FAST-MT Participation for the JOKER CLEF-2022 Automatic Pun and Humour Translation Tasks

Farhan Dhanani[1], Muhammad Rafi[2] and Muhammad Atif Tahir[3]

National University of Computer and Emerging Sciences (NUCES-FAST), Karachi, Pakistan

Abstract

This paper presents the solution proposed by team FAST Machine Translation to the shared tasks of JOKER CLEF 2022 Automatic pun and humour translation. State-of-the-art Transformer-based models are used to solve the three tasks introduced in the JOKER CLEF workshop. The Transformer model is a kind of neural network that tries to learn the contextual information from the sequential data by implicitly comprehending the existing relationships. In task 1, given a piece of text, we need to classify/explain any instance of wordplay is present in it or not. The proposed solution to task 1 combines the pipeline of token classification, text classification, and text generation. In task 2, we need to translate single words (nouns) containing a wordplay. This task is mapped to the problem of question answering (Q/A) on programmatically extracted texts from the OPUS parallel corpus. In task 3, contestants are required to translate the entire phrase containing the wordplay. Sequence-to sequence translation models are used to solve this task. The team has adopted different strategies for each task as they suited to the requirements therein. The paper reports proposed solutions, implementation details, experimental studies, and results obtained in JOKER CLEF 2022 automatic pun and humour translation tasks.

Keywords

Text Classification, Token Classification, Question Answering, Machine Translation, Transformers

1. Introduction

In our daily communications, humour is one of the most ubiquitous elements that we, as, a human comprehend comfortably with the help of previous cultural experiences and understandings. But, for the computers, this still remains one of the most daunting jobs as it is extremely difficult even for the expensive deep-learning-based solutions to decipher the double-meaning words which is a prominent feature of humour in different languages, and generate its appropriate parallel translation in the target language. This year JOKER CLEF-2022 team has come up with a unique set of challenges under the natural language processing domain. The workshop has brought professional translators and computer scientists together by presenting three different tasks to evaluate the understanding of translators and computer-based models about the humour. This paper will present our strategy to solve the given problems by fine-tuning transformer-based pre-trained deep learning models and then discuss the obtained results. The first task, named "Explain and classify instances of wordplay," presents tabular data with the

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

k214808@nu.edu.pk (F. Dhanani); muhammad.rafi@nu.edu.pk (M. Rafi); atif.tahir@nu.edu.pk (M. A. Tahir)
www.linkedin.com/in/farhan-dhanani-111253a3 (F. Dhanani)
0000-0002-7490-7655 (F. Dhanani); 0000-0002-3673-5979 (M. Rafi); 0000-0003-1366-8408 (M. A. Tahir)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3180/paper-136.pdf

following columns mentioned in the list below. The challenge is to predict the value for each column given an English text which possess the ***wordplay***, along with its ***id***.

- **ID**: An input value that uniquely identifies the associated wordplay text.
  ***Example*:** *pun_193*
- **WORDPLAY**: An input English text that contains a wordplay.
  ***Example*:** *Airline pilots make many friends in high places.*
- **LOCATION**: Words in the given English, which constructs the wordplay.
  ***Example*:** *high*
- **INTERPRETATION**: A possible explanation for the given wordplay in the English text.
  ***Example*:** *high (height)/high (addicted)/high (superior)*
- **HORIZONTAL/VERTICAL**:A binary categorical column to detect whether or not the target and source of the wordplay co-occur in the given English text.
  ***Example of horizontal, when source and target co-occur*:**
  They're called lessons (source) because they lessen (target) from day to day
  ***Example of vertical, when source and target gets collapsed into single word*:**
  Airline pilots make many friends in high (source/target) places.
- **MANIPULATION_TYPE**: A categorical variable to detect that the source and target of the wordplay are exact equivalents of each other **(Identity)**, or weakly resemble each other **(Similarity)**, or both possess different ordering **(Permutation)**, or its a group of initials that forms funny meaning **(Abbreviation)**.
  ***Example of Identity*:** *Airline pilots make many friends in high places.*
- **MANIPULATION_LEVEL**: A categorical variable to detect that the wordplay given in the English text is a kind of phonological manipulation **(Sound)**, or it is a kind of textual-based written manipulation **(Written)**, or if the detected wordplay is of some other form.
  ***Example of Sound*:** *Airline pilots make many friends in high places.*
- **CULTURAL_REFERENCE**: A boolean variable **(true/false)** to detect the existence of cultural reference in the given wordplay of the English text.
  ***Example of False*:** *Airline pilots make many friends in high places.*
- **CONVENTIONAL_FORM**: A boolean variable **(true/false)** to detect whether the given wordplay in the English text belongs to conventional form or not.
  ***Example of False*:** *Airline pilots make many friends in high places.*
- **OFFENSIVE**: This is a non-evaluated categorical variable. And we have ignored it throughout our experiments. Its purpose is to classify the given wordplay in the English text into offensive categories **(None, "Racist," "Possibly," "Sexist," "Other")**.
  ***Example of None*:** *Airline pilots make many friends in high places.*

The second task, named "Translate single words containing wordplay," specifies to predict an equivalent French version of a given English noun. Usually, the authors of anime, movies, comics, and video games choose the names of their story characters very carefully to advertise the core nature of their characters and the role they will be playing ahead in the story. Successful movies, comics, and video games, also get translated into other languages to attract audiences

from different regions. So, the authors try extremely hard to preserve original emotions by appropriately translating the name of their story characters while keeping the cultural background of the target language in context. The Pokemon series is a suitable example to explain this concept. For example, consider the following Pokemon shown in the figure 1 below.

