

# Comparative Evaluation of Humour Translation from English to Spanish: A Study with BLOOM and Googletrans

Notebook for the JOKER Lab at CLEF 2024

Olga Popova<sup>1,\*</sup>

<sup>1</sup>University of Cadiz, 9 Paseo Carlos III St., Cadiz, 11003, Spain

## Abstract

The problem of accurately translating wordplay in automatic humour analysis remains challenging, as highlighted by the CLEF 2024 JOKER Track. This problem is particularly interesting because humour is deeply cultural and context-dependent, making it difficult for language models to handle. Our study compares last year's and this year's results to determine if there have been improvements in the automatic translation of wordplay using BLOOM and Googletrans. The findings indicate that while numerical metrics show high precision, recall, and F1 scores, the actual translation and conveyance of jokes' meanings still rely heavily on coincidence, underscoring the need for further training and enhancement of language models.

## Keywords

Translation, pun translation, automatic translation, CLEF2024, JOKER

## 1. Introduction

According to the Overview of the CLEF 2024 JOKER Track (Automatic Humour Analysis), this Lab was established in 2022 at the Conference and Labs of the Evaluation Forum (CLEF) [1]. The author of this article is participating in the tasks of this laboratory for the second consecutive year [2]. In this way, this article will attempt to compare last year's results with this year's results to see if the automatic translation of wordplay using language models and automatic translators has improved.

Before sharing the methodology and the results of this year's runs in comparison with last year's, we will describe the JOKER tasks for this year, which include three different tasks:

- Task 1: Humour-aware information retrieval
- Task 2: Humour classification according to genre and technique
- Task 3: Translation of puns from English to French

This year, task 3 includes translation only from English to French, whereas in CLEF 2023, it also included Spanish. However, despite the exclusion of Spanish this year, the organizers of JOKER Track at CLEF 2024 allowed the author of this article to carry out the task using the Spanish language. Therefore, these working notes will focus solely on describing task 3 of JOKER Track.

The overall objective of this paper is to translate puns from English to Spanish using BLOOM and Google Translate. BLOOM is the only language model used, as the previous study found that it produced better results than all other models, including GPT, Simple T5, EasyNMT-Opus, and EasyNMT-mbart [2].

These working notes are organized as follows: after the introduction is the Experimental Setup section dedicated to the approach (data and method description); the third section is Experimental Results that includes some tables with metrics and scores, comparing the results of the task 3 of CLEF 2023 and CLEF 2024; and the fourth section is the discussion of the results and conclusions.

---

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

\*Corresponding author.

✉ olga.popova@uca.es (O. Popova)

🆔 0000-0001-7084-3140 (O. Popova)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 2. Experimental Setup

In this section we will give a brief description of the data provided to perform task 3, the methods used to solve it and the evaluation methods.

### 2.1. Data description

The task will provide an updated test set of English punning jokes, for French language the training data is also available. However, considering that the training data for Spanish is not needed for this experiment, we only use the test data. The test data is provided in JSON format.

**Table 1**  
Data size for task 3

entries	
test data (jokes) in English	5727

As we can see, we were provided with 5,727 jokes in English to translate. The data itself looked as follows:

**Table 2**  
Test data for pun translation

index	id_en	text_en
0	en_8	“I find you guilty,” said the judge with conviction.
1	en_9	The student had such a big assignment, he had to burn his kindle at both ends.
2	en_10	Herbivores come in browns and graze.
3	en_11	I’ll corroborate that again, Tom reproved.
4	en_13	I used to do rock climbing as a youth, but I was much boulder back then.
5	en_14	She dumped him because of all their lousy dates. After all, whining and dining does get tiresome after a while.

Each entry in the table contains an identifying number for each joke, and a one- or two-sentence joke that includes a pun.

### 2.2. Method description

As mentioned in the introduction, to carry out task 3, we used BLOOM with two different prompts and Google Translate.

#### 2.2.1. BLOOM

BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) is a transformer-based language model [3]. Because the number of tokens is limited in BLOOM, each run was performed with only 100 puns.

The prompts used are provided below.

- Prompt 1:

```
"Original: Diabetics should not be allowed to have sweet dreams.\n\nTranslation: Los diabéticos no deberian tener dulces sueños.\n\n\n"
```

```
Original: I’m going to the guillotine at dawn and my wife has already collected my severance pay.\n\n"
```

Translation: Al amanecer me van a pasar por la guillotina y mi mujer ya ha firmado la separación.\n\  
 \n\  
 Original: After 5 years with the same chiropractor, I moved and had to change doctors. It was quite an adjustment.\n\  
 Translation: Me mudé y tuve que buscar otros médicos después de estar cinco años con el mismo quiropráctico. Fue un mero ajuste.\n\  
 \n\  
 Original: A scientist doing a large experiment with liquid chemicals was trying to solve a problem when he fell in and became part of the solution.\n\  
 Translation: Un científico que hacía un gran experimento con productos químicos líquidos estaba intentando solucionar un problema cuando cayó en que él se convertiría en parte de la solución.\n\  
 \n\  
 Original: Old electricians never die, they just keep plugging away.\n\  
 Translation:"

