NLPalma @ CLEF 2023 JOKER: A BLOOMZ and BERT Approach for Wordplay Detection and Translation

Notebook for the JOKER Lab at CLEF 2023

Victor Manuel Palma Preciado [1,2], Carolina Palma Preciado [1] and Grigori Sidorov [1]

[1] *Instituto Politécnico Nacional de México, Gustavo A. Madero, Ciudad de México, México*
[2] *Université de Bretagne Occidentale, HCTI, France*

Abstract

The following work has the purpose of describing the participation in the JOKER 2023 track in the classification Task 1 in which it is asked to classify sentences that contain wordplay and the translation Task 3 in which this same type of sentence has to be translated into Spanish trying to maintain the wordplay and the sense in a certain way tropicalizing the sentence to Spanish language from English. Given the constant use of models such as GPT and T5, it was decided to take the direction of language models, making a slight fine-tuning of BLOOMZ & mT5 models with which to address the problem of translation with a simple prompt and on the other hand the use of BERT as a model proposed for the classification task. The results were mixed since in the manual review of the BLOOMZ & mT5 translated sentences not many were found to contain the Spanish pun, in the case of classification somewhat similar results were obtained given the structure of the dataset. It is believed that given their performance the models can still be optimized and improve the accuracy with which they classify, in the case of the translations perhaps another methodology could be chosen to improve the results obtained in this task, in general, the results are satisfactory for a first approach.

Keywords

wordplay, humorism, pun, wordplay classification, wordplay translation

1. Introduction

The main objective of the work is to find robust methods, to achieve good results in the tasks provided, specifically for Task 1 and Task 3. In Task 1, the model must be able to classify between those sentences that contain wordplay and those that do not contain wordplay and therefore would not contain humor. This task is based on the Spanish dataset of JOKER 2023 [1].

In the case of Task 3, the objective was to correctly translate an English pun into Spanish in such a way that the produced translation also contains the pun. In this case, the translated pun should also make sense and be adapted to something appropriate within the context. In a certain way adapting the term to something according to what could be expected in the translation, since the terms that form the wordplay usually carry certain homophone elements that when translated do not have a homologous.

It is important to note that sometimes a direct translation may not have an exact equivalent for a wordplay, but it is expected that the model is able to see beyond the simple literal translation of term to term and find the connection between the wordplay.

The dataset used contains Spanish and English puns, which were fed to a large language model to obtain translations that capture the correct type of pun humor or, at the very least maintain some humor in the resulting translation. Additionally, for the classification task a BERT-like model was employed,

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: victorpapre@gmail.com (A. 1); c.palma.p0@gmail.com (A. 2); sidorov@cic.ipn.mx (A. 3)
ORCID: 0000-0001-8711-1106 (A. 1); 0000-0003-3253-4464 (A. 2); 0000-0003-3901-3522 (A. 3)

© 2023 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)

since in other instances of humor, this model has demonstrated a certain level of effectiveness. Therefore, this time we aim to test if it can really be used for different cases and if its use can be generalized.

An important aspect of language-based models, especially BLOOMZ & mT5 [2] is they are optimized to follow the next token predicted, in this case, following a prompt with instructions oriented to the translation task. It is believed that such models could be useful to maintain the wordplay in the translation. Although this type of tasks can become a little complicated since certain terms used in different languages can make it difficult how to perceive the wordplay of the original sentence to the translated one. This issue is later encountered, as these challenges can normally arise within a model.

The use of homophones can complicate the task even more, as there may not be a similar word in the translation that matches or has the same intention, altering in a certain way the interpretation of the joke with the pun, thus changing the original meaning. It could be said that the pun is maintained, but not the meaning. With the understanding that most puns become partial translations of what the original sentence proposes, we can say that there are different phenomena present in the translation of puns whether they are homophones, homographs, or this comes from a regional homophony of the sentence.