Figure 1:

The image displays a Pokemon [1] character. It is named "EKANS" in the English version of the series, but French version, it's called "ABO."

Figure 2:

The image displays a character from the Astrix comic series [2]. Its name is "Dogmatrix" in the English version, but in the French version, its name is "Idéfix."

The name of the character shown in the above figure 1 is "EKANS" in the English version of the Pokemon [1] series, but in the French version, it is named "ABO." You can notice that the character visually looks like a small snake, and the authors wanted to preserve this similarity linguistically to educate their audience about the nature of the character. Therefore, they named the character "EKANS" in the English version to make it an anadrome of "SNAKE." In the French language, the boa means a masculine snake. Thus, in the French version, the writers have translated the name of this character to "ABO," which is an anadrome of "BOA." Because the french audience doesn't understand the word "SNAKE," the authors have renamed the character to keep them engaged with the vocabulary of their own native language. The challenge in Task 2 is to learn this mapping and predict an appropriate French translation for a given English noun. Figure 2 illustrates another example to understand this task. It presents a character from the Asterix series [2] named "Idéfix" in French. The French pun in this name is on the phrase "idée fixe", which means a fixed idea that illustrates that this character has a single-minded obsession. Alternatively, in the English version, the name of this character is "Dogmatix," which correctly maps the pun on the English word dogmatic to express the same idea of single-mindedness. Lastly, the third task presents English phrases containing a wordplay, and the challenge is to predict the corresponding French translation. It's important to note that multiple valid French translations may exist for a given English sentence containing a wordplay, and the task is to predict any one of it correctly. For example, consider the following scenario where both French translations are correct for the given English sentence.

- **English**: *"Be still my hart" she murmured, thinking how magnificent and stag - like he was.*
- **French-1**: *"Mon cœur se cerf", murmura-t-elle en voyant ce beau et majestueux mâle.*
- **French-2**: *Elle murmura "Calme-toi mon destrier" en pansant combien il était magnifique.*

This paper is further divided into three sections. The next section presents our approach to solving the mentioned tasks. The subsequent section lists the names of transformer models we have applied to solve these tasks. It provides the details of our experiments and the observed results. Finally, the last section discusses the conclusion and our learning in this workshop.

2. Materials and Methods

This section presents an overview of our strategies and summarizes the rationale for selecting these approaches to solve the three tasks of the JOKER CLEF 2022 [3] workshop. Let's discuss our solution for task 1 first.

2.1. TASK-1: Classify and explain instances of wordplay

2.1.1. Proposed approach for task-1

We have trained seven distinct models independently to predict the value for each of the seven target columns of task 1 listed in the previous section 1, given a pair of English wordplay text with its id. Firstly, we have utilized the token-classification-based method for preparing a model to extract the words forming the wordplay in a given English text, as depicted in figure 3. We have treated the English wordplay text as a series of space-separated tokens to locate the words forming the wordplay by classifying each token into the following three categories.

- **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.*

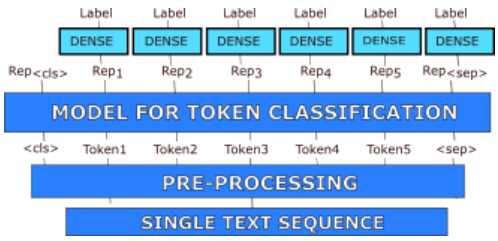

Figure 3:

The figure portrays our approach to mapping the task of finding the words forming the wordplay in the given English text onto the problem of token classification.

Figure 4:

The figure portrays our approach to mapping the task of predicting the labels for the five different categorical columns of the given tabular dataset onto the problem of text classification.

Employing this approach, we can apply several implementations of the token classification pipeline from the hugging face repository using auto encoding BERT-like transformer models such as *"bert-base-cased"* [4] to locate the words forming the wordplay in the given English text, as depicted in figure 3. Next, we have used the auto-regressive technique to construct a model to generate the interpretation for the extracted wordplay and used the text classification

scheme to build five separate models for inferring the values for the other five target columns, as illustrated in figure 4.

2.1.2. Data-set and processing

The JOKER CLEF 2022 [3] team has shared two versions of 10-column-based training and test sets for task 1 in both CSV and JSON formats. The tables below list the column names and the structure of both provided data sets.

No.	Column Names
1	ID
2	WORDPLAY
3	TARGET_WORD
4	DISAMBIGUATION
5	HORIZONTAL/VERTICAL
6	MANIPULATION_TYPE
7	MANIPULATION_LEVEL
8	CULTURAL_REFERENCE
9	CONVENTIONAL_FORM
10	OFFENSIVE

Table 1
The table shows the structure of the first training data set provided by the JOKER CLEF team for task 1 with the 531 records. The data set can be used for training the models to predict the value for each of the columns given a wordplay along with its Id.

No.	Column Names
1	ID (Given)
2	WORDPLAY (Given)
3	TARGET_WORD (Empty)
4	DISAMBIGUATION (Empty)
5	HORIZONTAL/VERTICAL (Empty)
6	MANIPULATION_TYPE (Empty)
7	MANIPULATION_LEVEL (Empty)
8	CULTURAL_REFERENCE (Empty)
9	CONVENTIONAL_FORM (Empty)
10	OFFENSIVE (Empty)

Table 2
The table shows the structure of the first test data set provided by the JOKER CLEF team for task 1 with the 4517 number of records. For each given pair of Wordplay and Id, the task is to predict the values of each empty column.

No.	Column Names
1	ID
2	WORDPLAY
3	LOCATION
4	INTERPRETATION
5	HORIZONTAL/VERTICAL
6	MANIPULATION_TYPE
7	MANIPULATION_LEVEL
8	CULTURAL_REFERENCE
9	CONVENTIONAL_FORM
10	OFFENSIVE

Table 3
The table shows the structure of the second training data set provided by the JOKER CLEF team for task 1 with the 2077 records. The data set can be used for training the models to predict the value for each of the columns given a wordplay along with its Id.

No.	Column Names
1	ID (Given)
2	WORDPLAY (Given)
3	LOCATION (Empty)
4	INTERPRETATION (Empty)
5	HORIZONTAL/VERTICAL (Empty)
6	MANIPULATION_TYPE (Empty)
7	MANIPULATION_LEVEL (Empty)
8	CULTURAL_REFERENCE (Empty)
9	CONVENTIONAL_FORM (Empty)
10	OFFENSIVE (Empty)

Table 4
The table shows the structure of the second test data set provided by the JOKER CLEF team for task 1 with the 3256 number of records. For each given pair of Wordplay and Id, the task is to predict values for each empty column.

The JOKER CLEF 2022 [3] team has first provided a smaller version of the training and test data set for task 1. Later as the competition timeline grew, they released an updated pair of more diverse training and test sets by adding more number of records in it. Both versions of the provided training data set for task 1 contain the correct values of all the remaining target columns against a given wordplay, along with its id. On the other hand, both the given test sets were for evaluation purposes to rank the approaches submitted by different teams that participated in the workshop. Thus, the test set only contains values under the wordplay and id column. All the remaining columns are empty, and for each listed wordplay in the English language, the challenge was to predict the values of the empty columns. It's important to note that both the versions of training and test sets have the same 10-column-based tabular structure apart from subtle naming differences. The ***"target_word"*** and ***"disambiguation"*** columns in the first pair of the given training and test sets map equally to the ***"location"*** and ***"interpretation"*** columns in the second pair of provided training and test set. We have used the provided training data sets to prepare our models and then used them to predict the values for each target column given a pair of English wordplay text and its id from the test data sets.

2.2. TASK-2: Translate single words containing wordplay

2.2.1. Proposed approach for task-2

We have mapped the task of learning the relation between English nouns and their corresponding French translations into the extractive Question/Answer (Q/A) problem by transforming the given three-column-based tabular data sets (Id, English Noun, French Noun) into extractive Question/Answer problem-styled data sets. To accomplish this transformation, we have utilized OPUS open-source parallel corpus [5] to artificially develop the context for all English/French noun pairs provided in the task 2 data set. We have iteratively selected each English/French noun pair listed in the provided data set. Then extracted those English/French parallel sentence pairs from the OPUS open-source parallel corpus [5] that contains the selected English noun in its English version and the translated French noun in its French version. In such a way, we have collected contexts for all English/French noun pairs. And transformed the task 2 data set where each record is composed of an English noun and its French translation, along with a list of extracted English/French parallel sentence pairs in the form of contexts. Now the task for the Question/Answering models is to use the English noun as a query and predict the location of its French translation in the French version of the extracted English/French parallel sentence pairs.

2.2.2. Data-set and processing

The JOKER CLEF 2022 [3] team has shared two 3-column-based tabular data sets for task 2 in both CSV and JSON formats. One is for training the models to predict an equivalent French version of the given English noun, along with a unique Id as input. The second is for testing/ranking the models submitted by different teams by evaluating their predictions for the given pairs of English nouns with a distinct id. It's essential to note that the test set only holds the list of English nouns and their corresponding ids. The associated French translations for each English noun are absent from the test set, and the challenge in this task is to predict these unknown French translations based on which the submissions from various teams will get

ranked. The training set contains 1164 records, while the test set holds 284 data points. Table 5 and 6 below tabulates the structure of the provided train and test data sets, respectively.

