Overview of the CLEF 2024 JOKER Task 3: Translate puns from English to French

Liana Ermakova[1,*], Anne-Gwenn Bosser[2], Tristan Miller[3,4] and Adam Jatowt[5]

[1]Université de Bretagne Occidentale, HCTI, France
[2]École Nationale d'Ingénieurs de Brest, Lab-STICC CNRS UMR 6285, France
[3]Department of Computer Science, University of Manitoba, Winnipeg, Canada
[4]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[5]University of Innsbruck, Austria

Abstract

This paper provides a comprehensive overview of Task 3 of the CLEF 2024 JOKER track on automatic houmous analysis. The overarching objective of the JOKER track series is to facilitate collaboration among linguists, translators, and computer scientists to advance the development of automatic interpretation, generation, and translation of wordplay. Task 3 specifically concentrates on the automatic translation of puns from English into French. This overview outlines the overall structure of the shared task we organised as part of the CLEF 2024 evaluation campaign. We discuss the approaches employed by the participants and present and analyse the results they achieved. We also describe the work of participants who used our data to translate puns from English to Spanish as part of the open task of the track.

Keywords

wordplay, puns, computational humour, machine translation

1. Introduction

This paper describes Task 3 of the CLEF 2024 JOKER[1] challenge, where the goal is to accurately translate puns between different languages. This is the final task of JOKER-2024 [1], following Tasks 1 [2] and 2 [3] on humour-aware information retrieval and humour classification according to genre and technique, respectively. The overall objective of the JOKER track series [4, 5], which started in 2022, is to facilitate collaboration among linguists, translators, and computer scientists to advance the development of automatic interpretation, generation, and translation of wordplay. Thus, this is the third edition of the JOKER track, with the pun translation task being its oldest and ultimate challenge.

A pun is a form of wordplay that exploits multiple meanings of a word or words with similar sounds but different meanings. Puns pose challenges in translation, as they often rely on language-specific nuances that may not have direct equivalents in other languages. Nonetheless, it can be important to preserve wordplay in the target text, even if the exact type of wordplay or the specific meaning is changed. In Task 3, the goal is to translate English punning jokes into French. The translations should aim to preserve, to the extent possible, both the form and meaning of the original wordplay – that is, to implement the pun→pun strategy described in Delabastita's typology of pun translation strategies [6, 7]. For example, "I used to be a banker but I lost interest" might be rendered into French as "*J'ai été banquier mais j'en ai perdu tout l'intérêt*". This fairly straightforward translation preserves the pun, since *interest* and *intérêt* share the same double meaning.

In the previous editions of the JOKER track, we observed that the success rate of wordplay translation is extremely low even in the case of LLMs, for both language pairs: English-French and English-Spanish [8]. For example, the highest success rate of translations that preserved both the form and sense of the original wordplay in the manually evaluated CLEF 2023 JOKER test set was 6% for French, while

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France
*Corresponding author.
0000-0002-7598-7474 (L. Ermakova); 0000-0002-0442-2660 (A. Bosser); 0000-0002-0749-1100 (T. Miller); 0000-0001-7235-0665 (A. Jatowt)

[1]https://www.joker-project.com

Spanish achieved 18% [8]. The challenge of translating wordplay even between well-studied languages with the use of LLMs, highlights the need for increased community focus on this difficult task.

This year eleven teams submitted 23 runs for Task 3 showing stable interest of the community in the pun translation task. Note that six submissions were made by translators rather than computer scientists (teams **Olga** and **UBO**).

In the following sections, we describe the data preparation process (Section 2) and participants' approaches (Section 3), and then present an analysis of their results (Section 4). In addition to the traditional machine translation evaluation measures, such as BLEU [9] and BERT Score [10], we examined the participants' performances using the dataset we created to identify words or phrases that have multiple meanings (pun locations) for the CLEF 2023 JOKER Task 2 [11, 4, 12]. Besides the official results of the translation into French, we describe the results of a team who submitted translations into Spanish which we compare with last year's participants' results. Section 5 concludes the paper.

