



# Machine Reading Comprehension

## Multiple Choice Question Answering

Final Mini-Project - NUANS

2022/2023

# Problem

## Story - Context

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

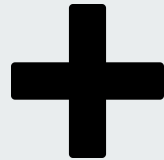
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

## Question

INPUT



What is the name of the trouble making turtle?

- A) Fries
- B) Pudding
- C) James
- D) Jane

Multiple  
Option  
Answers



MODEL

C) James

Right Answer

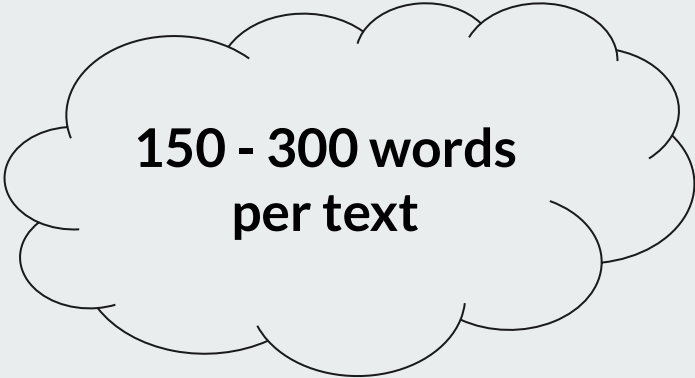
OUTPUT

# Dataset



## MC-500:

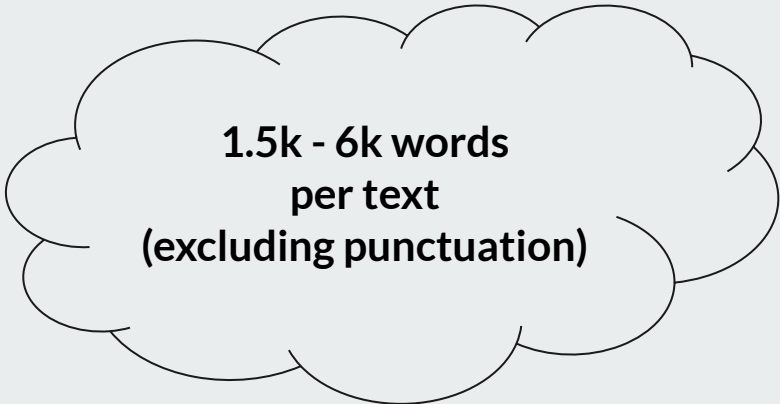
- 500 texts
- 2000 questions
- 4 option of choice



**150 - 300 words  
per text**

## QuALITY!:

- 4668 questions
- 4 option of choice



**1.5k - 6k words  
per text  
(excluding punctuation)**

# Issue



The length of the histories is too large respect the input of the natural language process algorithm, and all information in entire text are useless to answer a single precise question.

# Solution



**Extraction of contexts:**

- 1, 3, 5, 10, 15 rows

**Metrics:** Cosine Similarity, ROUGE values, BM 25

**Embeddings:**

- GloVe
- FastText
- TF-IDF
- DPR

## EXTRACTION CONTEXT

TO EXTRACT THE CONTEXT OF THE TEXTS FOR EACH QUESTION USING A DIFFERENT METRICS AND EMBEDDINGS TO CALCULATE THE SIMILARITY AND THE CORRELATION BETWEEN THE QUESTION AND THE TEXT.

FOR EACH TYPE WE PRODUCE A CSV FILES (1, 3, 5, 10, 15 rows of context) WHERE THERE ARE CONTEXT, QUESTION, ANSWER OPTIONS A, B, C, D, RIGHT ANSWER and LABEL.



