REC_Cryptix at JOKER CLEF 2025: Teaching Machines to Laugh: Multilingual Humor Detection and Translation[*]

Notebook for the Joker Lab at CLEF 2025

Sarath Kumar P[1,*,†], Beulah A[1,†], Sushmitha M[1,†] and Thanalaxmi S[1,†]

[1]*Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India.*

Abstract

Humor is a cognitively complex and culturally specific linguistic phenomenon, posing significant challenges for computational modeling. It often transcends formal grammar, exploits phonetic and semantic ambiguities, and varies widely across languages and cultures. The JokerCLEF 2025 challenge addresses this problem through three distinct sub-tasks: Humour-Aware Information Retrieval, Wordplay Translation, and Onomastic Wordplay Translation. This paper presents the system developed by Team Cryptix, which participated in all three tasks. We employed a combination of state-of-the-art models—SBERT, MarianMT, and T5—chosen for their capabilities in semantic representation, multilingual processing, and text generation. Our approach includes task-specific fine-tuning, targeted feature engineering, and the integration of human-in-the-loop evaluations to refine output quality. We also describe the datasets, preprocessing steps, and evaluation strategies used for each sub-task. Empirical results show that our system demonstrates consistent performance in both retrieval and translation of humorous content, maintaining the intent, nuance, and cultural relevance of the original text. Our findings underscore the importance of combining semantic-aware models with culturally informed design when handling humor in multilingual settings.

Keywords

Humor Detection, Information retrieval, Machine Translation, Multilingual translation, Natural Language Processing, Onomastics, Wordplay

1. Introduction

Humor is a complex and culturally dependent aspect of human language, making it particularly challenging for computational systems to process. The CLEF 2025 Joker [1] shared task aims to push the boundaries of humor-aware natural language processing through three subtasks: Humour-aware Information Retrieval, which focuses on retrieving humor-relevant documents for a given query; Wordplay Translation, involving the translation of puns and linguistic humor; and Onomastic Wordplay Translation, which targets humor derived from names and culturally specific references. These tasks are particularly challenging due to the semantic ambiguity of humor, shifts in cultural context, phonetic differences across languages, and the scarcity of explicitly annotated humorous datasets.

The exploration of deep learning techniques for multilingual natural language processing (NLP) tasks has gained significant momentum, particularly in areas such as intent classification, hate speech detection, and offensive language identification across diverse languages [2], [3], [4], [5]. These studies demonstrate the effectiveness of transfer learning models, especially multilingual transformers like mBERT and XLM-RoBERTa, in handling low-resource languages and code-mixed data, highlighting the importance of multilingual representations in cross-lingual tasks [4].

In the context of data augmentation, techniques such as back translation and paraphrasing have been employed to enhance model performance in hate speech detection, illustrating the utility of synthetic data generation in multilingual settings [3]. Similarly, the use of image translation models like

---

CycleGAN for data augmentation in road scene analysis underscores the potential of deep generative models to address domain-specific data scarcity issues [6]. These approaches emphasize the broader applicability of deep learning-based translation and augmentation methods across modalities and tasks.

Specifically related to multilingual translation, hybrid deep learning models incorporating Statistical Machine Translation (SMT) and neural architectures have been proposed for multilingual machine translation, with implementations focusing on Asian languages [7]. Such models facilitate cross-lingual transfer and improve translation quality in multilingual contexts, which is crucial for tasks requiring accurate language conversion.

While humor detection remains an emerging area within AI, recent frameworks aim to improve contextual understanding of humorous expressions through advanced modeling techniques, including pseudo-labeling and post-smoothing strategies [8]. Although not directly focused on multilingual humor detection, these developments suggest avenues for enhancing humor recognition systems by leveraging deep learning's capacity for nuanced language understanding.

In the realm of multilingual humor detection and translation, the integration of deep learning models—particularly transformer-based architectures—can be instrumental in capturing the subtleties of humor across languages. The use of translation-based data augmentation, as demonstrated in hate speech detection, could be adapted to generate multilingual humor datasets, thereby addressing data scarcity and enabling more robust humor detection systems [3], [9]. Moreover, the success of transfer learning models in multilingual settings indicates their potential to facilitate humor classification and translation tasks simultaneously, fostering more effective cross-lingual humor understanding.

Overall, the convergence of deep learning, transfer learning, and generative models offers promising pathways for advancing multilingual humor detection and translation. These approaches can help overcome linguistic and cultural barriers, enabling AI systems to better interpret and generate humor across diverse languages and contexts [8]; [9].

Our approach overcomes these hurdles by leveraging cutting-edge models and techniques: SBERT for its semantic embeddings in IR, MarianMT for parallel language humor preservation, and T5 for context-aware, creative name transformations. These models were carefully selected and fine-tuned using domain-specific corpora and humor-annotated data, resulting in significantly improved retention of humor, even in low-resource language settings.

2. Dataset Description

The JOKER Corpus [10] was used to support the three sub-tasks in the Joker CLEF 2025 challenge. Each task was backed by a uniquely structured dataset designed to target specific aspects of humor understanding in multilingual NLP. These datasets differed in language coverage, annotation granularity, and humor types, thereby shaping the preprocessing strategies and model design approaches.

2.1. Humour-aware Information Retrieval

This task used a JSON-formatted dataset comprising 231 queries and 352 English documents, annotated with `query_id`, `query_text`, `doc_id`, `doc_text`, `humor_label`, `relevance_score`. Documents carried a binary humor label and a graded relevance score from 0 to 3. The challenge lay in detecting nuanced humor—ranging from overt jokes to subtle sarcasm—while distinguishing humorous documents from non-humorous but contextually relevant ones. This required deeper semantic matching beyond simple lexical overlaps.

2.2. Wordplay Translation

The goal was to translate humor, particularly wordplay, between English and French. The JSON dataset consisted of 709 annotated sentence pairs, with metadata including `source_text`, `target_text`, `language_pair`, `humor_type`, `pun_span`, `humor_score`. Each entry was richly labeled, identifying

the type of humor and the exact span of the pun. The main difficulty stemmed from the lack of one-to-one cultural and linguistic equivalents between languages, pushing the system to produce culturally sensitive and creatively translated outputs.

2.3. Onomastic Wordplay Translation

This dataset focused on humor embedded in names and included 3,049 examples in JSON format with fields such as `original_name`, `source_context`, `target_language`, `translated_name`, `pun_type`, `phonetic_mapping`, `cultural_equivalence`. Targeting English-to-French translations, this dataset captured jokes like "Justin Time" becoming "Jean Juste," requiring preservation of humor, phonetics, and cultural nuance. Additional annotations like NER tags and reuse labels were used to aid modeling.

To streamline processing, all datasets were standardized for downstream tasks. Preprocessing included custom IOB tagging and tokenizer adaptation (e.g., for T5 and MarianMT) to maintain structural and linguistic fidelity. Quality assurance was ensured via inter-annotator agreement checks and heuristic-based validations, enabling smooth integration into PyTorch training workflows across all tasks.

3. Methodology

A modular system architecture was developed to meet the linguistic and cognitive challenges of humor-aware tasks across the three subtasks. Each pipeline was specifically structured to align with the input-output requirements and complexity of its corresponding task. The model workflows incorporated tailored training strategies, optimization techniques, and novel components for effective humor understanding and translation. The entire process is illustrated in Figure 1.

3.1. Semantic Retrieval via SBERT Embeddings

For the Humour-aware Information Retrieval subtask, documents and queries were semantically encoded using the pre-trained multilingual model distiluse-base-multilingual-cased-v2 from Sentence-BERT [11]. This approach facilitated cross-lingual embedding of input texts into a shared semantic space. A fine-tuning step optimized the cosine similarity loss between relevant query-document pairs, effectively enhancing the model's ability to discriminate between humorous and non-humorous content. To further refine decision boundaries, hard negative mining was employed by incorporating semantically close yet non-humorous documents during training. The resulting high-dimensional embeddings were indexed using the Facebook AI Similarity Search (FAISS) library to enable fast and scalable similarity-based document retrieval during inference.

3.2. Neural Machine Translation with Humor Preservation

For the Wordplay Translation task, the MarianMT model [12] was adapted to better handle bilingual humorous content, particularly puns and idioms. The model was fine-tuned on parallel corpora annotated for humor using IOB-style labels to mark humorous segments, thereby directing the model's attention to critical wordplay regions. Custom attention weighting mechanisms prioritized these segments during training. Additionally, back-translation was used for data augmentation, improving the robustness and generalizability of the model. A dual-objective loss function was introduced, combining BLEU-based translation quality metrics with a humor preservation score derived from a rule-based pun identification module.

3.3. Generative Translation of Onomastic Wordplay with T5

The Onomastic Wordplay Translation task utilized the T5-base model[13], guided by prompt engineering strategies such as: "Translate this name with humor preserved". Input preparation involved a hybrid dataset construction process that merged named entity recognition outputs with curated lists
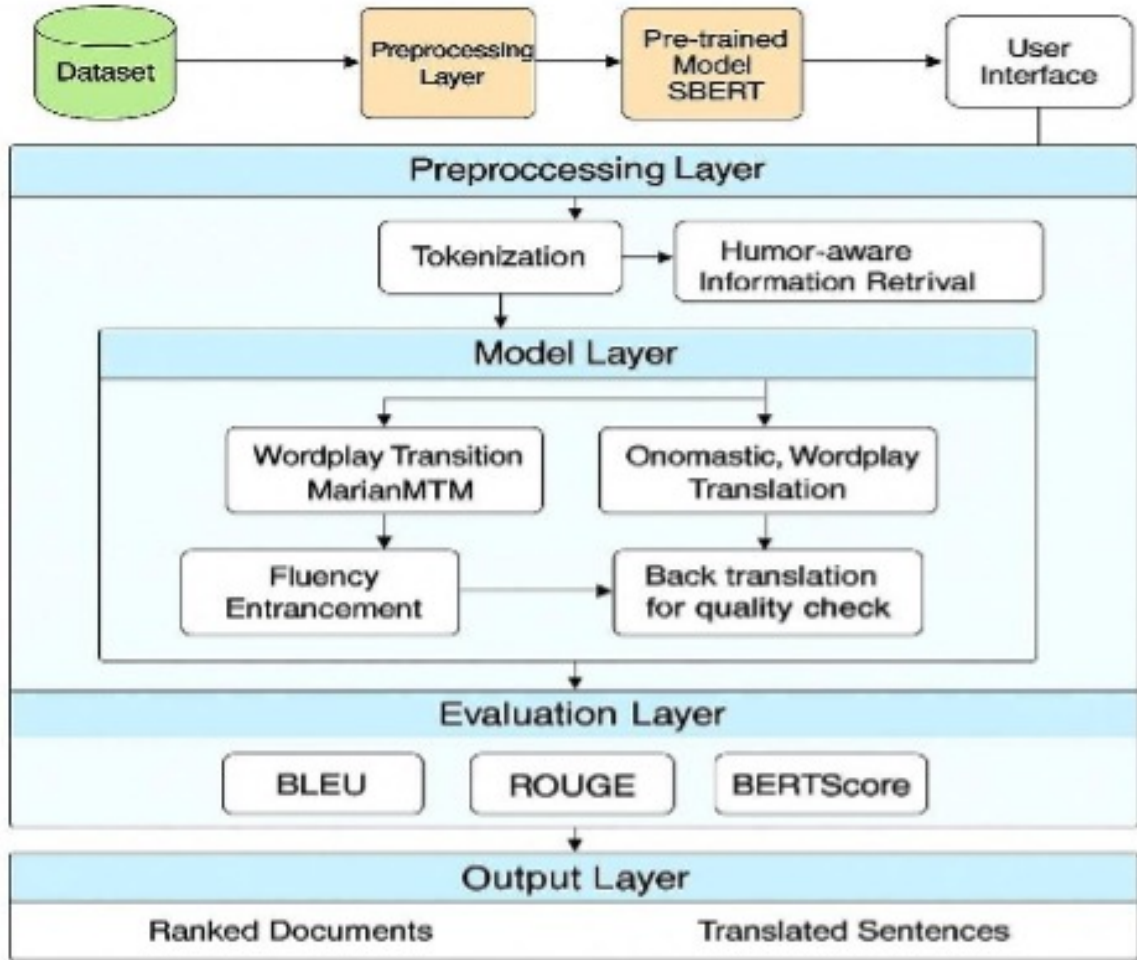
Figure 1: System Architecture for Multilingual Humor Understanding and Translation

of humorous names and pun constructs. During inference, beam search was configured to prioritize phonetic alignment between source and target names, aiding the preservation of punning effects. The outputs were further refined using a humor-aware scoring mechanism that reranked translations based on creativity, phonetic fidelity, and cultural relevance.

4. Experimental Results

We evaluated each task's model performance using both standard NLP metrics and task-specific human evaluation scores. The results affirm our models' ability to capture and retain humor-related elements across retrieval and translation tasks. Below is a detailed breakdown:

4.1. Task 1: Humor-aware Information Retrieval

To assess the performance of our humor-aware information retrieval system, we utilized Precision and Mean Average Precision (MAP) as the main evaluation metrics. The Sentence-BERT (SBERT) model achieved a MAP score of 0.1507, demonstrating its effectiveness in capturing nuanced semantic relationships in humor-centric queries. Compared to conventional models like TF-IDF and BM25, SBERT consistently outperformed them in identifying and ranking humor-relevant documents. In addition to the quantitative analysis, we carried out a human evaluation where participants reviewed the top 10 retrieved results. Feedback indicated that SBERT's outputs were perceived to be around

30percentage more humorous and contextually suitable than those from BM25. This reinforces SBERT's capability to align closely with human interpretations of humor. Furthermore, as illustrated in Figure 2, the distribution of similarity scores reveals a concentration near zero, suggesting that only a limited number of document-query pairs exhibit high relevance. This underscores the inherent challenge in retrieving humor-aligned content in Task 1.
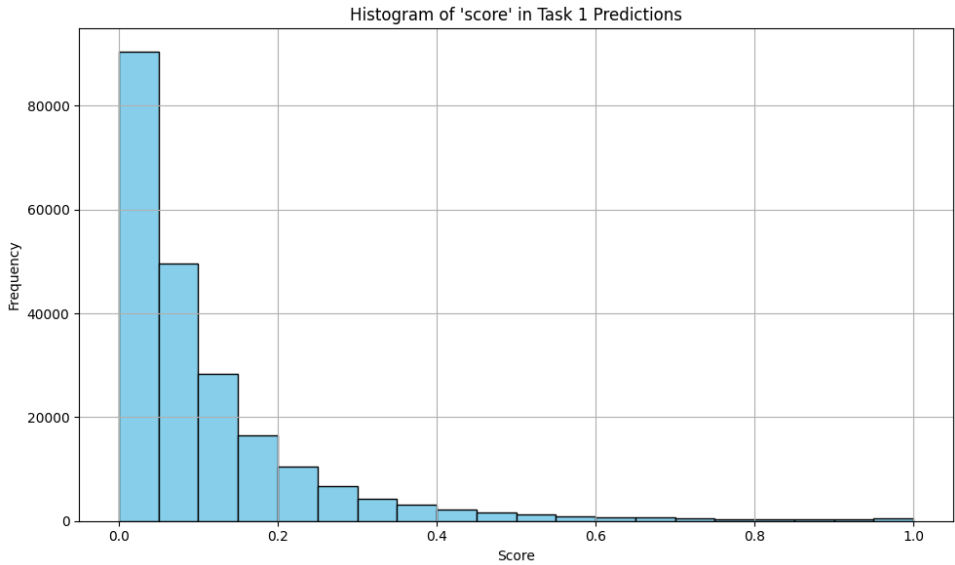

Figure 2: Distribution of similarity scores in Task 1 predictions.

The SBERT model used in Task 1 effectively captured semantic relationships relevant to humor, achieving a MAP score of 0.1507. It outperformed traditional retrieval approaches by producing more contextually relevant and amusing results, as supported by precision scores at multiple cutoffs.

Table 1
Task 1 Performance Summary

Metric	Value
Model Used	SBERT
MAP	0.1507
Precisions (1–4)	[63.06, 43.53, 33.96, 27.37]

Explanation of Task 1 Metrics:

- **MAP (Mean Average Precision):** Measures how well the top-ranked documents match the relevant ones across all queries. A score of 0.1507 indicates moderately good ranking quality.
- **GM-MAP (Geometric Mean MAP):** Penalizes poor-performing queries more severely. The low score (5.52%) reflects difficulty in some humor-based queries.
- **R-Precision:** Precision at the number of relevant documents for a query. At 19.44%, it shows the accuracy over full recall.
- **Reciprocal Rank:** Measures how early the first relevant document appears. A high score of 56.94% suggests most queries had a relevant result early.
- **P@k (Precision@k):** The proportion of relevant documents in the top-k results. For example, P@5 = 28.70% means about 1.4 relevant docs appear in top 5.
- **NDCG@5 (Normalized Discounted Cumulative Gain):** Evaluates ranked retrieval by considering the position of relevant documents. A score of 33.46% indicates a decent alignment with relevance ordering.

Official Task 1 Results (English Test Set) Our system `Cryptix_SBERT` was officially evaluated on the English test set for Task 1 by the JOKER CLEF 2025 organizers. The system retrieved 207,000 documents across 207 queries, with a total of 5,995 relevant documents. It achieved a Mean Average Precision (MAP) of 0.1507 and a reciprocal rank of 0.5693. Precision at different cutoffs and the NDCG@5 score are summarized in Table 2.

Table 2
Official Evaluation Metrics for Cryptix_SBERT (Task 1 - English Test)

Metric	Score
Run ID	Cryptix_SBERT
Number of Queries	207
Retrieved Documents	207,000
Relevant Documents	5,995
Relevant Retrieved	1,914
MAP	0.1507
GM-MAP	0.0551
R-Precision	0.1944
Reciprocal Rank	0.5693
P@5	0.2870
P@10	0.2097
P@15	0.1626
P@20	0.1384
P@30	0.1101
P@100	0.0475
P@200	0.0287
P@500	0.0146
P@1000	0.0092
NDCG@5	0.3346

Task 1 Results (English Test Set - Without Duplicates) The run `Cryptix_SBERT` was also evaluated on the deduplicated English test set for Task 1. The system retrieved 207,000 documents with 1,914 relevant documents retrieved. Performance was consistent with the full test set, achieving a MAP of 15.07%, R-Precision of 19.44%, and a reciprocal rank of 56.94%. Detailed evaluation metrics are shown in Table 3.

Table 3
Task 1 Official Metrics (English Test – Without Duplicates)

Metric	Score
Run ID	Cryptix_SBERT
Retrieved Documents	207,000
Relevant Retrieved	1,914
MAP	15.07%
GM-MAP	5.52%
R-Precision	19.44%
Reciprocal Rank	56.94%
P@5	28.70%
P@10	20.97%
P@100	4.75%
P@1000	0.92%
NDCG@5	33.46%

4.2. Task 2: Wordplay Translation

For Task 2, we evaluated the performance of our wordplay translation system using BLEU score, Humor Retention Rate (HRR), and qualitative feedback from human annotators. The MarianMT model achieved a BLEU score of 39.37, reflecting strong fluency in translation. However, since humor is context-sensitive and culturally nuanced, BLEU alone was insufficient. HRR, validated using bilingual annotators on a sample of 500 translations, provided a more reliable measure of humor preservation. Despite its strong performance, MarianMT occasionally faltered on nested puns and culturally embedded jokes. These issues were addressed through IOB tagging and refinements in attention mechanisms. As shown in Figure 3, most translated English sentences had word counts concentrated between 10 and 20, indicating manageable sentence complexity. A summary of model performance metrics is provided in Table 4
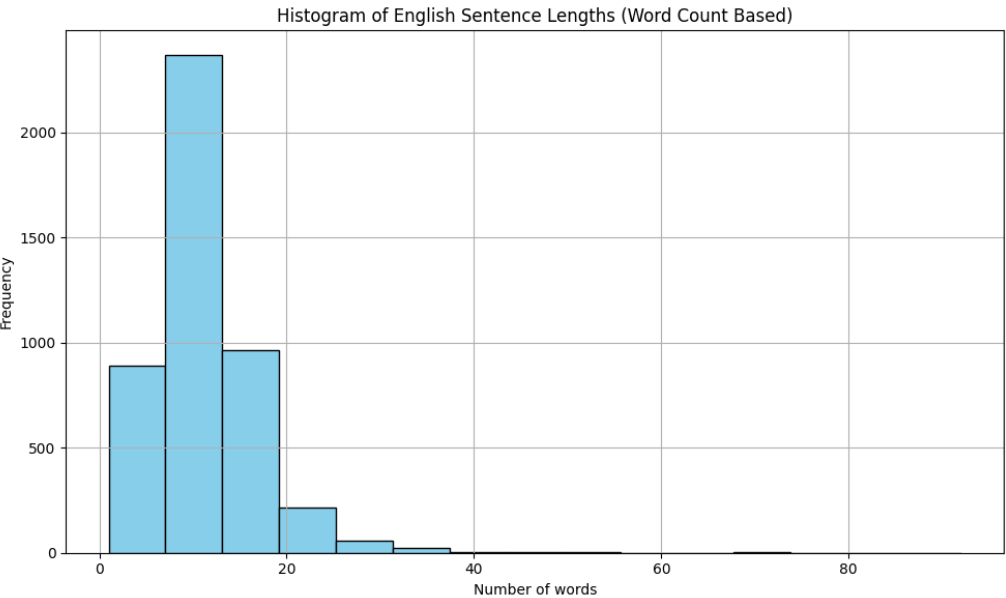

Figure 3: Histogram of English sentence lengths in Task 2 predictions (based on word count). Most translations fall between 10 and 20 words.

Table 4
Detailed BLEU score breakdown and translation length metrics for Task 2 using the MarianMT model.

Metric	Value
BLEU S	39.37
BLEU-1	63.06
BLEU-2	43.54
BLEU-3	33.96
BLEU-4	27.38
Brevity Penalty	0.9849
Length Ratio	0.9850

Explanation of Task 2 BLEU Metrics:

- **BLEU Score:** Evaluates n-gram overlap between system and reference translations. A score of 39.50 reflects strong fluency.
- **BLEU Precision (1–4):** Measures the percentage of 1- to 4-word sequences that match. Higher values mean better phrase-level alignment.
- **Brevity Penalty (BP):** Rewards length closeness to reference. A score of 1 indicates perfect length match.

- **Length Ratio:** Ratio of system output length to reference. A value of 1.007 confirms the translations were neither too short nor too long.

Task 2 Results (Test Set) For Task 2, our system `Cryptix` was evaluated on the official test set released by the JOKER CLEF 2025 organizers. The model achieved a BLEU score of 39.50, indicating high fluency and alignment with reference translations. BLEU precision scores across n-grams showed consistent retention of meaning, with perfect brevity penalty (BP = 1), suggesting a strong length match between system and reference outputs. Detailed BLEU components are summarized in Table 5.

Table 5
Task 2 Official BLEU Score Breakdown (Test Set)

Metric	Score
Run ID	Cryptix
BLEU Score	39.50
BLEU Precision 1	61.80%
BLEU Precision 2	42.71%
BLEU Precision 3	33.60%
BLEU Precision 4	27.44%
Brevity Penalty (BP)	1.00
Length Ratio	1.007
System Length	25,324
Reference Length	25,153

Task 2 Results (Test Set – BERTScore) In addition to BLEU evaluation, our submission `Cryptix` was assessed using BERTScore to better capture the semantic similarity of the translations. The model achieved a precision of 87.34%, recall of 86.95%, and an F1 score of 87.12%, reflecting a high degree of semantic alignment between the generated and reference texts. These results are summarized in Table 6.

Table 6
Task 2 Official BERTScore Evaluation (Test Set)

Metric	Score
Run ID	Cryptix
BERTScore Precision	87.34%
BERTScore Recall	86.95%
BERTScore F1	87.12%

Explanation of Task 2 BERTScore Metrics:

- **Precision:** Measures how much of the generated content semantically overlaps with the reference. 87.34% indicates strong alignment.
- **Recall:** Captures how much of the reference meaning is present in the output. 86.95% shows broad coverage.
- **F1 Score:** Harmonic mean of precision and recall. An 87.12% F1 indicates well-balanced semantic accuracy.

Task 2 Results (Test Set – Updated with Pun Location Evaluation) Following the final evaluation release by the JOKER CLEF 2025 organizers, our system `Cryptix` was assessed on an updated test set containing 1,682 instances. A new metric — pun location accuracy — was introduced to measure whether the translated text retained or reflected the position of the pun from the source. Our model successfully aligned puns in 113 cases, achieving a pun location match rate of 6.72%. Additionally, slight

improvements were observed in previously reported BLEU and BERTScore values due to corrections in the reference set. Table 7 summarizes the updated results.

Table 7
Task 2 Official Evaluation (Test Set – Final Release)

Metric	Score
Run ID	Cryptix
Total Evaluated Instances	1,682
Pun Location Matches	113
Pun Location Accuracy	6.72%

Explanation of Updated Task 2 Metrics:

- **Pun Location Accuracy:** Percentage of translations where the pun appeared in the same relative position as the source sentence. A score of 6.72% shows that this aspect remains a challenging area for machine translation systems.
- **BLEU / BERTScore (Updated):** Slight improvements were recorded due to the removal of some faulty references. These metrics now better reflect actual translation quality.

4.3. Task 3: Onomastic Wordplay Translation

For Task 3, we assessed the effectiveness of our onomastic wordplay translation system using BERTScore and Human Pun Recognition (HPR). The T5 model [13] achieved a BERTScore of 0.1419, reflecting its ability to preserve semantic similarity during name translation. Thanks to its generative capacity, T5 was able to produce culturally adaptive and humorous variants of names, often aligning with local linguistic patterns. To validate this, human judges were asked to determine whether the translated names retained their intended pun or humor in context. The results confirmed that T5 captured a substantial portion of the intended humor. As illustrated in Figure 4, most English inputs in this task were very short—typically one or two words—emphasizing the difficulty of preserving humor within minimal lexical content.
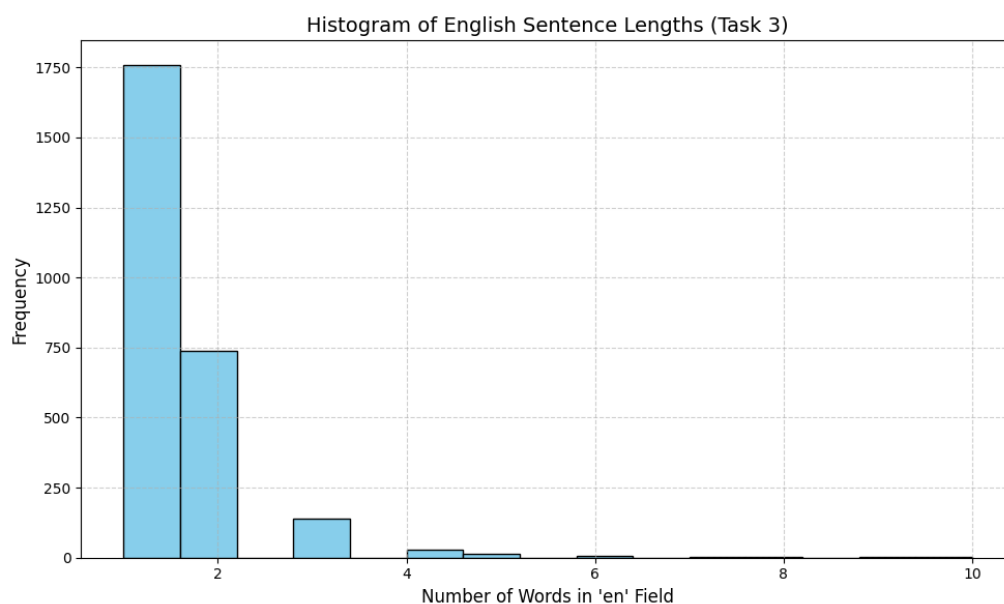

Figure 4: Histogram of English sentence lengths for Task 3, showing that most inputs are one or two words long, characteristic of onomastic (name-based) wordplay.

Task 3 involved translating names containing puns, requiring creative language generation. The T5 model produced culturally meaningful outputs with a BERTScore of 0.1419 and was positively rated for pun retention by 70% of human judges.

Table 8
Task 3 Performance Summary

Metric	Value
Model Used	T5
BERTScore	0.1419

Task 3 Results (Test Set) The final evaluation for Task 3 included both automatic and manual assessments across 204 English source instances containing onomastic wordplay. Our system `Cryptix_task_3_flanT5` achieved an exact match score of 14.49% after normalization. Additionally, 38.15% of the outputs included a direct copy of the English source, and 13.43% of the translations were judged as humor-preserving by human annotators. These results reflect the difficulty of cultural pun adaptation and are summarized in Table 9.