The task of classification presents a different challenge because puns can be confused by those sentences that unintentionally exhibit similar characteristics to a pun. Understanding the type of sentences that our model needs to screen can give us a broader idea of what needs to be done. In this case, the classification of puns appears to include certain sentences that emulate puns, but in reality, they are not. This can hinder and confuse the model, as it introduces a certain amount of noise with which we must deal.

2. Approach to the task

The tasks proposed by JOKER in CLEF 2023 primarily focus on the automatic analysis of puns. In this edition, a corpus consisting not only of English and French text but also of Spanish was available. Based on previous successful works [4, 5], it was decided to use the transformers approach [3]. Due to time constraints, of the three available tasks, only Task 1, involving wordplay detection, and Task 3, focusing on translation, were executed. It should be noted that tasks were performed solely on Spanish and English texts from the training corpus provided by JOKER.

The first task focused on pun detection, which can be solved as a binary classification task since it has yes and no labels that indicate whether a text is a pun or not. Initially, it was considered to train a multilingual model such as multilingual BERT [7], which is pre-trained in 104 languages. However, due to the low performance obtained during evaluation, it was decided to keep the model but apply a separate approach.

As a result, two models were trained: one specifically for English and the other for a mixture of English-Spanish, both utilizing the same BERT architecture [6]. In the end, the models were trained with BERT and multilingual BERT. For this task, the model was loaded from the Hugging Face transformers module with the help of the Ktrain [8] wrapper, which facilitated the process of loading and fine-tuning the model in a fast and simple manner.

Due to the fact that transformers do not require extensive preprocessing, the training process was relatively straightforward. However, during the fine-tuning phase, it was necessary to experiment with different parameters to achieve the best results in validation accuracy and loss. In the end, the best models were saved in a Keras H5 format for future reference.

For the training of the BERT-like models, the following parameters were utilized: a batch size of 8, 3 training epochs, and a learning rate of 7e-6 for the BERT model trained on English data. On the other hand, the parameters for BERT-multilingual trained on English-Spanish data were a batch size of 32, 3 training epochs, and a learning rate of 5e-5. Considering the text length from the JOKER corpus, a maximum length of 70 tokens was declared.

To develop Task 3, consisting of the translation of word plays from English to Spanish, BLOOMZ & mT5 was implemented, which is a more robust version of BLOOM, offering an improvement in task performance. This model can follow human instructions through the use of a prompt such as the examples used in the experimentation: “Translate into Spanish:”, “Translate the text into Spanish:”, “Translate into Spanish the next text:”, and “Translate into Spanish the next sentence:”. It is worth

mentioning that the result obtained depends on the prompt used. In this case, four passes or runs were made on the text to achieve a better result, each pass corresponding to one of the provided prompts.

These iterations were necessary because in some passes the results were not favorable, either because the model fail to produce a translation (resulting in blank output), generating incomplete translation, or because the model produced more text than necessary (generating new text beyond the necessary translation).

The queries were conducted using the Hugging Face API, specifically the Inference API [9], which facilitates the execution of queries using machine learning models. As mentioned, multiple iterations were performed due to the results obtained in each run, some sets of wordplays were either not translated or did not yield a proper result. It was required to try different prompts to get the best possible result, yet some puns failed to generate a translation, resulting in a blank query even after multiple iterations.

Although the generation of the translation was fully automated, a sample of the results were manually reviewed to identify potential errors. These errors were taken into consideration to determine which texts needed a re-runs. For instance, in the case of the pun "What do you call a doctor who treats retired soldiers? A veteran-arian." the first two attempts did not produce any results. It was in the third run that the text was successfully translated to *"¿Qué es un doctor que trata a soldados retirados? Un veterano-ario.*". Another example is the sentence "Recent survey revealed 6 out of 7 dwarf's aren't happy." which was generated on the second pass: "*Una reciente encuesta reveló que 6 de cada 7 enanos no son felices.*"

3. Resources employed

Among the resources used to train and evaluate the models, Google's Colab environment was employed. This platform enables Python programming and execution, while also providing easier use of GPU. The use of GPU resources allowed for faster execution in the performance of the described tasks, by enabling multiple simultaneous computations. The server used has the following specifications: GPU NVIDIA-SMI 525.85.12, CUDA v12.0, and 25 of RAM.