No.	Column Names
1	ID
2	EN
3	FR

Table 5
The table shows the structure of the training data set provided by the JOKER CLEF team for task 2 with the 1164 records.

No.	Column Names
1	ID (Given)
2	EN (Given)
3	FR (Empty)

Table 6
The table shows the structure of the test data set provided by the JOKER CLEF team for task 2 with the 284 number of records.

We have transformed the provided 3-column-based tabular training and test data set for task 2 into the extractive question-answer styled data set by utilizing OPUS parallel corpus [5]. In order to transform the training set, we have iteratively selected each of the listed English/French noun pairs. And then pulled out those English/French parallel sentence pairs from the OPUS parallel corpus [5] that possess the selected English noun in its English version and the corresponding French translation in its French version. We have programmatically ensured that the English version sentence pulled from the OPUS parallel corpus [5] must hold at least one English noun from task-2's training data set. And its corresponding parallel French version must also contain a French translation of that English noun. So to visualize the transformation, suppose we have only one record in the task 2 training data set, as shown in the table 7.

Id	En	Fr
4	Obelix	Obélix

Table 7
The table describes the structure of a single record in the training data set of task 2.

Id	Context	Question	Answers
4	astérix et obélix ne devraient plus quitter le village.	Obelix	{"text": [Obélix], "answer_start": [11]}

Table 8
The table shows the conversion of the table 7 records into the extractive Q/A styled data. The transformation shows that the English noun has now become the question and the context column holds a French sentence extracted from the OPUS corpus. The answers column holds a JSON that contains the actual French translation and the position of its appearance in the French sentence of the context.

Given such a scenario, we can pull the following sentence (English/French) pair listed below from the OPUS parallel corpus [5] to generate the extractive Question/Answer problem-styled data set shown in the table 8.

- **English Version**: *asterix and obelix should stay in the village and not go in the forest!*
- **French Version**: *astérix et obélix ne devraient plus quitter le village.*

Observe that the English and French version of the extracted pair contains the English and French nouns mentioned in table 7, respectively. After the transformation, we can utilize popular pre-trained extractive question answering models from the hugging face [6] repositories to predict the French translation for a given English noun. The models will use the English nouns from the JOKER CLEF [3] task 2 training's data set as the input question, along with the corresponding French sentence pulled from the OPUS parallel corpus [5] as their context. And now, the task for the deep learning models is to learn to locate the exact position of the French translation in the French text for the queried English noun. After the process of training completes, we have again transformed the test data set for task 2 into the extractive Question/Answer styled test data set by applying a similar strategy. The test set of task 2 holds test records for which we don't know the correct French translation of the given English noun. Because of this, we have only ensured the existence of the given English noun in the English version of the extracted (English/French) sentence pairs and assumed that the corresponding French translation must also hold its equivalent French version. It's a weak assumption, but we have made this architectural choice to design the solution. Again to mentally simulate the process suppose we only have one record in the test data set given in the table 9. Then we can pull the following (English/French) sentence pair from the OPUS parallel corpus [5] to generate the extractive Question/Answer problem-styled test-data set shown in the table 10 below.

- **English Version**: *I went to loompaland looking for exotic new flavors for candy.*
- **French Version**: *j'étais venu à lumpaland pour chercher de nouvelles saveurs.*

Id	En
18	Loompaland

Table 9
The table shows the structure of a single record in the test data set of task 2.

Id	Context	Question
17	j'étais venu à lumpaland pour chercher de nouvelles saveurs.	loompaland

Table 10
The table describes the conversion of the table 9 records into the extractive question/answer styled test data. Note, unlike table 8, there is no answer column here because we are transforming test records, we don't know the correct translations for the queried English noun. During the test time, the trained models need to generate the answer.

2.3. TASK-3: Translate entire phrases containing wordplay

2.3.1. Proposed approach for task-3

The problem description of task 3 is a classical example of sequence to sequence prediction, where the model needs to predict an equivalent French translation for the given English text.

But, in this case, the challenge is there can be multiple valid possible translations for a given input English sentence containing a wordplay. We have utilized sequence-to-sequence models from the hugging face repository to learn any matching French translations for the given English sentence from the task 3 data set.

2.3.2. Data-set and processing

The JOKER CLEF 2022 [3] team has shared two 3-column-based tabular data sets for task 3 in both CSV and JSON formats. One is provided for training the model to make it learn to generate an equivalent French translation for the given English text which may possess a wordplay. The other is for testing/ranking the models submitted by different teams by evaluating their predictions for the listed English texts with a distinct id. Again the id is just for uniquely identifying each record in the provided data sets. Plus, the test set only holds the list of English sentences along with the associated ids. The French translations for each English sentence are absent from the test set, and the challenge in this task is to generate the French translations for each of the given English sentences in the test set based on which the submissions from various teams will get ranked. The training set contains 5115 records, while the test set holds 2378 data points. Table 11 and 12 below tabulates the structure of the provided train and test data sets, respectively.

