:	
○ ISiM	S
○ Word, character or syllable counts	S
○ FKGL and Kandel-Moles indices	S
• Scores based on linguistic models	
○ Breadth and depth of syntax trees	S
○ Perplexity	G, S
○ BERTscore	P, G
○ Sentence embedding similarity	P

Legend
P: meaning preservation
G: grammaticality
S: simplicity
ajuster EASSE
To do
Rejected

EVALUATION METRICS

How to measure and understand how well information have been preserved in a simplified text

The Question Game

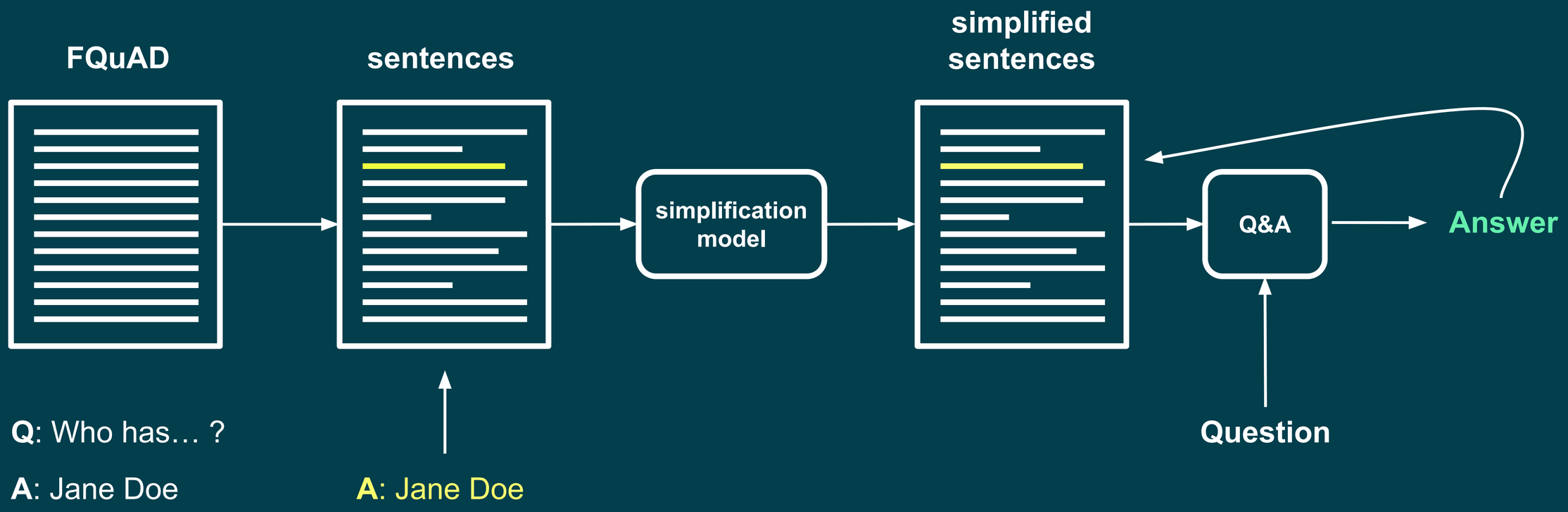
The Question Game consists of asking multiple-choice questions to users about a document content. Then the correctness of the answers is measured via different frameworks: if the readers have seen the initial corpus, if they have only read the generated texts, or if they have viewed both. It allows to understand how well the generated output replaces the most important facts conveyed by the input and how suited it is as an alternative source of information.

Principle applied in Text Summarization with automated Questions/Answering systems:
(Deutsch, D., Bedrax-Weiss, T., & Roth, D. (2021). Towards question-answering as an automatic metric for evaluating the content quality of a summary. Transactions of the Association for Computational Linguistics, 9, 774-789.)

