Hello!

Myself FARHAN, I am a Ph.D. Student

NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES (NUCES-FAST)

"FAST-MT PARTICIPATION FOR THE JOKER CLEF-2022 AUTOMATIC PUN AND HUMOUR TRANSLATION TASKS"

1. Classify and explain instances of wordplay

Given a Wordplay along with its Id, predict the values of the following columns

- *Location*
- *Interpretation*
- *Horizontal/Vertical*
- *Conventional Form*
- *Manipulation Type*
- *Manipulation Level*
- *Cultural Reference*
- *Offensive*

Locate the words forming wordplay via Token Classification

- We have classified each token of the given wordplay into the following types.
- ***word_play_token_B***:
  - To identify the word which begins the wordplay.
- ***word_play_token_I:***
  - To identify the other remaining words in the wordplay.
- ***Other_token***:
  - To identify all the words which don't belong to the wordplay.

Example of Token classification

Example:

- **English Text:** *Follow your knows.*
- **Processed Tokenized Text:** **[***Follow, your, knows***]**
- **Expected Output:** ***[****Other_token, Other_token, word_play_token_B***]**

Models for Token classification

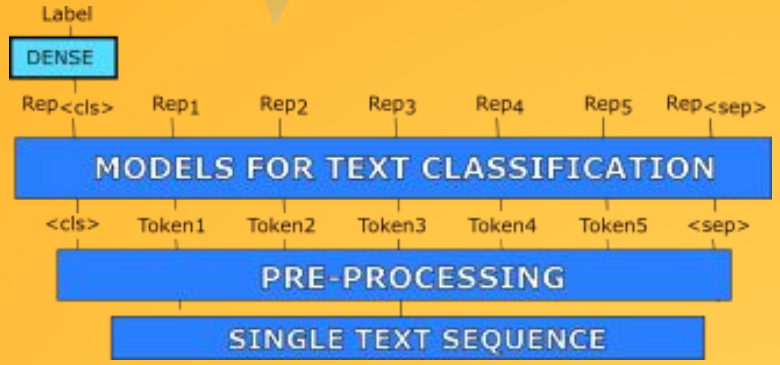
Models used for token classification

- Pre-trained BERT BASE
- KEY BERT: with two embeders
  - With fine-tuned BERT BASE
  - With pre-trained all-MiniLM-L6-v2

Example of Text classification

Example:

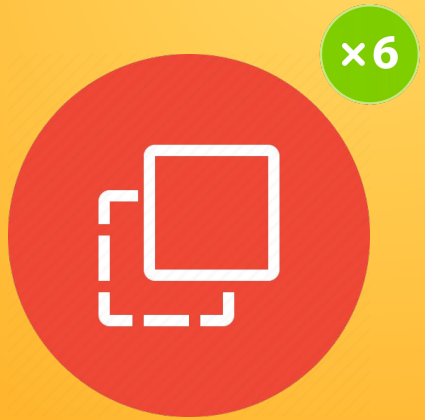
- **English Text:** *Follow your knows.*
- **Processed Tokenized Text: [***Follow, your, knows]*
- **Expected Output:**
  1. **Horizontal/Vertical:** “vertical”
  2. **Manipulation type:** “similarity”
  3. **Manipulation level:** “sound”
  4. **Cultural reference:** “false”
  5. **Conventional form:** “false”
  6. **Offensive:** “none”



Models for Text classification

Models used for text classification

- Pre-trained DistilBERT
- Make its 6 copies and fine tune them for each categorical target column.



Models for Text Generation

We have fine tuned GPT-2 to generate interpretation for a given wordplay

Example:1

Model Input:

- “follow your knows”

Expected Output:

- (knows/nose)

Example:2

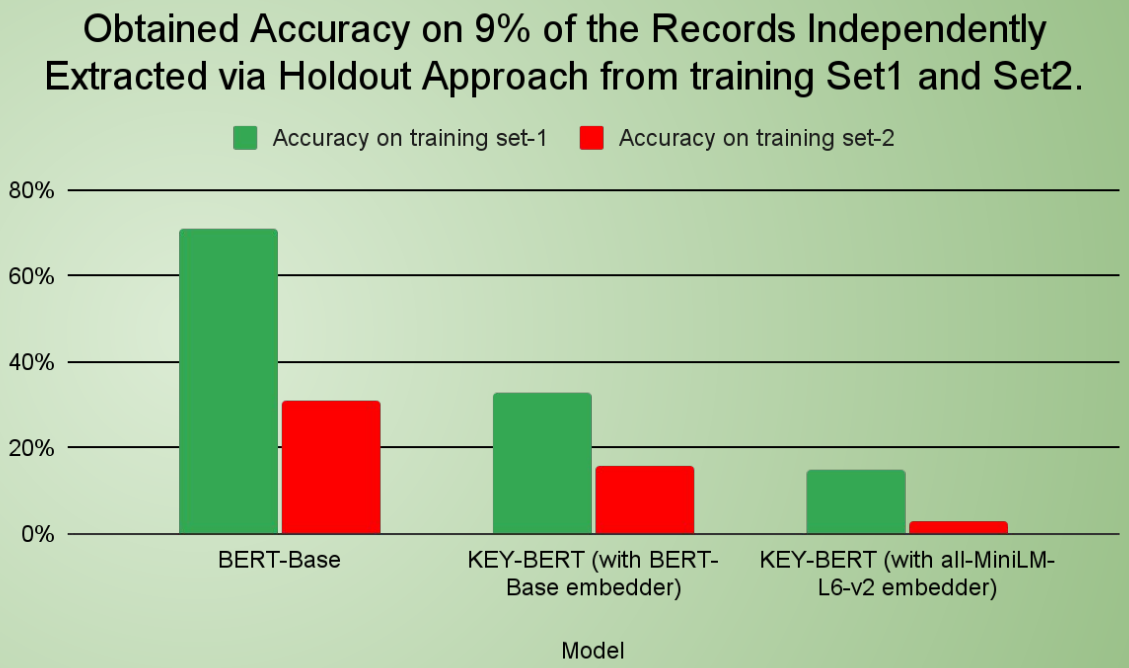
Model Input:

- “in the dark follow the son”

Expected Output:

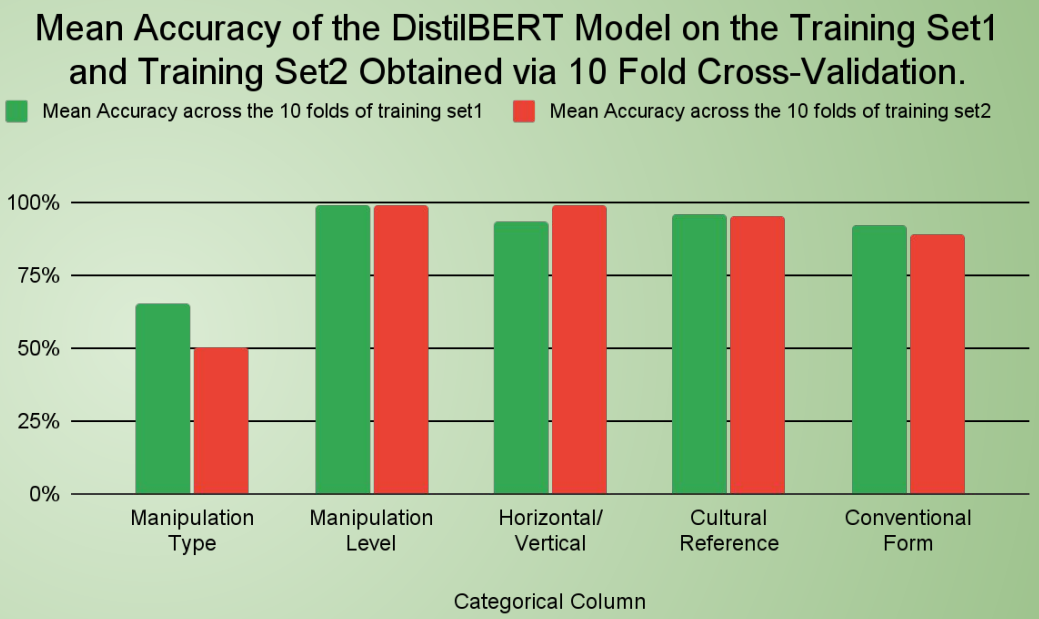
- (son/sun)

Performance of the Models



BERT-Base model provides the best performance to locate the wordplay in the given English text via token classification.

Performance of the Models



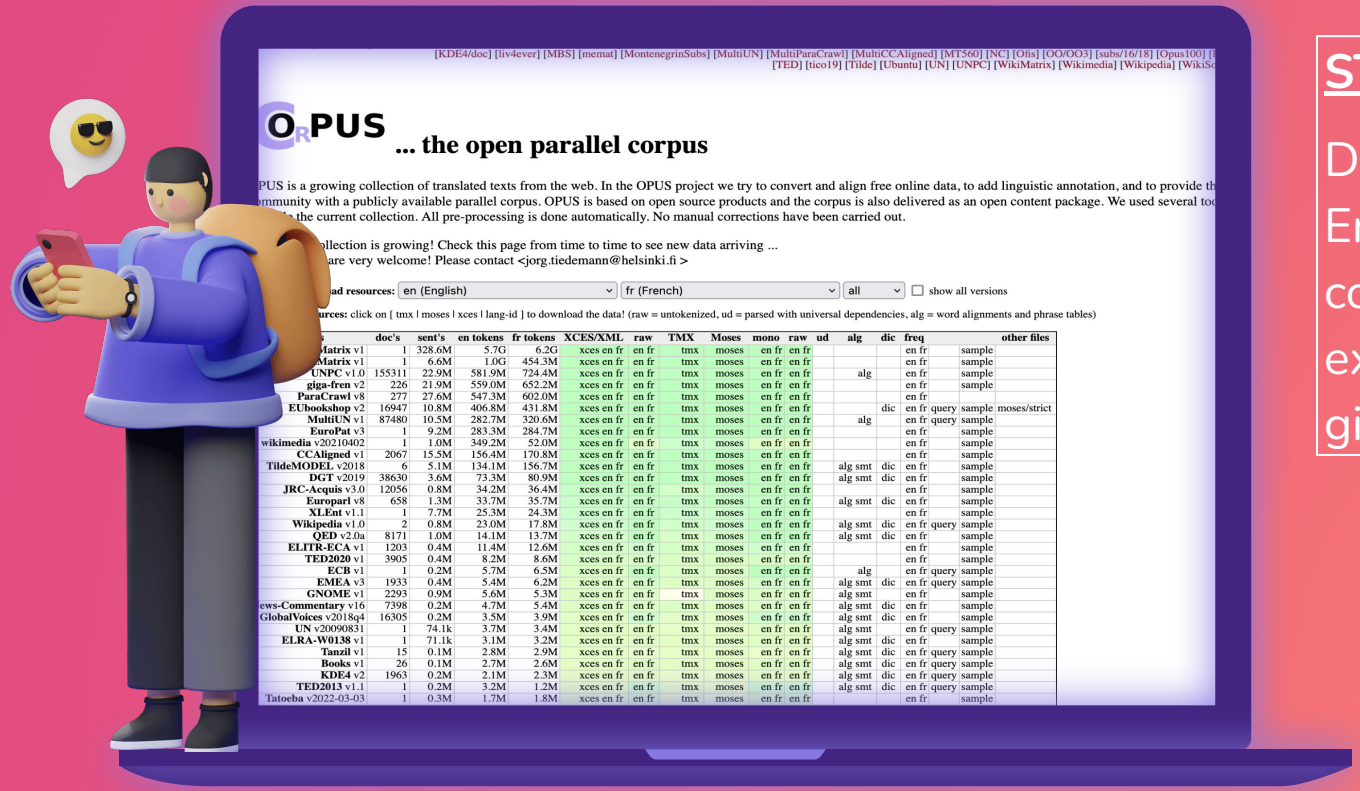
Overall the DistilBERT model has performed better on the training set1 for predicting labels of categorical columns via text classification.