- Prompt 2:

"Request: Translate from English to Spanish: Diabetics should not be allowed to have sweet dreams.\n\  
 Answer: Los diabéticos no deberían permitirse soñar dulce (expresión figurativa para decir que no deberían permitirse deseos excesivos o indulgencias).\n\  
 \n\  
 Request: I'm going to the guillotine at dawn and my wife has already collected my severance pay.\n\  
 Answer: Me voy a la guillotina al amanecer y mi esposa ya ha recogido mi indemnización (expresión figurativa para decir que me estoy preparando para un evento desafortunado y mi esposa ya ha preparado los asuntos financieros para el futuro sin mí).\n\  
 \n\  
 Request: After 5 years with the same chiropractor, I moved and had to change doctors. It was quite an adjustment.\n\  
 Answer:"

It is curious that when we created the second prompt, BLOOM itself, along with the Spanish translation in brackets, provided an explanation of the wordplay:

- 'expresión figurativa para decir que no deberían permitirse deseos excesivos o indulgencias' = 'figurative expression to say that excessive desires or indulgences should not be allowed'
- 'expresión figurativa para decir que me estoy preparando para un evento desafortunado y mi esposa ya ha preparado los asuntos financieros para el futuro sin mí' = 'figurative expression to say that I am preparing for an unfortunate event and my wife has already prepared the financial affairs for the future without me'

Thus, it appears that BLOOM has attempted to apply a sort of translator's explanatory note.

### 2.2.2. Googletrans

Googletrans is a free and unlimited python library that implements Google Translate API [4]. Since we use Google Colab for code execution, it was absolutely impossible to process all of them at once with over 5000 examples. Therefore, we were compelled to divide the test data into four parts.

### 2.3. Evaluation methods description

The task organizers used two methods for evaluating translations: BLEU and BERT Score [1].

BLEU (BiLingual Evaluation Understudy) measures the lexical similarity between a candidate translation and a reference translation [5]. The organizers utilized the sacreBLEU implementation<sup>9</sup> with the default tokenizer *13a*, which emulates the mteval-v13a script from Moses [6]. Their report includes the BLEU score (harmonic mean) and BLEU precisions for n-grams.

BERT Score, obtained from the python *bert-score* package<sup>10</sup> [7], presents mean values of precision, recall, and F1 scores.

### 3. Experimental Results

We divided the results section into two parts: numerical results and linguistic results. In the first subsection, we examine the evaluation metrics provided by the organizers, while in the second, we discuss some examples of translations produced using each method.

#### 3.1. Numerical results

Above all, we retrieved the table of results from CLEF 2023 for task 3 with BLOOM and Googletrans.

**Table 3**

Evaluation of task 3 CLEF 2023 with BLEU

run_id	count	BLEU	BLEU_1	BLEU_2	BLEU_3	BLEU_4
BLOOM	5.0	24.49	39.36	28.09	21.43	15.19
Googletrans	215.0	51.38	70.58	55.09	46.097	38.94

**Table 4**

Evaluation of task 3 CLEF 2023 with BERT score

run_id	count	BERT_score_P	BERT_score_R	BERT_score_F1
BLOOM	8.0	0.74	0.82	0.779
Googletrans	644.0	0.86	0.86	0.86

**Table 5**

Evaluation of task 3 CLEF 2024 with BLEU

run_id	count	BLEU	BLEU_1	BLEU_2	BLEU_3	BLEU_4
BLOOM_1	5.0	24.49	39.36	28.09	21.43	15.19
BLOOM_2	5.0	28.25	41.98	32.89	25.35	18.18
Googletrans	215.0	51.199	70.62	55.04	45.96	38.72

**Table 6**

Evaluation of task 3 CLEF 2023 with BERT score

run_id	count	BERT_score_P	BERT_score_R	BERT_score_F1
BLOOM_1	8.0	0.74	0.82	0.779
BLOOM_2	8.0	0.76	0.83	0.79
Googletrans	644.0	0.86	0.86	0.86

The results for BLOOM using prompt 1 remain the same for both BLEU and BERT score in CLEF 2023 and CLEF 2024. Googletrans shows superior metrics, but this advantage should be viewed with caution due to the larger number of evaluations. BLOOM\_2 performs slightly better than BLOOM\_1 in the few evaluations conducted, but both methods need more evaluations to provide a more reliable analysis. For a fairer and more accurate comparison, the number of evaluated examples should be equalized for all methods. With a larger amount of data, we might see a reduction in the performance gap between Googletrans and the BLOOM methods. Similarly, we can observe that there has been a slight change in the metrics performed by BLEU for Googletrans when comparing CLEF 2023 and CLEF 2024, this change should be analyzed qualitatively by comparing the translations.