As part of the resources, it was decided to use a combination of Google Colaboratory and Petals [10] for Task 3 translation, since we did not have the computational capacity to handle models with parameter size 10B, which requires a significant amount of computation beyond our capabilities. The use of Petals becomes indispensable for those jobs that do not have enough computational capacity, as it leverages the benefits of crowd computing to make inferences with large models in a relatively easy/semi-efficient way. The platform offers the flexibility of an API and the power of PyTorch, for seamless integration in the required tasks.

On the other hand, for other BERT-like models that are less demanding in the use of resources, it is necessary to consider the data volume and the desired width of tokenization. In the experimentation process BETO [11], BERT, and BERT-multilingual models were tested for the classification task, likewise, the training set was taken under three different approaches:

1. Spanish-only dataset.
2. English-only dataset.
3. And a mixture of English-Spanish data.

4. Results

As can be seen in the tables below, the probabilities of each of the tasks and a brief explanation of the cases are provided. Some of the sentences demonstrate a clear example of a pun, while others are not so straightforward in terms of their intended meaning. This variability in pun structures contributes to the challenge of accurately identifying puns. It is expected that the model may have some confusion in finding the characteristics between sentences that lack a pun's connotation but exhibit high similarity to those that do. This type of concern seems to be overlooked by the model to some extent. It is logical to evaluate the data set to get some insights into how the model works, which is exactly what is presented in the following tables.

Task 1: Detection of puns

In this section, the results obtained for the binary classification of English and Spanish puns are presented. As the test set labels were not provided, some inferences about the wordplays are made to explain the model performance when evaluating this type of sentence.

English corpus

The BERT model trained on the English corpus consisting of 5,292 wordplays, was divided into a training and validation set in a 80:20 distribution. Subsequently, the model was used to evaluate the test set provided by JOKER, which comprised 3,183 sentences. The probability of a text containing a wordplay was obtained, and the top 5 wordplays and non-wordplay were identified.

In Table 1, it can be seen that a positive result indicating a wordplay involves a specific structure, where a sentence contains references in pairs (the words indicating this event are marked in bold letters). The highest results obtained a probability of around 0.99, where three of the puns with the highest probability of being wordplay refer to a pun referring to lawyers, as an example:

*The lawyer asked a **loaded** question about **guns**.*

The pun talks about a loaded question, and it is easily understood that the play on the words lies in the relationship between “loaded” and “guns”. This example provides a small idea of how the model classifies such positive instances.

Table 1
Top positive cases for wordplay classification of the English corpus

Pun	Probability
She was only a lawyer's daughter, but what a will to break.	0.99865
She was only an Attorney's daughter, but what a will to break.	0.99862
The lawyer asked a loaded question about guns.	0.99860
Once ice cream was invented the problem was licked.	0.99860
Our Boy Scouts' knot - tying class went off without a hitch.	0.99859

Regarding the results with a low probability of being puns, that indicate a non-wordplay sentence. It can be observed that, unlike the previous table, these examples are not as obvious in terms of finding the wordplay. Therefore, if even for humans it is challenging to identify this word matching, it can be even more difficult for the model to understand the nuances that these sentences have with the use of puns. The results in Table 2 show how these sentences do not contain wordplay, indicating that the model correctly classified them as non-wordplay.