No.	Column Names
1	ID
2	EN
3	FR

Table 11
The table shows the structure of the training data set provided by the JOKER CLEF team for task 3 with the 5115 records.

No.	Column Names
1	ID (Given)
2	EN (Given)
3	FR (Empty)

Table 12
The table shows the structure of the test data set provided by the JOKER CLEF team for task 3 with the 2378 number of records.

Id	En	Fr
18	Tom said piously.	déclara Tom pi-eusement.
19	Tom said piously.	dit Tom pieusement.
20	Tom said piously.	Tom dit pieusement.

Table 13
The table shows the configuration of the task 3 data set when there exist multiple valid French translations for an English text.

The provided data sets for task 3 are in tabular form, due to which there are duplicate entries when multiple valid French translations are possible for a given English sentence. For example, consider the following English text. The multiple possible French translations for this English text can be arranged in the tabular structure as shown in table 13.

- **English Text**: *"Tom said piously."*

We have transformed the provided tabular training set into a JSON dictionary. The key in this dictionary is the English text, and its associated value contains the list of all possible French translations for the keyed English text from the training set. We have used the prepared JSON object to train sequence-to-sequence transformer models for learning the mapping between the English text and any of its corresponding valid French translations. Figure 5 visually expresses this transformation by converting the records listed in table 13 into a JSON object.

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

Figure 5: The figure shows the conversion of table 13 data records into a single JSON object.

3. Experiments and Results

This section will list the transformer architectures we have utilized to implement the approaches discussed in the previous section for solving tasks of the JOKER CLEF 2022 [3] workshop and their results. It's important to note that we have also shared our codebase on the public GitHub repository [7]. So, all the experiments can be easily re-executed to reproduce the mentioned results. Let's first discuss the implementation of the solution for task 1 and the obtained results.

3.1. IMPLEMENTATION OF TASK-1: Classify/explain instances of wordplay

We have used the listed pre-trained transformer models from the hugging face repository and fine-tuned them on the given training data sets for task 1 to make them learn to locate the words forming the wordplay in the given English text through token classification. The KEY-BERT [8] model can be utilized with different embedders. In this experiment, we have only used it with the fine-tuned "bert-base" [4] model and the pre-trained "all-MiniLM-L6-v2" [9] model.

- Pre-trained BERT-BASE [4]
- KEY-BERT [8] with fine-tuned BERT-BASE [4] as its embedder.
- KEY-BERT [8] with pre-trained all-MiniLM-L6-v2 [9] sentence transformer as embedder.

The JOKER CLEF 2022 team has provided two pairs of training and test sets for task 1. They have first released a smaller version of the data set, and then later in time, they have disclosed an updated bigger version of both the training and test set. We have processed them independently and employed the listed BERT-based transformer models on each of them separately. We have applied the hold-out approach and pulled aside 9% of the records from both the training data sets provided for task 1, where the length of the given English text containing the wordplay was more than two. We have kept them hidden from the model throughout its training and used them later to evaluate and rank the predictions of the fine-tuned models in locating the words forming the wordplay. Moreover, we have also further extracted the 4% of the data from both the training set for validation purposes and then fine-tuned two separate instances of the BERT base model for less than five epochs on the remaining records of the training sets. Lastly,

after fine-tuning, we have evaluated the predictions generated from the fine-tuned models for the 9% of the records, which were initially extracted from the training sets to estimate their performance on unknown data points, as shown in the table below. Overall our approach has generated comparatively good results for the first data set provided by the JOKER CLEF 2022 [3] team for task 1.

Model Name	**Accuracy on 9% of the records extracted from training set-1**	**Accuracy on 9% of the records extracted from training set-2**
BERT-BASE [4] (FINE-TUNED)	71%	31%
KEY-BERT [8] with fine-tuned BERT-BASE [4] as its embedder	33%	16%
KEY-BERT [8] with pre-trained all-MiniLM-L6-v2E [9] as its embedder	15%	3%

Table 14
The table shows the performance of selected BERT based transformers models for precisely identifying the words forming the wordplay with exact matches on 9% of the records extracted from the provided training data-sets for task 1 via hold out approach. If the model fails to locate even a single character of the wordplay in the given English text, then we have counted as an overall failure.

