University of Split and University of Malta (Team AB&DPV) at the CLEF 2024 JOKER Track: From 'LOL' to 'MDR' Using Artificial Intelligence Models to Retrieve and Translate Puns

Notebook for the JOKER Lab at CLEF 2024[1] by Team AB&PDV

Antonia Bartulović[1]*† and Dóra Paula Váradi[2]*†

[1] *University of Split Ul. Ruđera Boškovića 31, 21000, Split, Croatia*
[2] *University of Malta, Msida MSD 2080, Malta*

Abstract

The JOKER-2024 track aims to enhance the automatic processing of humorous wordplay, addressing the complexities involved in understanding and translating humour. The study comprises three tasks: humour-aware information retrieval, humour classification by genre and technique, and the translation of puns from English to French. Utilizing traditional classifiers, the research is tuned to these models on humour-specific datasets. The baseline approaches for JOKER 2024 track tasks which include TF-IDF for Task 1, the use embeddings with the help of Word2Vec and the use of Multilayer Perceptron for Task 2, and the use of Llama-2-7b for task 3. Despite promising initial results in information retrieval, the study found humour classification and pun translation to be challenging due to cultural and linguistic nuances. The research highlights the need for more sophisticated models and larger, diverse datasets to improve accuracy and effectiveness in automatic humour processing.

Keywords

Natural Language Processing, Computational Humour Detection, Humour Location, Machine Translation

1. Introduction

1.1. Introduction and overview

The CLEF JOKER-2024 Track [1] [2] focuses on the automatic processing of humorous wordplay, requiring cultural reference recognition, word formation knowledge, and double meaning discernment. This interdisciplinary effort aims to address the challenges in understanding and translating wordplay for both humans and machine users. For example, "LOL" is an acronym for "Laugh Out Loud," often used to indicate something is funny in English. On the other hand, "MDR" is an abbreviation for "Mort de Rire," which translates to "Dying of Laughter" in French.

The JOKER 2023 track involved three tasks:

- Task 1: Humour-aware information retrieval [3] [4]. The objective is to retrieve humorous texts from a document collection based on a query, ensuring relevance and wordplay presence.
- Task 2: Humour classification according to genre and technique [5] [6]. The objective is to classify texts into irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating, and wit-surprise.

[1] *CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*
* Corresponding author.
† These authors contributed equally.

antonia.bartulovic.00@fesb.hr (A. Bartulović); dora.varadi.21@um.edu.mt (D. P. Váradi)

- Task 3: Translation of puns from English to French [7] [8]. The objective is to translate English puns into French, preserving both form and meaning.

The motivation behind this research is to tackle the complexities and nuances involved in processing and understanding humorous wordplay, which poses significant challenges for both humans and machines. This involves recognizing cultural references, understanding word formation, and discerning double meanings, all of which are crucial for accurate humour detection and translation. By advancing the capabilities of natural language processing (NLP) systems in these areas, the research aims to improve the automatic retrieval, classification, and translation of humorous content, thereby enhancing user experiences in various applications, from entertainment to communication technologies. Additionally, improving cross-cultural communication and translation, ensuring that humour, which often relies heavily on cultural context, can be appreciated and understood universally are important research objectives.

The report highlights state-of-the-art works on humour awareness and translation then delves into the approaches used in this research, followed by an analysis of the results.

1.2 State-of-the-Art Overview

1.2.1. Humour-Aware Information Retrieval

Humour-aware information retrieval is a specialised and quickly expanding field in natural language processing. Conventional information retrieval algorithms predominantly depend on matching keywords and assessing semantic similarity to establish relevance. Nevertheless, these systems frequently encounter difficulties when it comes to processing hilarious content, mostly because of the intricate nature and subtle nuances of humour. Recent developments in this area involve the integration of more advanced models, such as transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers), which demonstrate exceptional proficiency in comprehending context and subtleties [9].