### 3.2. Linguistic results

Next, we will analyze some examples of the translations produced by BLOOM and Googletrans in 2023 and 2024.

**Table 7**  
Googletrans 2023 and 2024 translations

Original text	Googletrans 2023	Googletrans 2024
Some rappers are good but others are <i>Ludacris</i> .	Algunos raperos son buenos, pero otros son <i>ludacris</i> .	Algunos raperos son buenos, pero otros son <i>Ludacris</i> .
I've got to fix the automobile, said Tom mechanically.	Tengo que arreglar el automóvil, dijo Tom mecánicamente.	Tengo que arreglar el automóvil, dijo Tom mecánicamente.
Piano players know what bar they are in.	Los pianistas saben en qué bar se encuentran.	Los pianistas saben en qué bar se encuentran.

If we compare the translations produced by Googletrans in 2023 and 2024, we can observe that the vast majority of jokes are translated identically. In Example 1 of Table 7, the only difference is in the use of capitalization. We can infer that the differences in BLEU metrics are due to the fact that this year's test data does not include all the same jokes as last year's.

**Table 8**  
BLOOM 2023 and 2024 translations

Original text	BLOOM 2023	BLOOM 2024_1	BLOOM 2024_2
OLD POLICEMEN never die they just <i>cop out</i> .	Los viejos policías nunca mueren, <i>sólo se retiran</i> .	Los viejos policías nunca mueren, <i>sólo se retiran</i> .	Los viejos policías nunca mueren, <i>solo se jubilan</i> .
The boy swallowed a pillow, the hospital described <i>his condition as comfortable</i> .	El niño se tragó un cojín, el hospital describió <i>su estado como cómodo</i> .	El niño se tragó un cojín, el hospital describió <i>su estado como cómodo</i> .	El niño se tragó un cojín, el hospital describió <i>su condición como cómoda</i> .
I've never taken an elevator to the basement floor, <i>that's just beneath me</i> .	Nunca he cogido un ascensor hasta el sótano, <i>eso está por debajo de mí</i> .	Nunca he cogido un ascensor hasta el sótano, <i>eso está por debajo de mí</i> .	Nunca he tomado un ascensor al sótano, <i>eso está justo debajo de mí</i> .

Regarding the translations done by BLOOM, the outputs from 2023 and this year using prompt 1 are identical. In some cases, the joke is translated understandably, while in others, it is not. Concerning the translations with prompt 2, we observe that the target texts are more literal than those from prompt 1. While BLOOM\_1 attempts to adapt the vocabulary to make it more natural, BLOOM\_2 performs a highly literal translation.

## 4. Discussion and Conclusions

The numerical results, which are quite high in terms of precision, recall, and F1, suggest that the translations of the jokes are well-executed. However, when we analyze the results from a translational and linguistic perspective, we observe that the successful translation and conveyance of a joke's meaning are more coincidental. Although jokes are a cultural matter and usually require adaptation in translation, some jokes can be understood when translated literally, which explains some of the good translations in this task. By comparing the numerical and linguistic results, we can conclude that the more literal the translation, the higher the metrics. Therefore, we can conclude that language models still require significant training and improvement to accurately translate jokes in a "conscious" manner.

Future work could involve analyzing other language models, such as GPT, as it continues to improve and might offer different results. Additionally, exploring the translation of other types of jokes and wordplay, and attempting to train models with high-quality translations of jokes could be beneficial.

## 5. Acknowledgments

I would like to thank the organizers of CLEF 2024 in general and the organizers of JOKER in particular for providing us with the opportunity to continually improve our research, learn, and develop. Above all, I would like to express my gratitude to Professor Liana Ermakova for her invaluable advice and great support throughout the execution of this task.

This project has received a government grant managed by the National Research Agency under the program “*Investissements d’avenir*” integrated into France 2030, with the Reference ANR-19-GURE-0001.

## References

- [1] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of clef 2024 joker track on automatic humor analysis, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, LNCS, Springer-Verlag, 2024.
- [2] O. Popova, P. Dadić, Does ai have a sense of humor? clef 2023 joker tasks 1, 2 and 3: Using bloom, gpt, simplet5, and more for pun detection, location, interpretation and translation, in: *CLEF (Working Notes)*, 2023, pp. 1888–1908.
- [3] B. Workshop, Bloom (revision 4ab0472), 2022. URL: <https://huggingface.co/bigscience/bloom>. doi:10.57967/hf/0003.
- [4] Googletrans 3.0.0, ??? URL: <https://pypi.org/project/googletrans/>.
- [5] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: A method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. URL: <https://www.aclweb.org/anthology/P02-1040>. doi:10.3115/1073083.1073135.
- [6] M. Post, A call for clarity in reporting bleu scores, in: *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: <https://www.aclweb.org/anthology/W18-6319>.
- [7] T. Zhang, V. Kishore, F. Wu, K. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: *International Conference on Learning Representations*, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.