2. Translate single words (nouns) containing wordplay

Given an English noun, generate its corresponding French translation.

COLUMN	EXPLANATION
Id	A unique identifier for the given English noun.
En	An English noun containing a wordplay.
Fr	French translation for the English noun.

Transforming the problem of translation of single English nouns to French into Extractive Question/Answering



	doc's	sent's	en tokens	fr tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq		other files
Matrix v1	1	328.6M	5.7G	6.2G	xces en fr	en fr	tmx	moses	en fr	en fr				en fr	sample	
Matrix v1	1	6.6M	1.0G	454.3M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr	sample	
UNPC v1.0	155311	22.9M	581.9M	724.4M	xces en fr	en fr	tmx	moses	en fr	en fr		alg		en fr	sample	
giga-fren v2	226	21.9M	559.0M	652.2M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr	sample	
ParaCrawl v8	277	27.6M	547.3M	602.0M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		
EUbookshop v2	16947	10.8M	406.8M	431.8M	xces en fr	en fr	tmx	moses	en fr	en fr			dic	en fr	query	sample	moses/strict
MultiUN v1	87480	10.5M	282.7M	320.6M	xces en fr	en fr	tmx	moses	en fr	en fr		alg		en fr	query	sample	
EuroPat v3	1	9.2M	283.3M	284.7M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
wikimedia v20210402	1	1.0M	349.2M	52.0M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
CCAligned v1	2067	15.5M	156.4M	170.8M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
TildeMODEL v2018	6	5.1M	134.1M	156.7M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
DGT v2019	38630	3.6M	73.3M	80.9M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
JRC-Acquis v3.0	12056	0.8M	34.2M	36.4M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
Europarl v8	658	1.3M	33.7M	35.7M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
XLEnt v1.1	1	7.7M	25.3M	24.3M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
Wikipedia v1.0	2	0.8M	23.0M	17.8M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
QED v2.0a	8171	1.0M	14.1M	13.7M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
ELITR-ECA v1	1203	0.4M	11.4M	12.6M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
TED2020 v1	3905	0.4M	8.2M	8.6M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	
ECB v1	1	0.2M	5.7M	6.5M	xces en fr	en fr	tmx	moses	en fr	en fr		alg		en fr	query	sample	
EMEA v3	1933	0.4M	5.4M	6.2M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
GNOME v1	2293	0.9M	5.6M	5.3M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt		en fr		sample	
ews-Commentary v16	7398	0.2M	4.7M	5.4M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
GlobalVoices v2018q4	16305	0.2M	3.5M	3.9M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
UN v20090831	1	74.1k	3.7M	3.4M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt		en fr	query	sample	
ELRA-W0138 v1	1	71.1k	3.1M	3.2M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr		sample	
Tanzil v1	15	0.1M	2.8M	2.9M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
Books v1	26	0.1M	2.7M	2.6M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
KDE4 v2	1963	0.2M	2.1M	2.3M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
TED2013 v1.1	1	0.2M	3.2M	1.2M	xces en fr	en fr	tmx	moses	en fr	en fr		alg smt	dic	en fr	query	sample	
Tatoeba v2022-03-03	1	0.3M	1.7M	1.8M	xces en fr	en fr	tmx	moses	en fr	en fr				en fr		sample	

STEP-1

Download the English/French Parallel corpus from OPUS to extract contexts for the given English nouns

Transforming the problem of translation of single English nouns to French into Extractive Question/Answering

English/French nouns of the training set

id	en	fr
noun_1	Obelix	Obélix

English/French extracted sentence pairs from the OPUS parallel corpus

en	fr
asterix and obelix should stay in the village and not go in the forest	astérix et obélix ne devraient plus quitter le village.
asterix and obelix are cartoon characters.	astérix et obélix sont des personnages de dessins animés.
...	...
asterix and obelix are funny.	astérix et obélix sont drôles.