Objective : Transpose the idea of automatic evaluation to text simplification and measure how well the system is able to answer question with the original and the simplified text

EVALUATION METRICS

FQuAD as a means to evaluate the **meaning preservation** aspect of sentence simplification



TRAINING CORPORA

Dataset	Lng	Size	Description
Wikilarge	Anglais	296K	Automatic alignment of English Wikipedia and Simple English Wikipedia
MLSUM	Français	425K	Multilingual summaries of news
RFI	Français	105K	Transcripts of a news feed in simple French

EVALUATION CORPORA

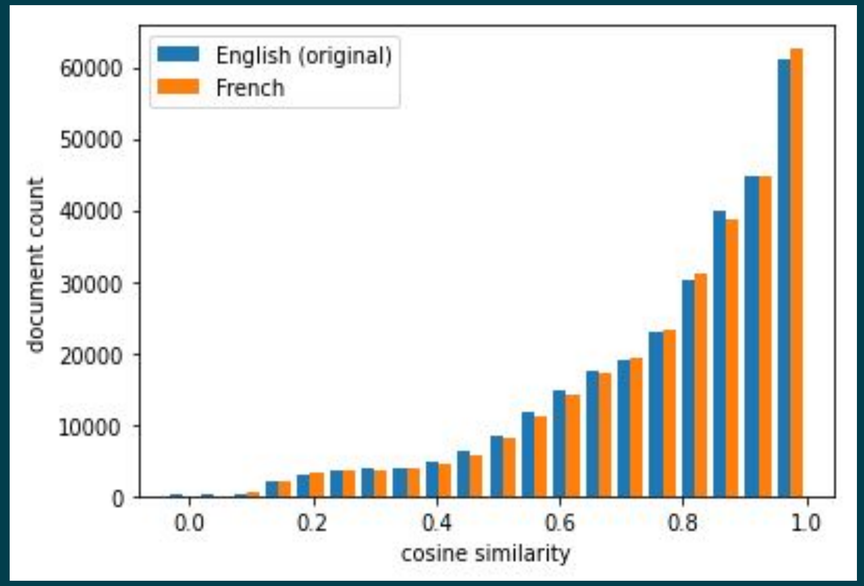
Dataset	Lng	Refs	Size	Translation	Description
Alector	FR	1	79 documents		79 texts simplified by experts
ASSET	EN	10	2359 phrases	FR	Wikilarge (validation set) + MechanicalTurk simplifications
TurkCorpus	EN	8	2359 phrases	FR	Wikilarge (validation set) + MechanicalTurk simplifications
OneStop English	EN	1	2166 phrases	FR	Aligned sentence pairs (The Guardian) from beginner, intermediate and advance levels

AUTOMATIC TRANSLATION

How to evaluate the impact of machine translated datasets?

- Should preserve the “difference” in complexity
- Complexity is difficult to quantify
- Verify whether some quantities related to complexity are preserved:
  - Flesch Reading Ease, Lexical richness (Rauf et al., 2020)
  - Semantic similarity (LASER), Edit distance, SARI

Wikilarge



CORPUS ALIGNMENT

ALECTOR

	42 fiction/stories	*849 aligned sentences*
	37 non-fiction	*595 aligned sentences*
TOTAL	**79 texts**	**1444 aligned sentences**

FICTION

COMPLEX SENTENCE L'oiseau quitta alors l'oreille **et** retourna dans son nid**, auprès de** ses enfants.

SIMPLE SENTENCE L'oiseau quitta alors l'oreille**. Il** retourna dans son nid **avec** ses enfants.