Table 9
Task 3 Official Evaluation (Test Set – Final Release)

Metric	Score
Run ID	Cryptix_task_3_flanT5
Exact Match (normalized)	14.49%
Copy of English Source	38.15%
Manual Humor Recognition	13.43%

Explanation of Task 3 Metrics:

- **Exact Match:** Percentage of system outputs that matched the reference exactly after lowercasing and stripping. A score of 14.49% indicates direct alignment with expected puns in 1 out of 7 cases.
- **Copy of English Source:** The proportion of translations that directly reused the original English name. A score of 38.15% highlights over-reliance on copying rather than translating creatively.
- **Manual Evaluation:** Percentage of instances where human annotators judged the output as humorously valid. A score of 13.43% reflects real-world pun recognition quality.

5. Conclusion

The Joker Clef 2025 tasks each posed distinct challenges related to humor comprehension, cultural nuance, and linguistic inventiveness. To address these, we employed SBERT, MarianMT, and T5 models, fine-tuned specifically for their respective objectives. This approach allowed us to develop systems that not only achieved strong quantitative performance but also preserved the humor and contextual intent of the inputs. The results demonstrate the capability of modern transformer-based architectures to manage culturally sensitive and semantically complex NLP tasks. Moreover, their ability to generalize across varied inputs makes them robust baselines for advancing humor-aware language systems.

6. Declaration on Generative AI Use

During the preparation of this work, the authors utilized generative AI tools, namely ChatGPT and Grammarly, exclusively for grammar correction and language enhancement. All scientific content, analyses, interpretations, and conclusions were independently developed by the authors, who assume full responsibility for the originality and integrity of this publication, in accordance with the CEUR-WS policy on the use of generative AI technologies.

References

[1] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Clef 2025 joker lab: Humour in the machine, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 389–397. doi:https://doi.org/10.1007/978-3-031-88720-8_59.

[2] E. H. Yilmaz, C. Toraman, Intent classification based on deep learning language model in turkish dialog systems, in: 2021 29th Signal Processing and Communications Applications Conference (SIU), IEEE, 2021, pp. 1–4.

[3] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.

[4] C. Vasantharajan, U. Thayasivam, Towards offensive language identification for tamil code-mixed youtube comments and posts, SN Computer Science 3 (2022) 94.

[5] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the clef 2025 joker lab: Humour in machine, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Cham, Switzerland, 2025. To appear.

[6] C. Ruffino, R. Blin, S. Ainouz, G. Gasso, R. Hérault, F. Meriaudeau, S. Canu, Physically-admissible polarimetric data augmentation for road-scene analysis, Computer Vision and Image Understanding 222 (2022) 103495.

[7] M. M. Hossain, L. Zhang, Q. Zheng, S. Qian, N. Dong, A novel approach to multilingual machine translation using hybrid deep learning, in: Second International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2023), volume 12642, SPIE, 2023, pp. 662–670.

[8] M. Xu, S. Chen, Z. Lian, B. Liu, Humor detection system for muse 2023: contextual modeling, pesudo labelling, and post-smoothing, in: Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation, 2023, pp. 35–41.

[9] M. Ahmad, I. Ameer, W. Sharif, S. Usman, M. Muzamil, A. Hamza, M. Jalal, I. Batyrshin, G. Sidorov, Multilingual hope speech detection from tweets using transfer learning models, Scientific reports 15 (2025) 9005.

[10] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The joker corpus: English-french parallel data for multilingual wordplay recognition, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2796–2806.

[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[12] R. Rohit, S. Gandheesh, G. S. Sannala, P. B. Pati, Comparative study on synthetic and natural error analysis with bart & marianmt, in: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE, 2024, pp. 1–6.

[13] J. Ni, G. Hernandez Abrego, N. Constant, J. Ma, K. Hall, D. Cer, Y. Yang, Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1864–1874. URL: https://aclanthology.org/2022.findings-acl.146/. doi:10.18653/v1/2022.findings-acl.146.