STEP-1

Download the English/French Parallel corpus from OPUS to extract contexts for the given English nouns

STEP-2

For each English/French nouns from the data set. Extract those English/French parallel sentences that contains the selected English/French nouns.

Transforming the problem of translation of single English nouns to French into Extractive Question/Answering

STEP-1

Download the English/French Parallel corpus from OPUS to extract contexts for the given English nouns

STEP-2

For each English/French nouns from the data set. Extract those English/ French parallel sentences that contains the selected English/French nouns.

STEP-3

Transform the extracted English/French parallel sentence pairs for each English/French nouns of the data set into extractive Q/A styled data

id	question	Context	Answers
1	Obelix	astérix et obélix ne devraient plus quitter le village.	{"text": [Obélix], "answer_start": [11]}
2	Obelix	astérix et obélix sont des personnages de dessins animés.	{"text": [Obélix], "answer_start": [11]}
...	...	...	...

Transformed English/French nouns of the training data set

Transforming the problem of translation of single English nouns to French into Extractive Question/Answering

Test set

id	en	fr
noun_1	Obelix	predict
nount_2	lompaland	perdict

STEP-1

Download the English/French Parallel corpus from OPUS to extract contexts for the given English nouns

STEP-2

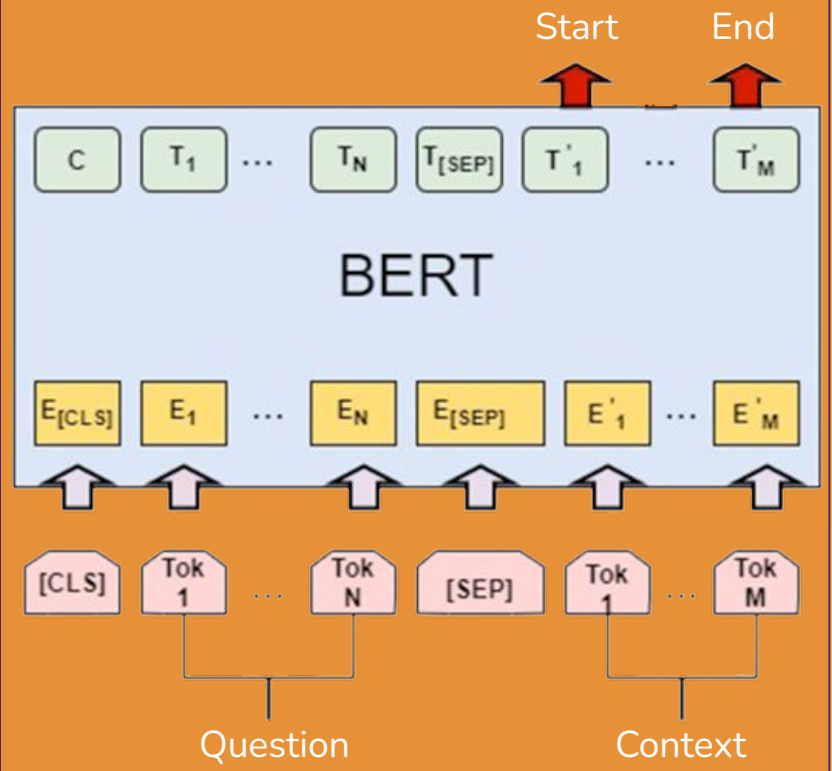
For each English/French nouns from the data set. Extract those English/ French parallel sentences that contains the selected English/French nouns.

STEP-3

Transform the extracted English/French parallel sentence pairs for each English/French nouns of the data set into extractive Q/A styled data

id	question	Context
1	Obelix	astérix et obélix ne devraient plus quitter le village.
2	lompaland	j'étais venu à lumpaland pour chercher de nouvelles saveurs.
...	...	...

Transformed English/French nouns of the test data set

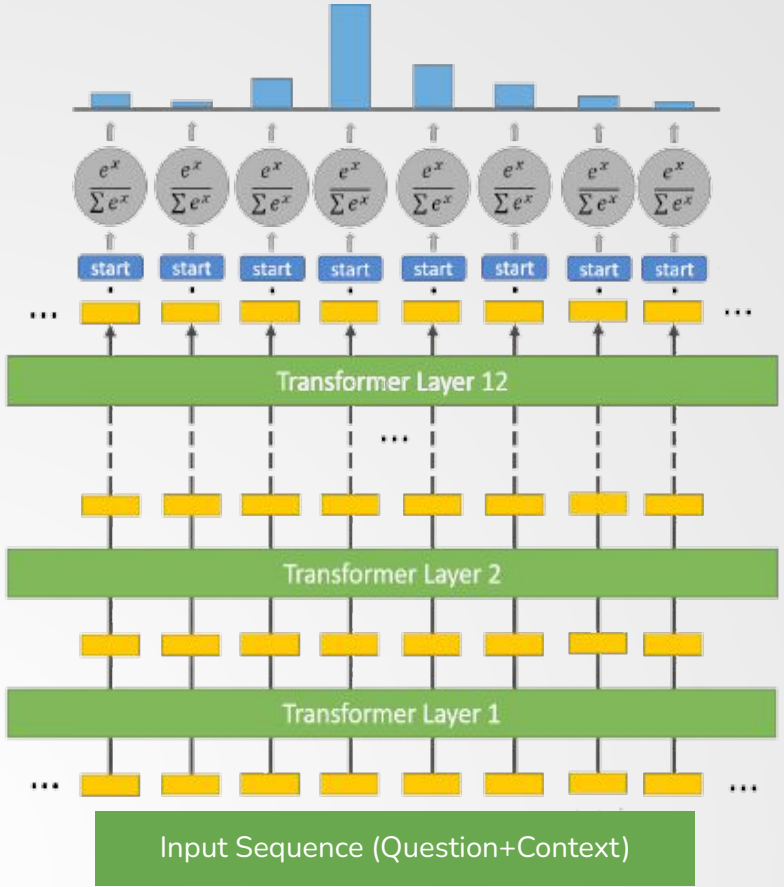


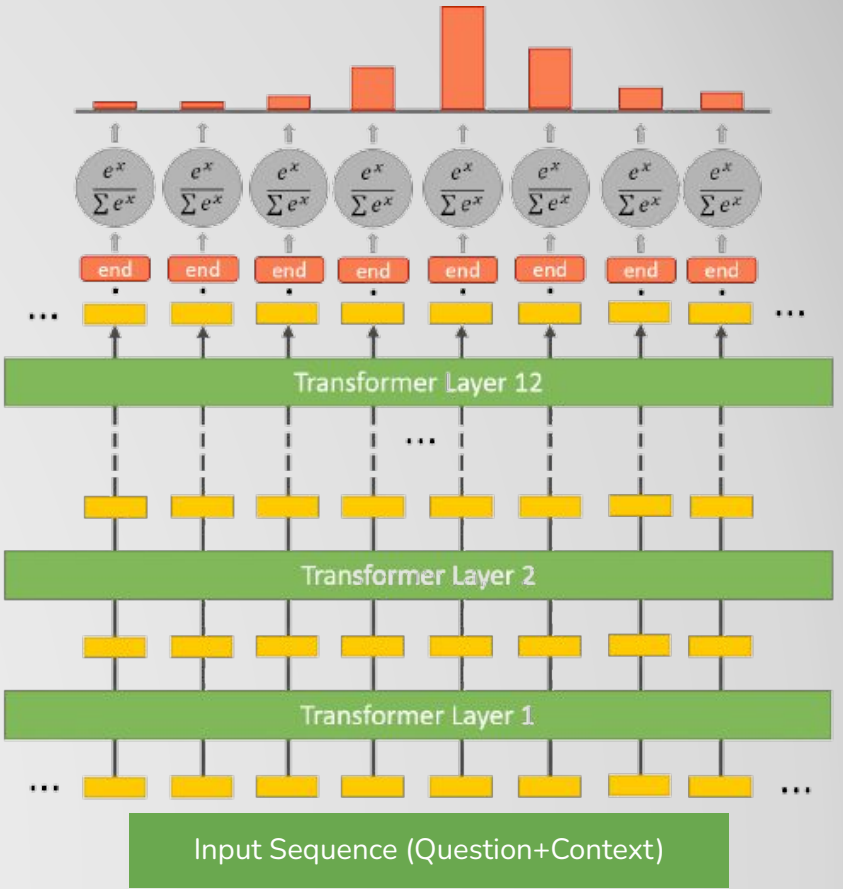
TASK-2 Extractive Q/A Models

Showcasing high level working of Extractive Q/A models

- *Question: Obelix (length of 1 token)*
- *Context: Astérix et obélix ne devraient plus quitter le village.*
- *Expected: start_value=11*
- *Expected: end_value=17*

Predicting start/end of Translation





Models for Extractive Q/A

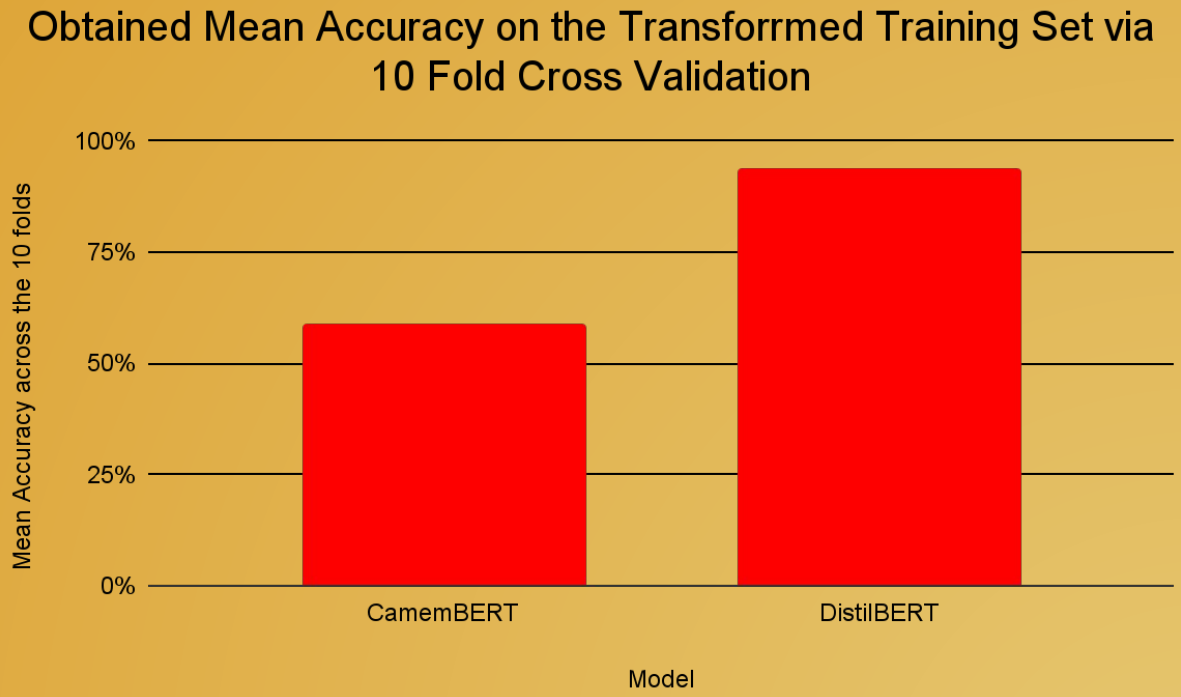
Transformed English/French nouns of the training data set