Table 2
Top negative cases for wordplay classification of the English corpus

Pun	Probability
What's the name of that street in Paris? asked Tom quietly.	0.00065
I wouldn't marry you if you were the only woman on earth, said Tom quietly.	0.00066
Give me some pre - packed cheese slices, said Tom professionally.	0.00068
''Let's take a vacation in the south of France,'' said Tom loudly.	0.00068
Can you read music? the bandleader asked calmly.	0.00068

Spanish corpus

The multilingual BERT model used to classify the Spanish test set (2,241 sentences) presents distinct curiosities compared to the problems encounter in the English dataset. This may be attributed to the fact that incorporating a mixture of datasets strengthened the model's ability to infer the type of wordplays, but more so the type of redundant wordplays that was able to correctly classify. The example below shows a sentence that was accurately classified as a wordplay:

*¿Sabes cuánto **pago de alquiler** por la frutería? - No, ¿cuánto? - **Pimientos** euros.*

In this instance, it is observed how the model was able to infer *alquiler* (rent), and connected in to the word "euros" which corresponds to the word *Pimientos* (bell peppers). The intended meaning is to convey *Pimientos* to *quinientos* (five hundred) in relation to a monetary amount, a very similar word to the word *pimientos*. This wordplay is easy to understand, unlike other types of puns that are presented in Table 3, like the sentence:

*Hey Jesús. - ¿Salió cara la **cena**? No, salió **mala**.*

To understand this pun, relate to religion, humans typically require prior knowledge about the subject. It can be challenging to identify, as it often involves cultural references. However, it is among the results with the highest probability of being a wordplay, which is correctly classified. As seen in the table, the confidence in the probability is lower than that of the English model, ranging between 0.80 percent in the probability that it will be a wordplay.

Table 3
Top positive cases for wordplay classification of the Spanish corpus

Pun	Probability
Los jugadores de baloncesto, famosos por su mal carácter, pillan muchos rebotes	0.82717
¿Sabes cuánto pago de alquiler por la frutería? - No, ¿cuánto? - Pimientos euros.	0.81454
Ejecutivo agresivo busca monedas antiguas para partirles la cara.	0.80890
Ejecutivo agresivo busca monedas antiguas para partirles la marca.	0.80807
Hey Jesús. - ¿Salió cara la cena? No, salió mala.	0.80444

In the case of the examples with lower probability, it can be noted that the presence of sentence structures where the word order is altered can make it more difficult to find the wordplay. This is evident in the example provided in Table 4:

*Un **Señor Ruiz**... que un **ruiseñor**.*

The use of *Señor* (Mr.) and then the surname *Ruiz* may create a pun that refers to a type of bird, the *ruiseñor* (nightingale bird), but the model was unable to recognize the relationship between these two statements, where only the order of the pun was altered. The other four sentences do not contain a pun, this shows that most of the non-wordplay text where correctly identify as such.

Table 4
Top negative cases for wordplay classification of the Spanish corpus

Pun	Probability
Mi nombre es Rob. Robo	0.05724
Un Señor Ruiz... que un loro.	0.08994
Mi nombre es Ladino. Soy ladrón.	0.10061
Un Señor Ruiz... que un ruiseñor.	0.10311
Mi nombre es Rob. Soy un ladrón.	0.11668

Task 3: Translation of puns

Translations posed an extremely interesting case due to the fact that translating a term that makes a play of words in English may be difficult to interpret in Spanish. For a translation to fulfill the same objective as its counterpart in Spanish, it is extremely important to pay attention to certain structures. For example, the use of proper names that refer to a particular characteristic can be difficult to understand when they are translated into Spanish.

Understanding these nuances often becomes the pivotal part that modifies the behavior of what one wants to translate but that on other occasions can play in favor of the translation by making sense in the other language. Without forgetting that normally these words are quite related to the internal structure of the word in both languages.

These observations can be seen in Table 5, where different results are presented. The first sentence "If there's one person you don't want to interrupt in the middle of a sentence it's a judge." although correctly translated to "*Si hay una persona que no quieres interrumpir en medio de una oración, es un juez.*" the pun is not maintained, resulting in a literal translation.