2. Data

The data is an extension of the JOKER parallel wordplay corpus [12]. The training data for Task 3 consists of 1,405 English wordplay instances, with a total of 5,838 professional human French translations. The maximum number of reference translations per English pun is 29. 72% of English puns have multiple reference translations. The histogram of the test references per English pun is shown in Figure 1.

The test files shared with participants consist of 4,501 English wordplay instances, in JSON format. Over these English puns, we used new 376 distinct source texts with 832 corresponding reference translations created by professional French native speaker translators. The maximum number of reference translations per English pun is eight. However, the majority of source texts have a single reference translation. The histogram of the test references per English pun is given in Figure 2. An example of the source data is as follows:


```
{
  "id_en":   "en_1007",
  "text_en": "Save the whales, spouted Tom."
}
```


The corresponding human reference translations are as follows:


```
{
  "text_fr": "\"Il faut sauver les baleines\", jeta Tom avant de se tasser."
},
{
  "text_fr": "\"Il faut sauver les baleines\", interjeta Tom."
},
{
  "text_fr": "Moi je sauve les baleines, Tom s'en venta."
},
{
  "text_fr":"Louis évent-a le projet de sauvetage des baleines."
},
{
  "text_fr":"\"Sauvez les baleines\", proclama Tom à tout évent."
},
{
  "text_fr":"\"Sauvez les baleines, cracha Toto, Cétacé!\""
}
```

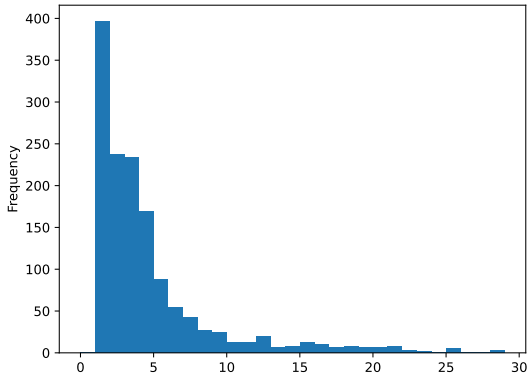

Figure 1: Histogram of translation references in French per English pun (train data)

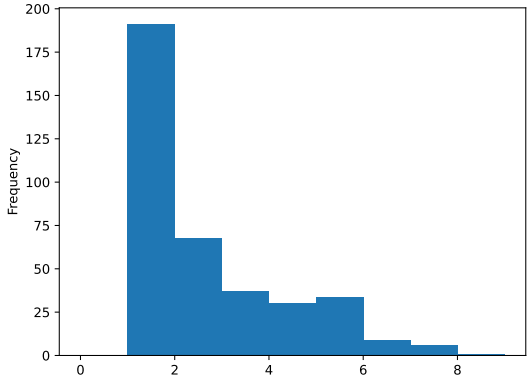

Figure 2: Histogram of translation references in French per English pun (test data)

3. Participants' approaches

Ten teams submitted 20 official runs for this task. In addition, the team **Olga** [13] used the data from this task to explore translation from English to Spanish as part of the open task. She submitted three additional runs. Statistics on the runs are summarised in Table 1. The approaches used were as follows:

The **UAms** team [14] submitted two runs. MarianMT – a sequence-to-sequence (`Seq2Seq`) model based on the Marian framework was used. The second run was based on the T5 (`t5-base`) model with the same standard preprocessing as for the first run.

The **Arampatzis** team submitted six runs for this task employing (among others) MarianMT, Google Translate, Helsinki-NLP Opus, and mBART.

The **Petra&Regina** team [15] submitted a single run. The authors relied on the EasyNMT library, which they used with the Helsinki-NLP Opus-MT model.

The **Tomislav&Rowan** team [16] preprocessed the data and used it to build prompts to translate the jokes with the translation pre-trained model (Helsinki-NLP Opus) through the MarianMT framework. The authors judged that EasyNMT was less effective for this task. Two runs were submitted.

The **Farhan** [17] team provided two runs. They used single-shot prompting techniques with GPT-4 and GPT-4o.

The **Frane** team submitted one run. They used neural machine translation models like MarianMT. The translations were refined with a custom module to preserve the pun elements. This module used bilingual dictionaries and contextual embeddings. A similar approach was taken by the **Dajana&Kathy**

Table 1
Statistics on the runs submitted for Task 3

Team	# of submitted runs
UAms	2
Arampatziz	6
AB&DPV	1
Petra&Regina	1
Tomislav&Rowan	2
Farhan	2
Frane	1
Dajana&Kathy	1
Jokester	1
UBO	3
Olga	3
Total	23

team.

The **AB&DPV** team [18] used simple prompts with Llama-2-7b and reported that in a number of instances the translations were found to be incomplete or mixing two languages. They submitted a single run.

The **Jokester** [19] team submitted one run. They also used the MarianMT framework.

The **UBO** team submitted three semi-manual runs using commercial models such as DeepL, Google Translate, and ChatGPT.

The **Olga** team [13] submitted three runs as part of the open task. The team explored the topic of translating humour from English to Spanish, comparing the BLOOM model with Google Translate. For the BLOOM translations, two different prompts were employed. We provide a comparison of her runs with the CLEF JOKER 2023 participants [4, 8] as this year we did not have a shared task on translation into Spanish.

4. Results

4.1. General results

Tables 2 and 3 show the results on the test data while Tables 4 and 5 display the results obtained on the training data for the pun translation task from English into French.

We evaluated the runs with the following machine translation metrics:

BLEU (BiLingual Evaluation Understudy), which measures the vocabulary overlap between the candidate translation and a reference translation [9]. We used the sacreBLEU implementation[2] with the default tokeniser *13a* which mimics the mteval-v13a script from Moses [20]. We report the BLEU score (harmonic mean) and the BLEU precisions for n-grams on 376 distinct English texts with corresponding 832 reference translations to French.

BERT Score from the Python bert-score package[3] [10]. We report mean values of BERT score precision, recall, and F_1 over all 832 references.

We make the following observations. First, the best results were obtained by participants who used the commercially produced machine translation engines (such as Google Translate and DeepL) integrated into SDL studio. The MarianMT models, which are similar to BART, showed very similar results. Second, the same models fine-tuned by different teams achieved different scores. Third, the

[2] https://github.com/mjpost/sacrebleu/
[3] https://pypi.org/project/bert-score/

Table 2
BLEU results for pun translation from English into French (test data)

run ID	count	BLEU	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Arampatzis_GoogleTranslate	376	65.23	78.96	67.48	61.59	57.52
Frane_TranslationModel	92	57.13	64.33	58.41	54.66	51.85
Dajana&Kathy_TranslationModel	376	58.45	71.94	60.27	54.11	49.73
UBO_SDL	312	13.17	71.90	57.17	49.13	43.24
Tomislav&Rowan_MarianMTModel	376	58.85	77.11	63.66	56.06	50.45
Arampatzis_MarianMT	376	58.85	77.11	63.66	56.06	50.45
UBO_ChatGPT	312	13.09	69.90	54.08	46.07	40.31
UBO_DeepL	312	11.97	68.53	50.32	41.38	35.11
UAms_T5-base_ft	376	48.74	71.75	54.57	45.18	38.05
Arampatzis_mBART	376	48.71	70.95	54.40	45.29	38.67
Arampatzis_M2M100	376	42.37	68.46	48.73	37.72	29.93
UAms_Marian_ft	376	25.69	47.05	28.47	20.74	15.69
Tomislav&Rowan_MarianMTModel	1	11.46	100.00	100.00	100.00	100.00
Farhan_2	376	14.33	23.68	15.84	12.05	9.32
Farhan_1	376	9.21	15.92	9.97	7.65	5.92
jokester_MarianMTModel	49	0.29	15.34	0.14	0.08	0.04
Arampatzis_opus_mt	63	0.29	15.04	0.23	0.06	0.03
Arampatzis_T5	63	0.32	11.35	0.17	0.10	0.06

Table 3
BERT score results (precision, recall, and F_1) for pun translation from English into French (test data)

run ID	count	P	R	F_1
Arampatzis_GoogleTranslate	832	91.93%	91.82%	91.85%
Frane_TranslationModel	279	92.06%	91.53%	91.77%
Dajana&Kathy_TranslationModel	832	91.35%	91.00%	91.15%
UBO_SDL	598	90.13%	90.21%	90.15%
Tomislav&Rowan_MarianMTModel	832	90.82%	89.19%	89.95%
Arampatzis_MarianMT	832	90.82%	89.19%	89.95%
UBO_ChatGPT	598	89.12%	89.34%	89.21%
UBO_DeepL	598	89.06%	89.31%	89.16%
UAms_T5-base_ft	832	89.53%	88.52%	89.00%
Arampatzis_mBART	832	88.95%	87.41%	88.13%
Arampatzis_M2M100	832	88.23%	87.23%	87.70%
UAms_Marian_ft	832	81.06%	82.53%	81.74%
Tomislav&Rowan_MarianMTModel	3	84.42%	71.23%	77.26%
Farhan_2	832	69.38%	77.14%	72.96%
Farhan_1	832	64.30%	73.18%	68.41%
jokester_MarianMTModel	112	67.30%	66.38%	66.80%
Arampatzis_opus_mt	157	66.98%	66.05%	66.47%
Arampatzis_T5	157	65.91%	64.79%	65.31%

BLEU scores of the UBO submission are very low, while the BERT scores are very high. More analysis is needed to investigate this difference.

4.2. Analysis of the presence of the punning words

For a fine-grained analysis of the generated translations, we decided to evaluate the translations based on the presence of words or phrases carrying multiple meanings (pun location) from the reference texts. This approach allows us to focus on specific elements of translation quality that standard evaluation metrics might overlook. Thus, for this analysis we used the data created within CLEF 2023 JOKER Task

Table 4
BLEU results for pun translation from English into French (training data)

run ID	count	BLEU	BLEU_1	BLEU_2	BLEU_3	BLEU_4
UAms_T5-base_ft	1,405	59.93	77.66	63.35	55.50	49.25
UAms_Marian_ft	1,405	68.56	77.50	70.09	65.84	61.79
Arampatzis_GoogleTranslate	1,405	42.19	67.50	46.29	35.76	28.37
Dajana&Kathy_TranslationModel	1,405	47.95	70.02	50.87	41.69	35.61
Arampatzis_MarianMT	1,405	48.55	70.52	51.47	42.50	36.71
Tomislav&Rowan_MarianMTModel	1,405	48.55	70.52	51.47	42.50	36.71
Arampatzis_M2M100	1,405	34.10	62.85	39.12	27.85	20.42
Arampatzis_mBART	1,405	33.93	62.38	38.66	27.73	20.26
Farhan_2	1,405	12.16	23.06	13.47	9.75	7.22
jokester_MarianMT	223	0.30	17.52	0.33	0.07	0.02
Arampatzis_opus_mt	229	0.32	17.42	0.40	0.07	0.02
Farhan_1	1,405	7.75	15.96	8.49	6.05	4.40
Arampatzis_T5	229	0.36	14.16	0.49	0.11	0.03

Table 5
BERT score results (precision, recall, and F_1) for pun translation from English into French (training data)

run ID	count	P	R	F_1
UAms_T5-base_ft	5,838	84.35%	83.33%	83.80%
UAms_Marian_ft	5,838	81.82%	82.84%	82.28%
Arampatzis_GoogleTranslate	5,838	82.36%	81.62%	81.96%
Dajana&Kathy_TranslationModel	5,838	81.98%	81.56%	81.73%
Arampatzis_MarianMT	5,838	82.16%	81.38%	81.72%
Tomislav&Rowan_MarianMTModel	5,838	82.16%	81.38%	81.72%
Arampatzis_M2M100	5,838	80.91%	79.90%	80.37%
Arampatzis_mBART	5,838	80.59%	80.01%	80.26%
Farhan_2	5,838	66.54%	72.97%	69.52%
jokester_MarianMT	945	67.37%	67.16%	67.24%
Arampatzis_opus_mt	956	66.78%	66.77%	66.74%
Farhan_1	5,838	62.20%	69.86%	65.76%
Arampatzis_T5	956	65.15%	64.48%	64.78%

2 – Pun Location and Interpretation [11, 4, 12].

Preserving wordplay is often crucial to maintain the sense of the text as in the pun from *Alice's Adventures in Wonderland by Lewis Carroll, : " 'That's the reason they're called lessons,' the Gryphon remarked: 'because they lessen from day to day.' "*. The official translation preserves wordplay by using the pair *cours/courts*: "*C'est pour cette raison qu'on les appelle des cours: parce qu'ils deviennent chaque jour un peu plus courts.*". While machine translation destroys the wordplay, resulting in the sentence becoming nonsensical: "*C'est la raison pour laquelle on les appelle leçons, remarqua le Griffon: parce qu'elles diminuent de jour en jour.*"

The train data contained 5,838 French translations of 1,405 distinct English puns. These translations had 4,355 distinct locations, i.e. words or phrases with multiple meanings. These locations were manually annotated by master's students specializing in translation who are native speakers of French. The location annotation was not shared with participants this year. However, in 2023 we released 2,000 annotations as train data [11, 4]. Each English pun has a maximum of 20 corresponding distinct locations in French while the maximal number of translations is 29. This means that multiple translations exploit the same location to create double meanings. The histogram of distinct locations in French per English pun for train data is given in Figure 3.

The overlap between the JOKER 2023 Task 2 pun location data and the test data of Task 3 this year reveals 13 distinct locations for 8 unique English puns.

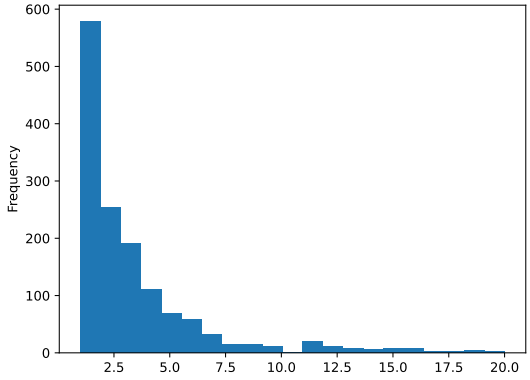

Figure 3: Histogram of distinct pun locations in French per English pun (train data)

For each English pun, we computed a boolean value that is *True* if the corresponding French translation contains at least one word from the set of locations corresponding to that English pun. We considered only exact string matches. The participants' results in terms of translations containing locations are presented in Table 6. As the overlap between the JOKER 2023 Task 2 pun location data and the test data of Task 3 of this year is small, the results are close to zero and are not entirely interpretable. However, on the training data, we observed similar results between the non-fine-tuned MarianMT and Google Translate. The T5 and MarianMT models fine-tuned by **UAms** [14] on our data show improved overlap with the punning words in the reference translations. The fine-tuned MarianMT [14] includes nearly twice as many translations containing location terms compared to T5. The fine-tuned MarianMT [14] achieves a higher BLEU score than T5, whereas the BERT Score shows the opposite trend. The BLEU score of both GPT-based runs submitted by the **Farhan** team [17] is notably low, as are the BERT Scores. However, one of the GPT-based runs has a higher number of locations than Google Translate, while another is close to the encoder-decoder model M2M100. In general, only a small percentage of translations contain at least one word identified as carrying multiple meanings in references. Models, fine-tuned on our training data achieve a maximum of 23% of translations containing at least one pun location word from reference translations. In contrast, non-fine-tuned models use pun location words in only 11% of cases. These results closely mirror those obtained last year [8, 4]. According to manual evaluation of the JOKER 2023 participants' runs, the highest success rate for preserving both the form and sense of the original wordplay in translations from English to French was 6% over the total evaluated test set. For the training set, the percentage of successful translations was less than 17%. These observations could open up new perspectives for the evaluation of machine translation in handling wordplay.