During the process of humour retrieval, these models undergo fine-tuning using datasets specifically designed for humour. This allows them to more effectively capture the fundamental aspects of wordplay and jokes. TF-IDF, a technique that stands for Term Frequency-Inverse Document Frequency, is frequently utilised in conjunction with sophisticated embeddings to improve the model's capacity to effectively identify and prioritise hilarious content. Integrating conventional methods such as TF-IDF with embeddings derived from models like Word2Vec, GloVe, and BERT improves the effectiveness of the system [10] [11]. Incorporating external knowledge bases that contain cultural allusions greatly enhances the model's performance, allowing it to comprehend and handle the intricacies of humour [12].

1.2.2. Humour Classification According to Genre and Technique

Categorising humour into distinct genres and techniques continues to be a difficult undertaking because of its subjective nature. Contemporary methods utilise machine learning algorithms, which encompass a variety of approaches such as traditional classifiers like Random Forests, as well as more advanced models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [13] [14].

Contemporary approaches utilise transformer models like BERT and RoBERTa, which are trained on extensive, annotated datasets specifically designed for humour analysis. These models excel at collecting intricate linguistic patterns and contextual information, which are essential for differentiating between various forms of humour, such as irony, sarcasm, and wit [15]. Transformer models have shown exceptional efficiency in tasks involving the classification of humour, thanks to their capacity to handle enormous amounts of data and comprehend contextual subtleties. The fine-tuning method entails training these models on datasets that are particularly labelled for various types of humour, allowing them to acquire knowledge of the nuanced distinctions between different hilarious styles.

Furthermore, the integration of textual and visual data, such as memes, using multimodal techniques, has demonstrated potential in enhancing the accuracy of classification. These methods utilise models that are capable of analysing and combining data from many sources, hence improving the capacity to categorise comedy that depends on both written language and visual content. For instance, Kiela et al. [16] illustrates how the combination of visual data and textual analysis can greatly enhance the comprehension and categorization of humour in memes, which frequently depend on both visual background and verbal punchlines.

1.2.3. Translation of Puns

Translating puns presents a particularly arduous task as it necessitates not just linguistic translation but also cultural adjustment. Puns frequently depend on the use of wordplay, homophones, and cultural allusions that are not readily translatable across other languages. Conventional machine translation methods, which primarily prioritise syntactic and semantic precision, frequently struggle to maintain the comedy and clever wordplay found in puns.

Current models in this field utilise transformer-based structures such as MarianMT and OpenNMT. These models are optimised using parallel datasets that consist of puns and their corresponding translations [17] [18]. These models utilise their advanced ability to learn and comprehend the intricate and situation-dependent characteristics of puns.

Recent progress has been made in the field by utilising Large Language Models like GPT-3 and LLaMA. These models are capable of producing translations by comprehending context and subtle distinctions [19]. These models utilise methods such as controlled creation using precise prompts and temperature settings to preserve the humour and significance of the pun in the desired language. Translators can manipulate these settings to exert control over the inventiveness and diversity of the translations, so safeguarding the whimsical elements of the original text.

Incorporating bilingual dictionaries and cultural allusions can enhance the accuracy and humour of translations. This method guarantees that the translations faithfully preserves the cultural context and humour of the original, which is essential for puns that largely depend on these components. The study conducted by Holtzman et al. investigates the use of controlled text generation techniques to preserve specific traits, such as humour, in translation [20]. This is achieved by carefully controlling the process of generating text.

2. Approach

2.1. Data Description

The data for each task is structured as follows:

Task 1: The dataset consists of a JSON file with short texts, training queries, and relevance judgments.

Task 2: The dataset consists of manually annotated JSON files containing humorous texts categorized by genre and technique.

Task 3: The dataset consists of a JSON files with English puns and their corresponding French translations.

2.2. Methodology

2.2.1. Task 1: Humour-aware Information Retrieval

The first task involves preprocessing the data by removing empty entries to avoid processing issues and tokenizing the text, converting all text to lowercase. Words are tagged to identify wordplay, and TF-IDF [21] is applied to represent the text and score the documents based on the presence of humour-related terms and their relevance to the query.

2.2.2. Task 2: Humour Classification

For the second task, the corpus of training data is merged with corresponding genre and technique classifications. A Random Forest classifier [22] is used initially. The text is tokenized and vectorised using Word2Vec [23], followed by training and testing the classifier with varying numbers of estimators. The estimators used were 50, 100, 250, 500,1000, and 2000.

An MLPClassifier [24] is then utilised, experimenting with different alliteration ranges (50, 100, 200, 500, 750, 1000, 1500, 2000, and 3000 neurons) and activation functions (Tanh). Other models, such as Gaussian Naive Bayes [25], Decision Tree Classifier [26], and LogisticRegression [27] were also tested, however, showed lower accuracy than MLPClassifier, highlighting the complexity of humour classification.

2.2.3. Task 3: Translation of Puns

For the third task, a LLM such as Llama-2-7b [28] is used. Each joke is input into the LLM using a specific prompt format. The temperature is set to 0.7 to balance randomness and coherence. Unnecessary characters are removed, and outputs are fine-tuned to ensure the preservation of humour and meaning in the translations. The following prompts were used:

- “You are a translator that outputs in JSON. You always use the following format: \{ 'translation': 'joke' \}. You use \" quotes.”
- “Translate the following joke from English into French, ensuring that the humor and punchline are preserved as much as possible while considering cultural differences and linguistic nuances. Feel free to adapt the joke as needed to make it work in the target language.”

3. Results

3.1. Results of Task 1

The TF-IDF scores were utilized to represent the text and score the documents based on the presence of humour-related terms and their relevance to the query. These scores provided valuable insights into the importance of specific terms within the context of humour retrieval. By utilising TF-IDF, we were able to effectively identify and rank humorous texts within the document collection, thus contributing to the success of the information retrieval task

Table 1: Task 1 Retrieval Results Using TF-IDF

run_id	map	ndcg	R@5	R@10	R@15	R@20	R@30	R@100	R@200	R@500	R@1000	bpref	recip_rank	P_1	P_5	P_10
AB&DPV_task_1_TFIDF	0,08612132498	0,2413229819	0,0728292012	0,1257164331	0,1514077895	0,1862047957	0,2204701972	0,3251114011	0,342960853	0,3630872118	0,3674296218	0,1028903169	0,2540405555	0,1333333333	0,1155555556	0,1444444444

3.2. Results of Task 2

Despite the efforts, the testing results indicated limited accuracy with different classifiers and estimators. The below table showcases highest levels of accuracy achieved with different estimators.

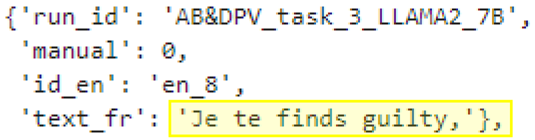
Table 2: Task 2 Classification Results

run_id	accuracy	weighted avg_precision	weighted avg_recall	weighted avg_f1-score	weighted avg_support
AB&DPV_task_2_MLP3000params	0,48	0,45	0,48	0,44	722,00
AB&DPV_task_2_RandomForestClassifier250	0,38	0,38	0,38	0,29	722,00
AB&DPV_task_2_RandomForestClassifier500	0,38	0,36	0,38	0,29	722,00
AB&DPV_task_2_MLP2000	0,37	0,15	0,37	0,21	722,00
AB&DPV_task_2_MLP3000	0,37	0,15	0,37	0,21	722,00
AB&DPV_task_2_DecisionTreeClassifier	0,29	0,29	0,29	0,28	722,00
AB&DPV_task_2_GaussianNB	0,27	0,29	0,27	0,25	722,00

MLP Testing with 2000 neurons achieved 33% accuracy, and 3000 neurons achieved 41% accuracy. A possible further increase could have produced better results, however due to limited resources this could not be done.

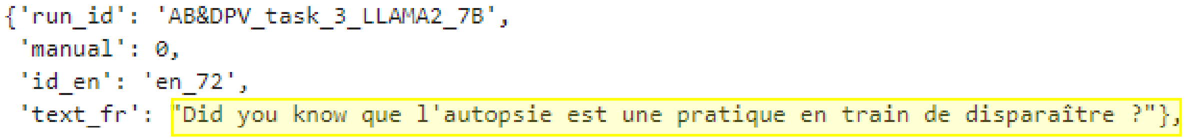
3.3. Results of Task 3

The results were not desirable as Llama-2-7b doesn't understand humour, so translation is inaccurate. In quite a few instances there were cases when not the entire pun was translated but only a few words as shown in below figures 1 and 2. Furthermore, judging if the translations retain puns could not be judge due to language barriers.



```
{'run_id': 'AB&DPV_task_3_LLAMA2_7B',
 'manual': 0,
 'id_en': 'en_8',
 'text_fr': 'Je te finds guilty,'},
```

Figure 1: EN to FR translation Example 1



```
{'run_id': 'AB&DPV_task_3_LLAMA2_7B',
 'manual': 0,
 'id_en': 'en_72',
 'text_fr': "Did you know que l'autopsie est une pratique en train de disparaître ?"},
```

Figure 2: EN to FR translation Example 2

Section 4. Conclusions

The project encountered several challenges, including the inherent complexity of humour detection and classification due to cultural and linguistic nuances. The accuracy of classification models indicates a need for more refined features and larger, more diverse training datasets. Future work could explore advanced transformer models like GPT-4 for improved understanding and generation of humour, as well as incorporating more contextual and cultural information to enhance humour detection and translation.

This project demonstrates the potential and challenges of automatic humour processing in NLP. While the initial results are promising, particularly in humour-aware information retrieval and pun translation, further advancements are needed to achieve higher accuracy and better handle the intricacies of humour across languages and cultures.

References

[1] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma-Preciado, G. Sidorov and A. Jatowt, “Overview of {CLEF 2024 JOKER} track on Automatic Humor Analysis,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, 2024.

[2] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma-Preciado, G. Sidorov and A. Jatowt, “Overview of JOKER - CLEF-2023 Track on Automatic Wordplay Analysis,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, 2023.

[3] L. Ermakova and et al., “Overview of the {CLEF 2024 JOKER} Task 1: Humour-aware information retrieval,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.

[4] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma-Preciado, G. Sidorov and A. Jatowt, “Overview of JOKER 2023 Automatic Wordplay Analysis Task 1 - Pun Detection,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, 2023.

[5] V. M. Palma-Preciado and et al., “Overview of the {CLEF 2024 JOKER} Task 2: Humour classification according to genre and technique,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.

[6] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma-Preciado, G. Sidorov and A. Jatowt, “Overview of JOKER 2023 Automatic Wordplay Analysis Task 2 - Pun Location and Interpretation,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, 2023.

[7] L. Ermakova and et al., “Overview of the {CLEF 2024 JOKER} Task 3: Translate puns from English to French,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.

[8] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma-Preciado, G. Sidorov and A. Jatowt, “Overview of JOKER 2023 Automatic Wordplay Analysis Task 3 - Pun Translation,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, 2023.

[9] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171-4186.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013, pp. 3111--3119.

[11] J. Pennington, R. Socher and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[12] J. West and C. T. Bergstrom, “Misinformation in and about science,” *Proceedings of the National Academy of Sciences,* vol. 116, no. 16, pp. 7657-7662, 2019.

[13] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[14] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv preprint arXiv:1907.11692,* 2019.

[16] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh and D. Testuggine, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2611--2624.

[17] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev and et al., “Marian: Fast Neural Machine Translation in C++,” in *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 116-121.

[18] G. Klein, Y. Kim, Y. Deng, J. Senellart and A. M. Rush, “OpenNMT: Open-Source Toolkit for Neural Machine Translation,” in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 67-72.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and et al., “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877-1901.

[20] A. Holtzman, J. Buys, L. Du, M. Forbes and Y. Choi, “The Curious Case of Neural Text Degeneration,” *arXiv preprint arXiv:1904.09751,* 2019.

[21] [Online]. Available: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[22] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[23] [Online]. Available: https://www.tensorflow.org/text/tutorials/word2vec.

[24] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[25] [Online]. Available: https://builtin.com/artificial-intelligence/gaussian-naive-bayes

[26] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[27] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[28] [Online]. Available: https://huggingface.co/meta-llama/Llama-2-7b