Table 14 showcases that the fine-tuned BERT-BASE [4] model delivers the best performance compared to other variants of KEY-BERT models on both versions of the training data sets. So, we have used this model to locate the words forming the wordplay among the listed English sentences in the provided test data sets for task 1. We have used the hold-out approach instead of the K cross-validation to evaluate the performance of BERT based transformer models for locating the wordplay in the given English sentences because the provided training data sets contain numerous instances where the length of the English sentences was one. And in all such instances, it was apparent that the given English sentence was itself the wordplay. So, obviously, the deep learning models will achieve a perfect score in such scenarios. Thus, using such instances for evaluating BERT based transformer models will result in an unfair boost in their performance. To mitigate this effect, we have used the hold-out approach and selectively extracted those records from the training set in which the length of the English sentence was more than two. Additionally, we have downloaded the pre-trained DistilBERT [10] model from the hugging face repository to perform the text classification on the provided English sentence to predict the categorical label for the remaining target columns of the task 1 listed in table 15. We have made five separate copies of the pre-trained DistilBERT [10] model and trained them individually to infer the label of each of the five categorical target columns of task 1. We have removed the "nan" values from both versions of the provided training data sets for task 1 and used them to evaluate the performances of all five copies of the DistilBERT [10] model independently via 10-fold cross-validation. Table 15 shows the mean accuracy of the DistilBERT [10] model on both versions of the training data set for correctly predicting the categorical labels of all the five mentioned target columns. Comparatively, we can observe that

the DistilBERT [10] model has provided more promising results for the first data set provided by the JOKER CLEF 2022 [3] team for task 1.

Column Name	Mean Accuracy of 10 fold cross validation on training set-1	Mean Accuracy of 10 fold cross validation on training set-2
MANIPULATION_TYPE	65%	50%
MANIPULATION_LEVEL	99%	99%
HORIZONTAL/VERTICAL	93%	99%
CULTURAL_REFERENCE	96%	95%
CONVENTIONAL_FORM	92%	89%

Table 15
The table shows the mean accuracies of the DistilBERT [10] model across ten folds of the training sets of task 1 for predicting labels of the categorical columns after getting fine-tuned for less than five epochs.

Column Name	Obtained Scores on test set-2 having 3256 records
LOCATION	1455
MANIPULATION_TYPE	1667
MANIPULATION_LEVEL	2437
HORIZONTAL/VERTICAL	68
CULTURAL_REFERENCE	Not evaluated
CONVENTIONAL_FORM	Not evaluated
OFFENSIVE	Not evaluated

Table 16
The table states the obtained scores on the predictions generated by our fine-tuned models for the test records of test set-2 of task 1. The scores were allotted by evaluators from the JOKER CLEF 2022 [3] team, and the scoring criterion was straightforward. They have given a point for correctly predicting the value of a target column for each of the provided English texts in the second test set of task 1.

Lastly, we have selected the fine-tuned BERT-BASE [4] model for locating the words forming the wordplay in the English sentences listed in both versions of the test sets. Along with five separate copies of the fine-tuned DistilBERT [10] model for predicting labels of the remaining five categorical target columns of task 1 and submitted our predictions to the evaluators of the JOKER CLEF 2022 [3] workshop. We were the only team that successfully submitted the predictions for the test set-1. Thus for the first test set, we have implicitly got the first rank, so the evaluators haven't released any other statistical details or scores for the test set-1 of task 1. Furthermore, for test set-2, we have managed to get ourselves among the top three positions. The table 16 reveals the scores for the generated predictions from our fine-tuned models to correctly predict the values of all the target columns for each English text listed in the test set-2 of task 1. It's important to note that the evaluators have awarded a score of one point for predicting a correct value for each target column of task 1 against an English text containing a wordplay from test set-2. Plus, the evaluators have not evaluated the predictions for the *"cultural reference"*, *"conventional form"*, and *"offensive"* columns.

3.2. IMPLEMENTATION OF TASK-2: Translate words containing wordplay

We have downloaded pre-trained CamemBERT [11] and DistilBERT [12] models and fine-tuned them to perform the task of extractive question answering for task 2. The CamemBERT [11] model is pre-trained in the French language to retrieve answers for the provided French queries in the French context. Contrastingly, the DistilBERT [12] model is pre-trained in the English language to extract answers in the English context for the given English questions. The reason for choosing these two distinct pre-trained models designed for different languages is because the contents of the transformed data set for task 2 consist of both French and English language. In the previous section, we have observed that our approach has transformed the problem of translating single-word English nouns to their equivalent French versions into the extractive question answering domain. As a result of this transformation, the extractive question-answering models have to process the input question in English and the associated context in the French language. So now, the task of the extractive question-answering models is to extract the French translation for the given English query from the French context. Because of this heterogeneity of different languages in the transformed data set, we have utilized two different English and French pre-trained extractive question-answering models and compared their performance. We have used 10-fold cross-validation to evaluate the performance of both the DistilBERT [12] and CamemBERT [11] models on the transformed training data set. Table 17 shows the mean performance of both models by stating the average percentages of predictions that exactly matched the expected translations in French for the given English nouns.