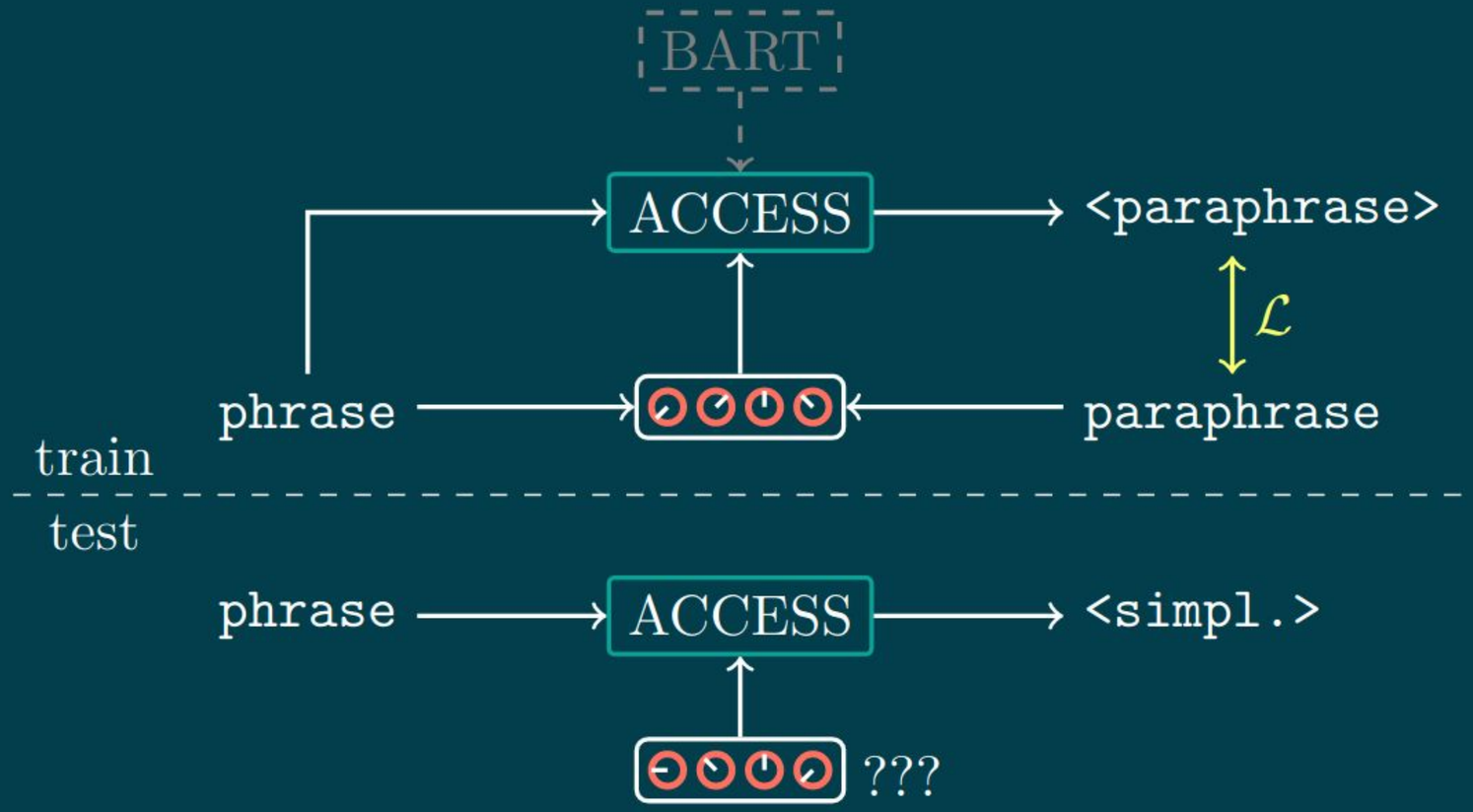
NON-FICTION

COMPLEX SENTENCE Le tonnerre est le bruit sourd qui **accompagne** la foudre.

SIMPLE SENTENCE Le tonnerre est le bruit sourd qui **vient avec** la foudre.