id	question	Context	Answers
1	Obelix	astérix et obélix ne devraient plus quitter le village.	{"text": [Obélix], "answer_start": [11]}
2	Obelix	astérix et obélix sont des personnages de dessins animés.	{"text": [Obélix], "answer_start": [11]}
...	...	...	...

Models used for Extractive Q/A

- DistillBERT
  - English Query and English context
- CamemBERT
  - French Query and French context

Performance of the Models



In our experiment, DistilBERT Model has given the best results.

3. Translate entire English phrases containing wordplay

Given an English sentence, generate its corresponding French version.

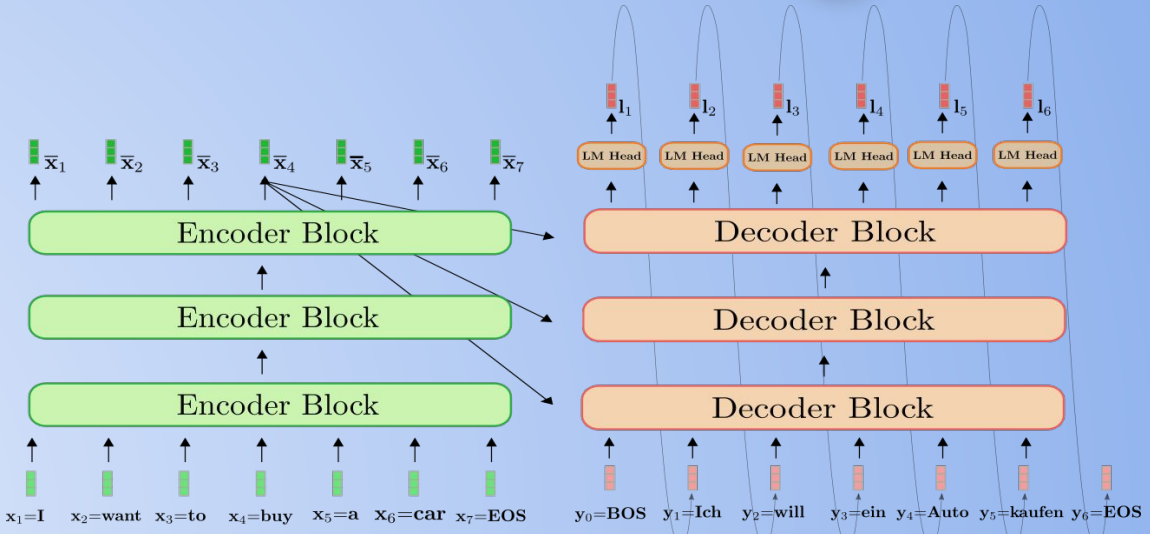
COLUMN	EXPLANATION
Id	A unique identifier for the English phrase.
En	An English phrase containing a wordplay.
Fr	French translation for the English phrase.

Transforming tabular data into a JSON dictionary with a total of 1185 keys, for further processing.