## *Model - Architecture*



### PRE-PROCESSED INPUT

For each question:

CONTEXT + QUESTION + OPTION A

CONTEXT + QUESTION + OPTION B

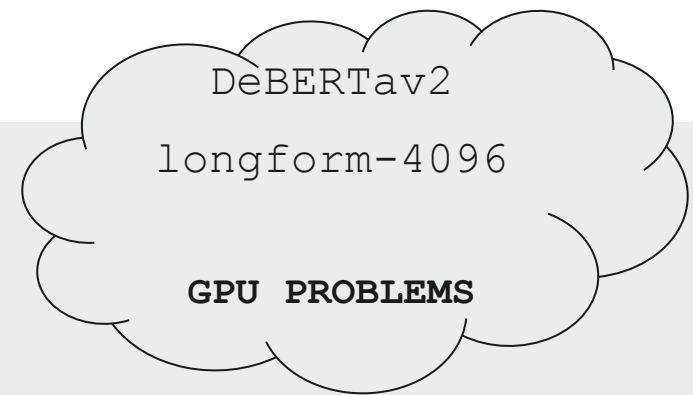
CONTEXT + QUESTION + OPTION C

CONTEXT + QUESTION + OPTION D



LABEL

right option



## BERT

- bidirectional transformer pretrained with a combination of masked language modeling objective and next sentence prediction,
- learns an inner representation of the language that can then be used to extract features useful for downstream tasks.

## RoBERT

- modifies key hyperparameters and training with much larger mini-batches and learning rates,
- same architecture as BERT, but uses a byte-level BPE as a tokenizer (same as GPT-2) and uses a different pretraining scheme.

## ALBERT

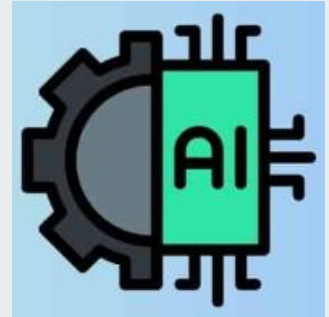
- presents two parameter-reduction techniques to lower memory consumption (e.g layers are split in groups that share parameters) and increase the training speed of BERT,
- computational cost remains similar to a BERT-like architecture.

## Train

TRAINING PHASE DONE ON ALL THE EXTRACTED CONTEXT FIRST ON MCTEST-500 DATASET AND THEN ON QUALITY DATASET.

BERT and RoBERTa using with MCTEST-500 DATASET

ALBERT using with QUALITY DATASET



## Evaluation Metrics

ACCURACY

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Multiple Choice Question Answering Problem: only one answer option is the right one



# Results

## TRAINING PHASE ACCURACY MCTEST-500 ROBERTA

Best results for the longer contexts.

		1	3	5	10	15
CONTEXT EMBEDDINGS	BM 25	55%	58%	55%	63%	72%
	ROUGE	56%	58%	61%	68%	71%
	GloVe	41%	49%	54%	58%	70%
	FastText	68%	57%	57%	66%	70 %
	TF-IDF	54%	56%	62%	68%	69%
	DPR	51%	62%	67%	70%	70 %

		NUMBER OF SENTENCES				
		1	3	5	10	15
CONTEXT EMBEDDINGS	BM 25	38%	37%	38%	36%	33%
	ROUGE	35%	39%	38%	30%	35%
	GloVe	36%	37%	37%	35%	38%
	FastText	33%	35%	27%	37%	38%
	TF-IDF	37%	38%	37%	38%	37%

## TRAINING PHASE ACCURACY QuALITY ALBERT

Best results are observed not only for longer contexts.

## TESTING PHASE

During the testing phase, using the previously computed weights (in yellow), we assessed the accuracy for all the extracted contexts.

In Table there are the best values of accuracy reach during the testing.

All the results can be observed by the following [link](#)

MCTest-500			
Embedding	Weight	Context	Accuracy
DPR	DPR	15	0.74
ROUGE		15	0.73
DPR	FastText	15	0.73
QuALITY			
Embedding	Weight	Context	Accuracy
all	ROUGE	15	0.40
BM25		10	0.40
GloVe		10	0.40
BM25	TF-IDF	1	0.41
GloVe		1	0.41
TF-IDF		3	0.41
BM25	GloVe	15	0.41
GloVe		15	0.41
FastText		15	0.40
TF-IDF		15	0.40
BM25	BM25	15	0.40
GloVe		15	0.40
Rouge		10/15	0.40



**Thank you!**