4.3. Translation into Spanish

The team **Olga** [13] submitted three additional runs providing translations of English puns into Spanish. The translations into Spanish were generated by the commercial translation engine Google Translate and the LLM BLOOM with few-shot prompting. Both BLOOM runs were partial.

We provide the evaluation of her submissions in comparison with last year's participants' results on the train data as we have not constructed new references in Spanish this year. However, as the models used by **Olga** have not been trained on our data, last year's train references are not problematic. For this evaluation, we used 215 puns in English with 644 corresponding references.

Tables 7 and 8 display the results obtained on the CLEF 2023 JOKER training data for the participants of both 2023 and 2024. For the details of the previous year's runs refer to the corresponding JOKER overview papers [4, 8] and participants' working notes respectively.

Among the runs submitted in 2023, those using mBART, OpusMT, and Google Translate produced

Table 6
Presence of identified punning words (locations) in generated translations

	Training data			Test data		
run ID	Total	# Location	%	Total	# Location	%
UAms _Marian _ft	1,405	317	23%	8	0	0%
UAms _T5-base _ft	1,405	179	13%	8	0	0%
Dajana&Kathy _TranslationModel	1,405	158	11%	8	1	13%
Tomislav&Rowan _MarianMTModel	1,405	157	11%	8	1	13%
Arampatzis _MarianMT	1,405	157	11%	8	1	13%
Farhan _2	1,405	143	10%	8	0	0%
Arampatzis _GoogleTranslate	1,405	141	10%	8	1	13%
Arampatzis _mBART	1,405	121	9%	8	1	13%
Arampatzis _M2M100	1,405	115	8%	8	0	0%
Farhan _1	1,405	106	8%	8	0	0%
Arampatzis _T5	229	0	0%	2	0	0%
Arampatzis _opus _mt	229	0	0%	2	0	0%
jokester _MarianMTModel	223	0	0%	2	0	0%

Table 7
BLEU results for pun translation from English into Spanish (training data)

run ID	count	BLEU	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Olga_ES_BLOOM_1	5	24.49	39.36	28.09	21.43	15.19
Olga_ES_Googletranslator	215	51.20	70.62	55.04	45.96	38.72
Olga_ES_BLOOM_2	5	28.25	41.98	32.89	25.35	18.18
LJGG_es_mt5_base_auto	215	40.14	60.67	45.30	38.19	32.18
LJGG_es_t5_large_no_label_auto	215	47.90	68.25	51.90	42.81	35.52
LJGG_Google_Translator_EN_ES_auto	209	52.26	71.88	56.22	47.04	39.77
LJGG_es_mt5_base_no_label_auto	215	37.93	61.75	45.00	35.72	28.58
LJGG_es_t5_large_auto	11	0.76	14.15	0.53	0.30	0.17
TheLangVerse_j2-grande-finetuned	215	38.81	63.33	43.31	32.82	25.19
Smroltra_EN-ES_GPT3	5	46.15	74.07	53.06	40.91	28.21
Smroltra_EN-ES_BLOOM	5	24.49	39.36	28.09	21.43	15.19
Smroltra_EN-ES_GoogleTranslation	215	51.38	70.58	55.09	46.10	38.94
Smroltra_EN-ES_EasyNMT-Opus	215	53.95	71.86	57.55	49.08	42.48
Smroltra_EN-ES_SimpleT5	215	25.76	53.68	29.74	19.73	13.97
Smroltra_EN-ES_EasyNMT-mbart	215	36.72	62.01	41.32	30.81	23.03
Croland_EN_ES_GPT3	3	25.78	46.67	29.63	25.00	19.05
ThePunDetectives_EN-ES_OpusMT	65	54.18	73.58	58.06	50.00	42.61
ThePunDetectives_EN-ES_M2M100	65	39.67	65.51	43.15	33.29	26.33

the best results for Spanish according to our manual evaluation in terms of the number of successful translations – i.e., translations preserving, to the extent possible, both the form and sense of the original wordplay [4, 8]. The non-fine-tuned OpusMT and Google Translate showed the best results, with BLEU scores exceeding 50% and BLUE_1 going up to 74%. The BERT score for all models is very high, with an F1 score always exceeding 77%. However, the number of successful translations evaluated manually was a maximum of 18% [4, 8]. These results suggest that further research is needed in the field of machine translation evaluation measures and the development of more sophisticated and reliable measures is essential.

Table 8
BERT score results (precision, recall, and F_1) for pun translation from English into Spanish (training data)

run ID	count	P	R	F_1
Olga_ES_BLOOM_1	8	74.36%	81.92%	77.94%
Olga_ES_Googletranslator	644	86.26%	85.93%	86.07%
Olga_ES_BLOOM_2	8	75.96%	83.13%	79.36%
LJGG_es_mt5_base_auto	644	83.10%	81.46%	82.24%
LJGG_es_t5_large_no_label_auto	644	85.61%	85.05%	85.30%
LJGG_Google_Translator_EN_ES_auto	626	86.81%	86.40%	86.59%
LJGG_es_mt5_base_no_label_auto	644	83.74%	81.14%	82.37%
LJGG_es_t5_large_auto	29	79.00%	76.69%	77.81%
TheLangVerse_j2-grande-finetuned	644	84.66%	84.43%	84.52%
Smroltra_EN-ES_GPT3	8	91.01%	90.23%	90.62%
Smroltra_EN-ES_BLOOM	8	74.37%	81.93%	77.95%
Smroltra_EN-ES_GoogleTranslation	644	86.27%	85.96%	86.10%
Smroltra_EN-ES_EasyNMT-Opus	644	86.31%	86.14%	86.21%
Smroltra_EN-ES_SimpleT5	644	81.25%	80.64%	80.92%
Smroltra_EN-ES_EasyNMT-mbart	644	84.04%	83.94%	83.97%
Croland_EN_ES_GPT3	4	77.58%	80.97%	79.21%
ThePunDetectives_EN-ES_OpusMT	185	86.07%	85.74%	85.88%
ThePunDetectives_EN-ES_M2M100	185	84.61%	83.72%	84.14%

5. Conclusion

In this paper, we have described Task 3 of the JOKER track at CLEF 2024. The task aims to advance the automation of wordplay translation. The EN→FR parallel JOKER corpus we used in the previous edition [12, 4] contained 5,838 French translations of 1,405 distinct English puns. This year, we expanded the corpus by introducing 376 new distinct source texts with 832 corresponding reference translations created by professional French native-speaker translators.

This year, eleven teams submitted 23 runs for Task 3, demonstrating a stable interest from the community in the pun translation task. Participants mainly used LLMs, commercial machine translation engines, and out-of-the-box translation models.

We evaluated the participants' results using automated measures, specifically BLEU and BERT scores. According to these automatic measures on the test data, the best results were achieved by commercial machine translation models, with Google Translate producing the top results. While on the training data, the fine-tuned models largely outperform Google Translate. The BLEU scores showed a lot of variation across the runs, while the BERT scores produced results that could be grouped into two distinct levels.