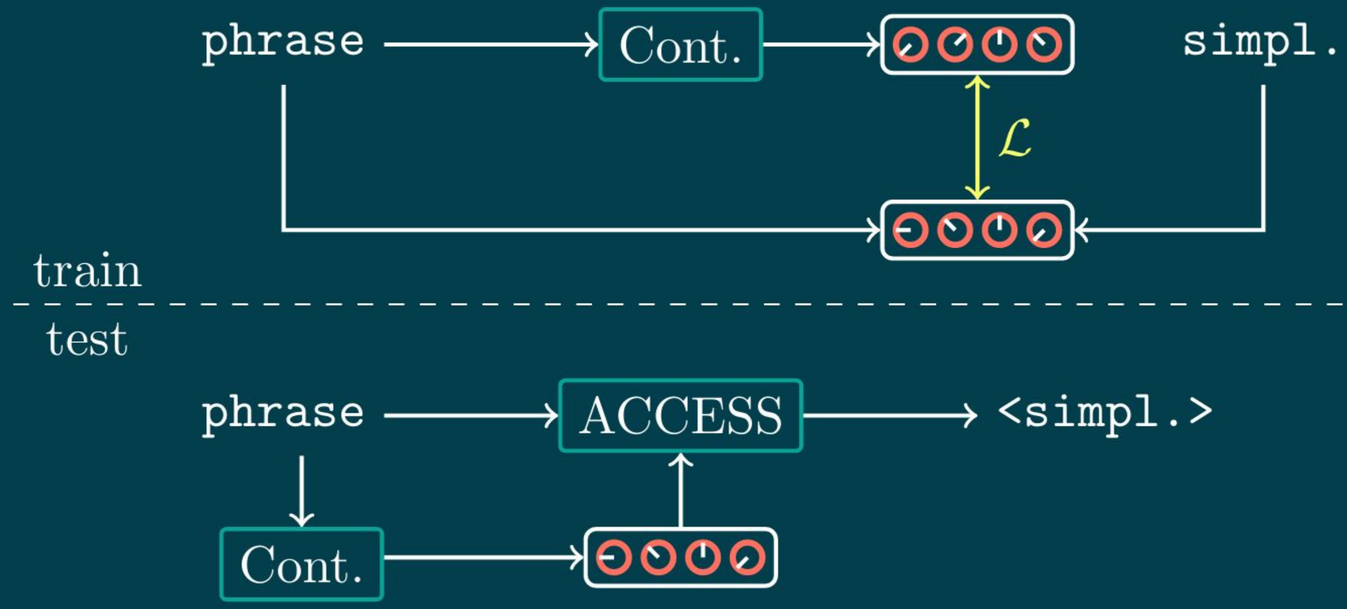
ANALYSED SOLUTION: SEQUENCE-TO-SEQUENCE MODEL

MUSS (Martin et al., 2021)



PROPOSED SOLUTION : SEQUENCE -TO-SEQUENCE MODEL

- MUSS is trained to **paraphrase**, not **simplify**
- Control values for simplification : inferred at corpus level from validation
- Idea : learn to produce control values as a function of input



OTHER POSSIBILITIES

Denoising autoencoders:

- Applied successfully to multiple task :
  - Grammar Correction : Wan, Z., et al. (2020). Improving grammatical error correction with data augmentation by editing latent representation.
  - Text Summarization : Fevry, T., & Phang, J. (2018). Unsupervised sentence compression using denoising auto-encoders.
  - Machine translation : Liu, Y., et al. (2020). Multilingual denoising pre-training for neural machine translation.
- Objective: add noisy signal to input, autoencoder must recreate original input
- Model learns to drop noise and becomes more robust

OTHER POSSIBILITIES

Application for text simplification :

- By studying some standardized dataset, we know what types of operation are performed, how often, and how many tokens they concerned
  - 50% of sentences undergo 1 to 3 modifications
  - 14% of modifications are synonym replacement
  - etc.
- We can consider complexity as non desirable noise/information.
  - For any sentence S, we can add controlled noise. Specifically, we can complexify that sentence using some common transformation (adding adjective, using a rarer synonym, add dependent clauses, etc.)
- We thus create a noisy sentence that can be considered as complex and can serve as input for an sequence-to-sequence model

OTHER POSSIBILITIES

Application for text simplification :

- During training :
  - Input = Noisy complex generated sentence
  - Reference = Original sentence that can be seen as the simple sentence output (We actually don’t know the simplicity/complexity of that sentence, we just know that it is more simple than the input and is grammatical)
- During generation :
  - Input = Original sentence in the test sets
  - Output = The model has learned to generate coherent sentences by reducing noise from the input. We expect the output to be a sentence where the model delete unnecessary adjectives, change synonyms, delete dependent clauses, etc.

RC LAB

MERCI

QUESTIONS