Model Name	Mean Accuracy of 10 fold cross validation on the transformed training set
CamemBERT [11]	59%
DistilBERT [12]	94%

Table 17
The table shows the performance of the DistilBERT[12] and CamemBERT[11] models for performing extractive question answering across the ten folds of the transformed training data set of task 2 after getting fine-tuned for less than five epochs.

Table 17 also reveals that the fine-tuned DistilBERT [12] model generates more satisfactory predictions compared to the CamemBERT [11] model. We have used the DistilBERT [12] model to generate the French translations for the English nouns of the test set, and our submission ranked first in the competition. The test set consists of a total of 284 English nouns that include named entities from official movies, and novels for which official French translations were available. However, the evaluators have also considered unofficial translations for a few English nouns, which were popular among the native French audience. The evaluators have used simple case insensitive string matching to evaluate official French translations for English-named entities with their expected French versions. But string matching can not be used to assess unofficial translations because they are not part of authentic literature. So, the evaluators have manually assessed the unofficial translations based on lexical field preservation, sense preservation, comprehensibility, and the formed wordplay. The table 18 below demonstrates

the obtained score of test submissions against each of the mentioned parameters.

Metric	Score	Explanation
Total	284	Total number of records in the test set.
Not translated	0	Total number of records that are either missing or not translated in the submission file.
Official	250	Number of the official named entities that are correctly translated in the submission file.
Not Official	34	Number of translations in the submission file that are unofficial.
Lexical Field Preservation	16	Number of translations that preserve the lexical field of the source wordplay in the submission file.
Sense Preservation	13	Number of translations that preserve the sense of the source wordplay in the submission file.
Comprehensible Terms	26	Number of translations that do not exploit any specialized terms in the submission file.
Wordplay form	3	Number of translations that are itself wordplay in the submission file.

Table 18
The table shows the obtained scores on the translations produced by our fine-tuned models for the test records of task 2, along with a brief description of what the mentioned numbers against each testing parameter represent. The official paper released by JOKER CLEF [3] 2022 organizers discusses the mentioned results in more detail. And also showcase a thorough comparison of the performance of all participating teams on the provided test set.

3.3. IMPLEMENTATION OF TASK-3: Translate entire phrases containing wordplay.

The training data set provided for task 3 only comprised 1,185 unique English sentences for which there exist multiple valid French translations. Because the data set was not huge, we have decided not to fine-tune pre-train models on the training data set. Our goal was to select the pre-trained model that generates predictions that, on average, provide the highest BLEU [13] scores and least TER [14] scores for the given English phrases from the training set without fine-tuning. We have assumed that the high BLEU [13] scores and low TER [14] scores indicate that the generated French translations by the model for the given English phrases are more similar to expected French translations. It's a weak assumption, but still, we have made this architectural choice to design the solution for task 3. We have downloaded four popular pre-trained sequences to sequence transformer models from the hugging face repositories [6] listed in the table 19 and evaluated their performance on the provided training set of task 3. Table 19 mentions the names of the selected models and states their performance with the help of BLEU [13] and TER [14] scores. We have iteratively extracted English texts from the data-set of task 3 and provided it as an input to these four models to evaluate their predictions. We have recorded BLEU [13] and TER [14] scores for each generated prediction by the four models and then calculated the average to rank the overall performance of the models on the entire training set.

Model Name	AVERAGE BLEU SCORE	AVERAGE TER SCORE
Helsinki-NLP/opus-mt-en-fr [15]	18.43	0.80
GOOGLE T5 BASE [16]	12.64	0.84
GOOGLE T5 SMALL [17]	11.07	0.85
GOOGLE T5 LARGE [18]	11.90	0.84

Table 19
The table presents the average BLEU [13] and TER [14] scores achieved by the selected pre-trained sequence-to-sequence transformer models for translating the English texts provided in the training data set of task 3 to their equivalent French versions along with the composed wordplays.

Metric	Score	Explanation
Total	2378	Total number of records in the test set.
valid	2120	Total number of submitted translations that are valid.
Not translated	103	Total number of missing or invalid translations in the submission file.
Nonsense	220	Total number of the translations in the submission file that don't make sense to native French speakers.
Syntax problem	58	Total number of the translations in the submission file that contains syntactical errors.
Lexical problem	79	Total number of the translations in the submission file that contains lexical errors.
Lexical field Preservation	1739	Number of translations in the submission file that have preserved the lexical field of the source wordplay.
Sense preservation	1453	Number of translations in the submission file that have preserved the sense of the source wordplay.
Comprehen-sible terms	867	Number of translations in the submission file that haven't exploited very specialized terms.
Wordplay form	345	Number of correct translations provided in the submission files that are itself wordplay.
Identifiable wordplay	318	Number of translations provided that are itself wordplay and understandable by the audience.
Over-translation	1	Number of translations in the submission file that contains useless words and are unnecessarily long.
Style shift	12	Number of translations that have a style shift. For example when vulgarism exists in the source sentence or the produced translation but not in both.
Hilariousness shift	765	Number of French translations that are much less or much funnier than the source sentence.