To conduct a fine-grained analysis of the generated translations, we evaluated them based on the presence of words or phrases with multiple meanings (pun locations) from the reference texts. To this end, we combined the reference translations in French with pun location annotations from the data created for CLEF 2023 JOKER Task 2 – Pun Location and Interpretation [4, 11]. In general, only a small percentage of translations contain at least one word identified as carrying multiple meanings from the references, despite high BLEU and BERT Scores. For example, Google Translate achieved an 82% BERT Score and 42 BLEU Score, while it shares only 10% of punning words with the references. Fine-tuned models on our training data achieve up to 23% of translations containing at least one pun location word from reference translations, whereas non-fine-tuned models incorporate pun location words in only 11% of cases. These findings closely align with those from the previous year [8, 4]. These observations suggest potential new perspectives for evaluating machine translation performance in handling wordplay.

We observe that the success rate of wordplay translation remains extremely low, even in the case of LLMs. However, one of the major obstacles in the development of wordplay machine translation is its

evaluation. Destroying the wordplay may result in the text becoming nonsensical. The existing metrics do not take into account punning words which can reward translations with completely lost sense. In future work, we will explore new perspectives on evaluating wordplay in machine translation based on the data constructed within the JOKER track.

Additional information on the track is available on the JOKER website: https://www.joker-project.com/

Acknowledgments

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. We thank all other colleagues and students who participated in data construction, the translation contests, and the CLEF JOKER track.

References

[1] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of CLEF 2024 JOKER track on automatic humor analysis, in: L. Goeuriot, G. Q. Philippe Mulhem, 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), Lecture Notes in Computer Science, Springer, 2024.

[2] L. Ermakova, A.-G. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER Task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] V. M. Palma Preciado, G. Sidorov, L. Ermakova, A.-G. Bosser, A. Jatowt, Overview of the CLEF 2024 JOKER Task 2: Humour classification according to genre and technique, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[4] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER – CLEF-2023 track on automatic wordplay analysis, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14163, Springer Nature Switzerland, Cham, 2023, pp. 397–415. doi:10.1007/978-3-031-42448-9_26.

[5] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, E. Mathurin, G. L. Corre, S. Araújo, J. Boccou, A. Digue, A. Damoy, 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), volume 13390 of *Lecture Notes in Computer Science*, 2022, pp. 447–469.

[6] D. Delabastita, There's a Double Tongue: an Investigation into the Translation of Shakespeare's Wordplay, with Special Reference to Hamlet, Rodopi, Amsterdam, 1993.

[7] D. Delabastita, Wordplay as a translation problem: a linguistic perspective, in: Ein internationales Handbuch zur Übersetzungsforschung, volume 1, De Gruyter Mouton, 2008, pp. 600–606. doi:10.1515/9783110137088.1.6.600.

[8] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 3 – pun translation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1818–1827.

[9] 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, 2002, pp. 311–318. URL: https://www.aclweb.org/anthology/P02-1040. doi:10.3115/1073083.1073135.

[10] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[11] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 2 – pun location and interpretation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1804–1817.

[12] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER Corpus: English–French parallel data for multilingual wordplay recognition, in: SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.

[13] O. Popova, Comparative Evaluation of Humour Translation from English to Spanish: A Study with BLOOM and Googletrans, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[14] L. Buijs, M. Cazemier, E. Schuurman, J. Kamps, University of Amsterdam at the CLEF 2024 Joker Track, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[15] R. Elagina, P. Vučić, Convergential approach in machine learning for effective humour analysis and translation, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[16] R. Mann, T. Mikulandric, CLEF 2024 JOKER Tasks 1–3: Humour identification and classification, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[17] F. Dhanani, R. Abbas, Translating English puns to French, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[18] D. P. Varadi, A. Bartulović, JOKER 2024 by AB&DPV: From 'LOL' to 'MDR' using AI models to retrieve and translate puns, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[19] H. Baguian, H. N. Ashley, JOKER Track @ CLEF 2024: The Jokesters' approaches for retrieving, classifying, and translating wordplay, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[20] 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.