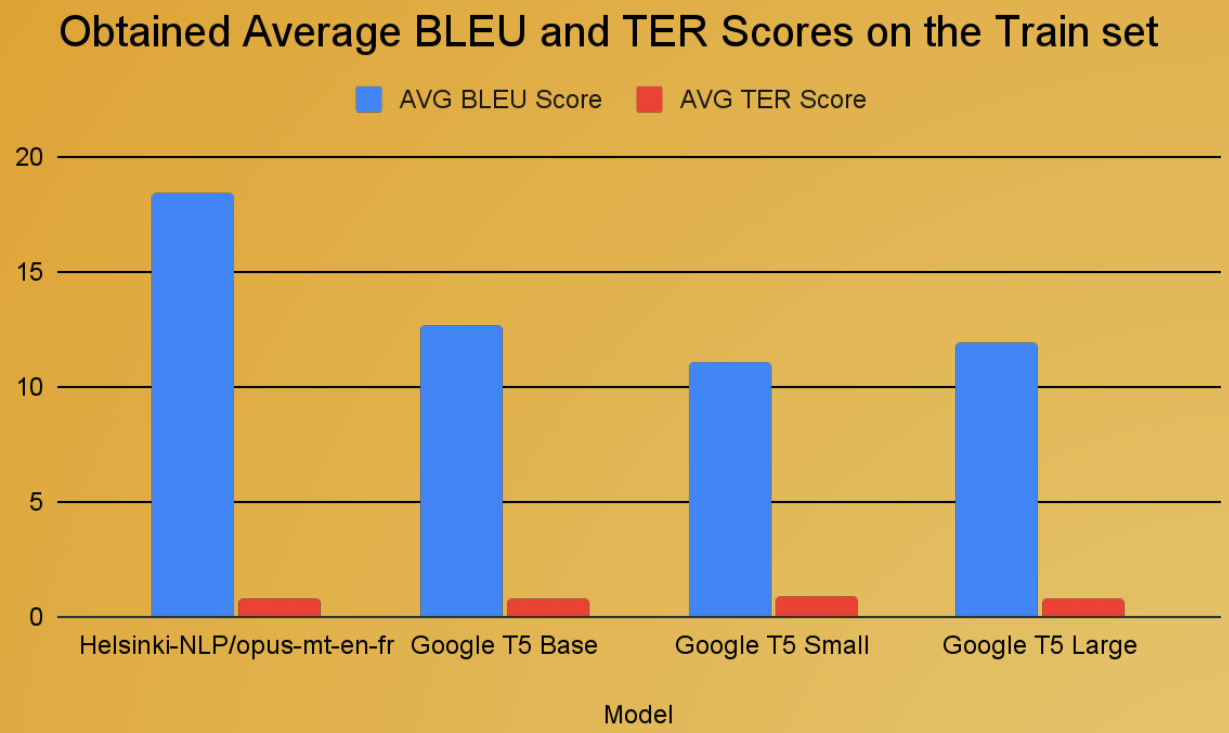
```
{
        " Tom said piously " :
        [
                " declara Tom pi - eusement . " ,
                " dit Tom pieusement." ,
                " Tom dit pieusement."
        ],
        …
}
```

TASK-3: Sequence to Sequence Models

1. Helsinki-NLP/opus-mt-en-fr
2. Google T5 Base
3. Google T5 Small
4. Google T5 Large

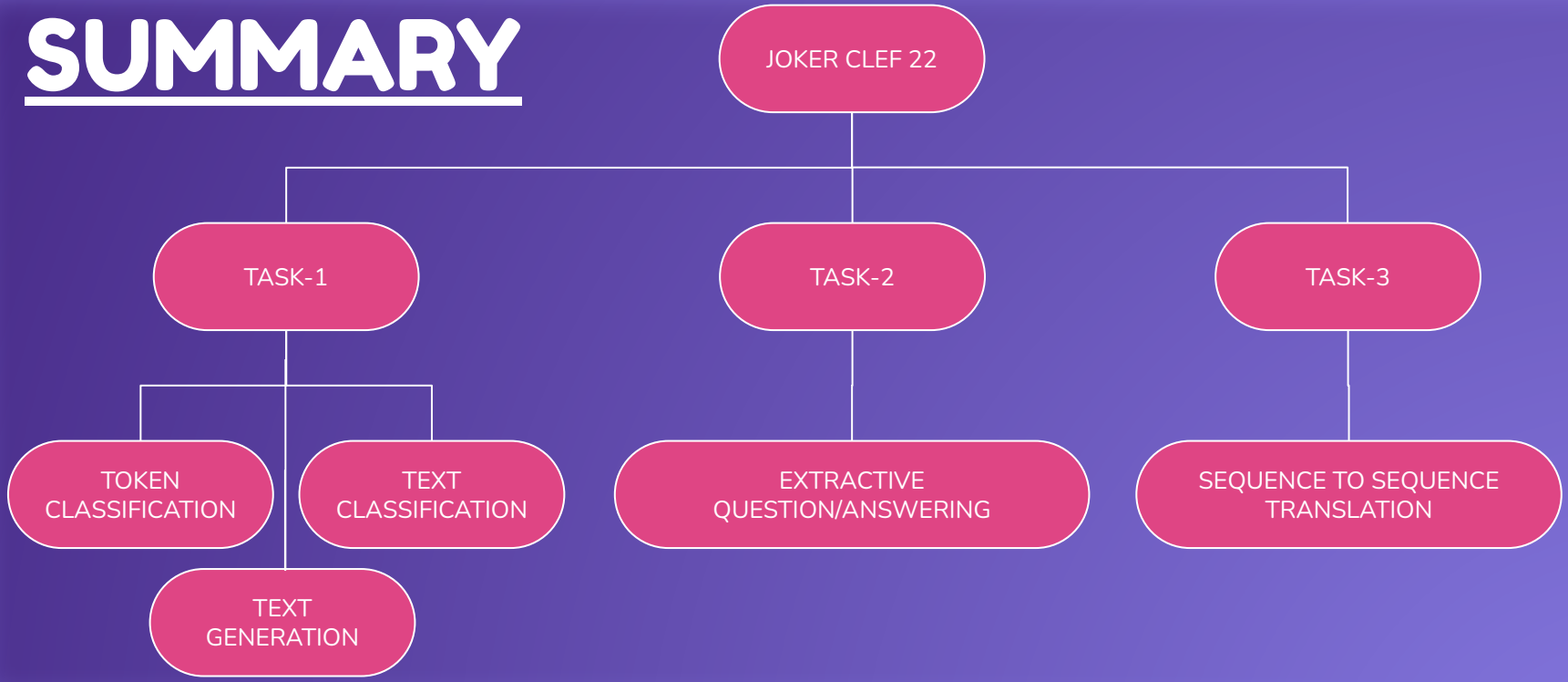


Performance of the Models

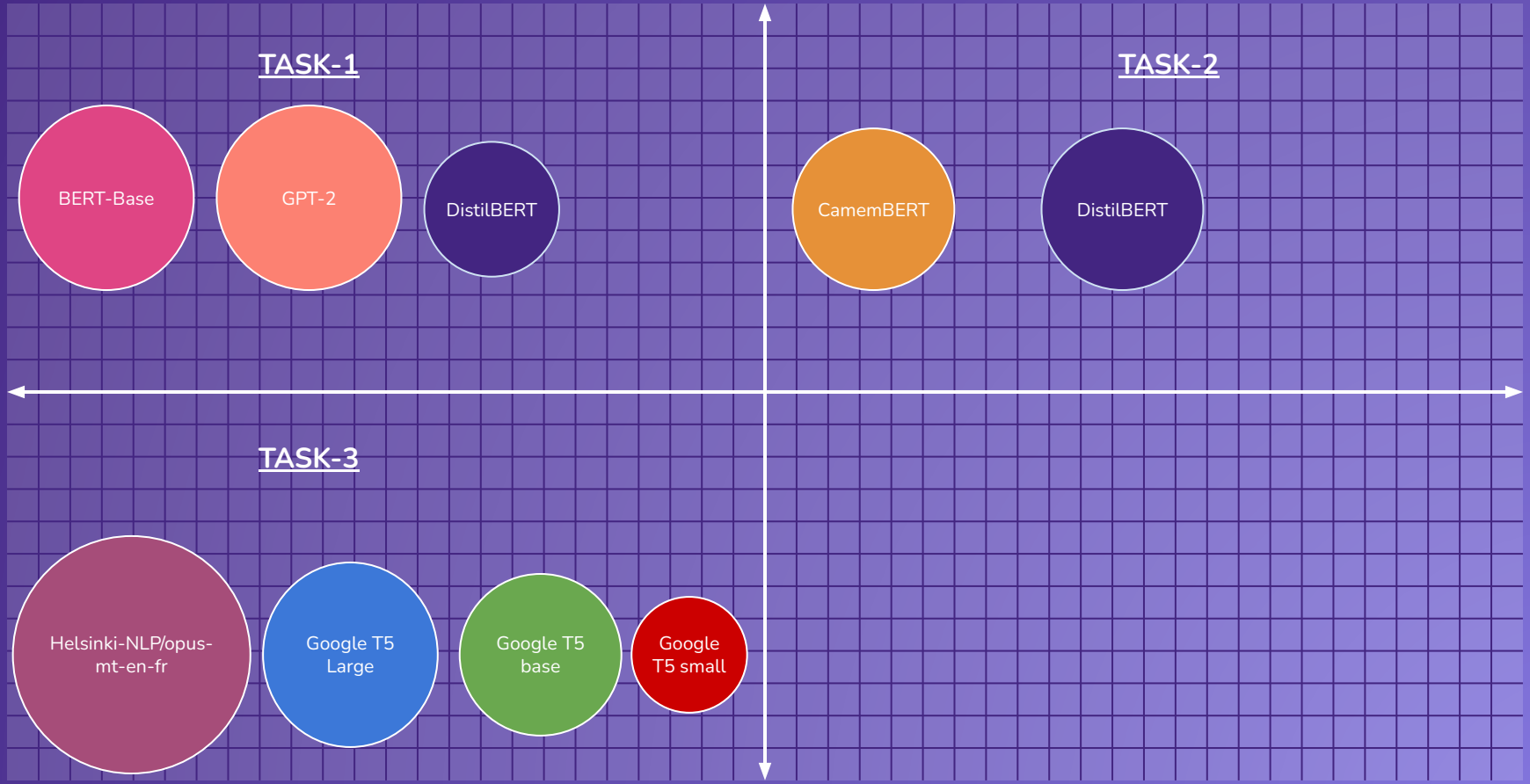


In our experiment, Helsinki-NLP Model has given the best results.

SUMMARY


JOKER CLEF 22
TASK-1
TASK-2
TASK-3
TOKEN CLASSIFICATION
TEXT CLASSIFICATION
TEXT GENERATION
EXTRACTIVE QUESTION/ANSWERING
SEQUENCE TO SEQUENCE TRANSLATION

SUMMARY


TASK-1
BERT-Base
GPT-2
DistilBERT
TASK-2
CamemBERT
DistilBERT
TASK-3
Helsinki-NLP/opus-mt-en-fr
Google T5 Large
Google T5 base
Google T5 small

Thanks!

Dr. Muhammad Rafi
JOB TITLE
Professor and Coordinator
FAST NUCES

Dr. Muhammad Atif Tahir
JOB TITLE
Professor and Director
FAST NUCES

Liana Ermakova
SPECIAL THANKS TO
JOKER 22 TEAM

Any questions?

You can find me at k214808@nu.edu.pk