Table 20
The table shows the obtained scores on the translations produced by our fine-tuned Helsinki-NLP/opus-mt-en-fr [15] model for the test records of task 3, along with a brief description of what the mentioned numbers against each testing parameter represent. The official paper released by JOKER CLEF [3] 2022 discusses the mentioned results in more detail. And also showcase a thorough comparison of the performance of all participating teams on the provided test set.

The table 19 entails that the Helsinki/NLP/opus-mt-en-fr [15] model has given the best performance and produced more desired translation as compared to other models. We have used this model to generate predictions for English phrases listed in the test set of task 3 and received second position in the contest. There were a total of 2378 English phrases provided in the test set, and we have generated a French translation for each of them. So, our submission file contains a list of 2378 pairs of English and French sentences. The evaluators have manually scored each of the submitted French translations using thirteen different parameters to rate the quality of the generated predictions. Table 20 shows a list of these parameters and explains how the evaluators have used each of them. The table also lists the obtained scores of our submitted predictions against each of the listed thirteen parameters.

4. Conclusion

The BERT-Base model has given the best performance for locating the words forming the wordplay in the English texts of the task 1 data set. And the DistilBERT model has achieved the best performance in predicting classification labels for the remaining target columns of task 1. The DisitilBERT model successfully delivered more than 89% accuracy in predicting all of the categorical target columns of task 1, except the manipulation type of wordplay. We think in the future, the results of the DistilBERT model can be compared with other popular text classification BERT alternatives to rank its performance. Additionally, we haven't employed the large BERT variants to locate the wordplay in the given English text of the task 1 data set, but we strongly think that they will boost the performance of our approach. The main highlight of our work is our designed technique for solving task 2 in Question/Answering style. And our submission of task 2 is also ranked top in the competition. An extension of this work can be to test the approach with different language pairs because, in this paper, we have only evaluated it on English/French noun pairs as per task 2 requirements. The DistilBERT model again delivers the best performance for solving task 2. And accurately predicts 96% of the French translations for the given English nouns in the extractive question/answer style. Lastly, we have concluded that Helsinki-NLP/opus-mt-en-fr model has provided the best performance on the task 3 data set by achieving 18.43 and 0.80 averaged BLEU and TER scores. The data set of task 3 doesn't contain a large number of records in it. We think in the future, increasing the size of the English/French parallel corpus containing wordplay and humour will benefit in excelling the research and will immensely help in training and evaluating the sequence-to-sequence models.

5. Acknowledgments

This research is supported by the school of computing National University of Computer and Emerging Sciences (FAST-NU). I would like to thank my supervisor, Dr. Muhammad Rafi, and Dr. Atif Tahir, director of the university, for providing technical insights and expertise that greatly assisted the research. We also like to recognize the efforts of the JOKER 2022 team to organize this workshop and develop practical tasks along with datasets. I believe this workshop has provided a platform where students can apply their NLP knowledge to solve challenging problems and evaluate their understanding. We look forward to participating again in this event

next year and wish this kind of event should happen more. Furthermore, we are also immensely grateful to Liana Ermakova for keeping us posted about the updates of the event and solving our queries timely to help us throughout the event.

References

[1] Pokémon, Pokémon — Wikipedia, the free encyclopedia, 2001. URL: https://en.wikipedia.org/wiki/Pok%C3%A9mon, online; accessed May 24, 2022.

[2] Asterix, Asterix — Wikipedia, the free encyclopedia, 2001. URL: https://en.wikipedia.org/wiki/Asterix, online; accessed May 24, 2022.

[3] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, Mathurin, G. L. Corre, S. Araújo, J. Boccou, A. Digue, A. Damoy, P. Campen, B. Jeanjean, Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation workshop, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), 2022, p. 25.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[5] J. Tiedemann, Parallel data, tools and interfaces in opus, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[7] FARHAN, Fast-mt team submission for joker clef 2022, 2022. URL: https://github.com/FarhanDhanani/joker-clef-22-FAST-MT.

[8] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[10] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019). URL: https://huggingface.co/distilbert-base-uncased, accessed: 2022-05-24.

[11] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[12] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (for question answering), in: NeurIPS EMC Workshop, 2019. URL: https://huggingface.co/distilbert-base-cased-distilled-squad, accessed: May 24, 2022.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[14] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 2006, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25.

[15] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020. URL: https://huggingface.co/Helsinki-NLP/opus-mt-en-fr, accessed: May 24, 2022.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer (t5 base), Journal of Machine Learning Research 21 (2020) 1–67. URL: https://huggingface.co/t5-base, accessed: May 24, 2022.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer (t5 small), Journal of Machine Learning Research 21 (2020) 1–67. URL: https://huggingface.co/t5-small, accessed: May 24, 2022.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer (t5 large), Journal of Machine Learning Research 21 (2020) 1–67. URL: https://huggingface.co/t5-large, accessed: May 24, 2022.