On the contrary, in the case of the sentence "I tried to learn how to drive a stick shift but couldn't locate the manual." that was correctly translated to "*Intenté aprender a conducir con una* ***caja de cambios manual*** *pero no pude encontrar el* ***manual****.",* maintained not only the sense, but also the funniness of the wordplay (as indicated in bold letters).

Table 5
Pun translation from English to Spanish

English	Spanish
If there's one person you don't want to interrupt in the middle of a sentence it's a judge.	Si hay una persona que no quieres interrumpir en medio de una oración, es un juez.
I tried to learn how to drive a stick shift but couldn't locate the manual.	Intenté aprender a conducir con una caja de cambios manual pero no pude encontrar el manual.
The bride's best friend is so proud, she's practically made of honor.	La mejor amiga de la novia está tan orgullosa que parece que fuera la madrina.
Some rappers are good but others are Ludacris.	Algunos raperos son buenos, pero otros son Ludacris.
When the fog burns off it won't be mist.	Cuando se disipe la niebla, no será vapor.

5. Conclusions

In conclusion, we can say that the approaches implemented to cope with the classification and translation of text with puns is not misguided, perhaps more work could be done to strengthen and further explore and improve the results in these tasks. Especially in the case of translation, as the aspect of translating while maintaining the original context of the language to another makes this type of task a very interesting one to delve into, and it calls for continued research to find better methods and techniques.

Moreover, the integration of more advanced language models, such as BLOOMZ & mT5, showcased the potential of the use of prompts to enhance the translation of puns, which are instructions carried out by humans but this itself influences the result obtained. Regarding the classification of wordplays, the utilization of wordplays with BERT-like showed good results but they could be improved with better fine-tuning and exploring different variations of BERT-like architectures. Overall, this study obtained good results that demonstrate the potential of transformer-based models and language models in the classification and translation of puns.

6. References

[1] Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER - CLEF-2023 track on Automatic Wordplay Analysis. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, Nicola Ferro (Eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023) J. Cohen (Ed.), Special issue: Digital Libraries, volume 39, 1996.

[2] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z., Schoelkopf, H., Tang, X., Radev, D. R., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2022). Crosslingual Generalization through Multitask Finetuning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.01786

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. En arXiv (Cornell University) (Vol. 30, pp. 5998-6008). Cornell University. https://arxiv.org/pdf/1706.03762v5

[4] Mahurkar, S., & Patil, R. (2020). LRG at SemEval-2020 Task 7: Assessing the Ability of BERT and Derivative Models to Perform Short-Edits Based Humor Grading. https://doi.org/10.18653/v1/2020.semeval-1.108

[5] Victor Manuel Palma Preciado, Grigori Sidorov, Carolina Palma Preciado Assessing Wordplay-Pun classification from JOKER dataset with pretrained BERT humorous models, JokeR: Automatic Wordplay and Humour Translation, pages (1828-1833), CLEF (2022)

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Retrieved from http://arxiv.org/abs/1810.04805

[7] Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? https://doi.org/10.18653/v1/p19-1493

[8] Arun S. Maiya (2020). ktrain: A Low-Code Library for Augmented Machine Learning. arXiv preprint arXiv:2004.10703. BigScience Workshop. (2022). BLOOM (Revision 4ab0472)

[9] Inference API - Hugging Face. (s. f.). https://huggingface.co/inference-apiCañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., & Pérez, J. (2020). Spanish PreTrained BERT Model and Evaluation Data. In PML4DC at ICLR 2020.

[10] Borzunov, A., Baranchuk, D., Dettmers, T., Ryabinin, M., Belkada, Y., Chumachenko, A., Samygin, P., & Raffel, C. (2022). Petals: Collaborative Inference and Fine-tuning of Large Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2209.01188

[11] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA. workshop paper at PML4DC, ICLR 2020. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf