CLEF 2024 JOKER Lab: Automatic Humour Analysis

Liana Ermakova Anne-Gwenn Bosser Tristan Miller
Victor Preciado Grigori Sidorov Adam Jatowt

CLEF, Grenoble, September 11, 2024

JOKER Track Motivation

- Humor remains one of the most di cult aspects of intercultural communication & translation
- Applications: Machine translation (Google Translate, DeepL,...), conversational agents (Siri, Alexa,...), humour study, social listening reputation monitoring, recommendation, fake news and hate speech detection (sexism, racism,...)...
- SOTA AI models are wordplay- and humour-agnostic

Goals

- To provide appropriate reusable data and benchmarks for automatic wordplay analysis.
- To provide a discussion platform to address technical & evaluation challenges of automatic wordplay analysis
- Use cases:
  - Computer-Assisted Translation of wordplay
  - Corpus-based analysis of wordplay in the humanities
    - literary criticism
    - language education
    - translation studies
    - humor studies
  - Wordplay-aware Information Retrieval

JOKER@CLEF Shared Tasks

TASK 1: Humour-aware information retrieval

TASK 2: Humour classi cation according to genre and technique

TASK 3: Translation of puns from EN to FR

CLEF'24 JOKER Track Participation

Of over 53 registered teams, 22 teams submitted 103 runs

Team	Task 1	Task 2	Task 3	Total
jokester	1	1	1	3
LIS	1			1
Arampatzis	10	8	6	24
Frane	1	1	1	3
AB&DPV	1	7	1	9
Dajana&Kathy	1	1	1	3
Petra&Regina	1	1	1	3
Tomislav&Rowan	1	3	2	6
UAms	8	1	2	11
RubyAiYoungTeam	1	1		2
ORPAILLEUR		9		9
NaiveNeuron		3		3
HumourInsights		1		1
CYUT		3		3
CodeRangers		2		2
VayamSolveKurmaha		2		2
DadJokers		3		3
NLPalma		3		3
PunDerstand		4		4
Olga			3	3
Farhan			2	2
UBO			3	3
Total	26	54	23	103

Task 1: Humour-aware IR

Retrieving short humorous texts from a document collection

Use case: to search for a joke on a speci c topic

Queries = locations of wordplay from JOKER 2023 Task 2

Collection: 61,268 documents

- 4,492 humourous texts (3,507 texts from JOKER 2023 + 985 new wordplay)
- 4,954 negative examples from JOKER 2023
- 12,523 texts generated using Llama 2
- 39,299 sentences from Wikipedia extracts

Evaluation: traditional IR metrics (MAP, NDCG, ...)

Task 1: Data statistics

Train: 12 queries

Test: 45 queries

11,831 documents topically relevant to all 57 queries

1,730 were considered to be humorous and relevant

Table 1: Statistics of relevant humourous texts per query

count	57
mean	30
std	43
min	1
25%	8
50%	18
75%	38
max	281

Figure 1: Histogram of # relevant humourous texts

Task 1: O cial Results

10 teams, 26 runs

only non-0 scored runs are scored

run ID	map	ndcg	R5	R10	R100	R1000	bpref	MRR	P1	P5	P10
UAms_rm3_T5_Filter2	.12	.28	.09	.15	.36	.43	.18	.26	.13	.11	.13
UAms_rm3_BERT_Filter	.12	.27	.09	.14	.35	.42	.16	.27	.16	.11	.12
UAms_rm3_T5_Filter1	.11	.27	.09	.15	.36	.42	.16	.23	.11	.09	.11
UAms_bm25_BERT_Filter	.09	.24	.06	.12	.37	.40	.12	.19	.09	.05	.08
AB&DPV_TFIDF	.09	.24	.07	.13	.33	.37	.10	.25	.13	.12	.14
UAms_Anserini_rm3	.08	.27	.06	.08	.38	.50	.09	.20	.11	.06	.06
jokester_1_TFIDF_LogRegr	.08	.19	.09	.09	.10	.16	.21	.51	.44	.23	.14
UAms_Anserini_bm25	.08	.24	.06	.08	.37	.42	.09	.19	.11	.05	.06
UAms_bm25_CE100	.04	.17	.03	.04	.37	.37	.06	.08	.00	.04	.03
UAms_rm3_CE100	.04	.18	.03	.04	.38	.38	.06	.07	.00	.04	.03
LIS_MiniLM-T5	.02	.05	.03	.04	.05	.05	.05	.13	.04	.06	.04

Topical relevance results on TEST

run ID	map	ndcg	R5	R10	R100	R1000	bpref	MRR	P1	P5	P10
UAms_Anserini_rm3	.37	.60	.06	.10	.39	.64	.64	.82	.73	.61	.61
AB&DPV_TFIDF	.36	.53	.07	.12	.36	.50	.50	.83	.73	.69	.69
UAms_Anserini_bm25	.35	.55	.07	.11	.38	.56	.56	.79	.64	.61	.60
UAms_bm25_BERT_Filter	.30	.48	.07	.11	.35	.46	.46	.77	.62	.62	.60
UAms_rm3_T5_Filter1	.25	.44	.06	.10	.30	.40	.40	.86	.78	.69	.63
UAms_rm3_CE100	.22	.40	.05	.10	.39	.39	.39	.79	.64	.56	.55
UAms_rm3_BERT_Filter	.22	.39	.06	.09	.27	.34	.34	.84	.76	.68	.61
UAms_bm25_CE100	.22	.39	.05	.10	.38	.38	.38	.78	.62	.56	.55
UAms_rm3_T5_Filter2	.22	.38	.06	.10	.27	.34	.34	.80	.64	.71	.63
jokester_TFIDF_LogRegr	.03	.09	.03	.03	.04	.05	.07	.63	.62	.39	.24
LIS_MiniLM-T5	.01	.05	.02	.02	.03	.03	.03	.33	.18	.20	.15

Results on TRAIN

run_id	map	ndcg	R5	R10	R100	R1000	bpref	MRR	P1	P5	P10
Arampatzis_DecisionTree	.40	.55	.24	.30	.44	.45	.42	.92	.92	.68	.53
Arampatzis_SVM	.36	.52	.25	.28	.44	.45	.39	.83	.75	.68	.52
Arampatzis_kNN	.36	.50	.23	.28	.44	.45	.38	.71	.50	.60	.51
Arampatzis_GaussianNB	.35	.50	.24	.28	.44	.45	.38	.72	.58	.63	.51
UAms_rm3_T5_Filter2	.23	.39	.14	.25	.44	.52	.35	.34	.17	.28	.28
UAms_rm3_BERT_Filter	.23	.42	.12	.23	.50	.60	.36	.37	.17	.23	.23
UAms_rm3_T5_Filter1	.21	.37	.13	.24	.40	.49	.29	.38	.25	.25	.27
UAms_bm25_BERT_Filter	.19	.37	.07	.19	.49	.59	.27	.22	.08	.12	.18
UAms_Anserini_rm3	.17	.37	.09	.18	.45	.63	.30	.24	.08	.17	.18
Arampatzis_NeuralNetwork	.17	.34	.09	.17	.43	.45	.14	.41	.33	.28	.25
Arampatzis_LSTM	.17	.33	.09	.19	.44	.45	.11	.20	.08	.18	.19
ABDPV_TFIDF	.17	.34	.07	.14	.39	.50	.21	.26	.17	.15	.16
UAms_Anserini_bm25	.16	.35	.07	.17	.46	.60	.24	.19	.08	.12	.16
jokester_TFIDF_LogRegr	.16	.34	.11	.12	.14	.36	.49	.59	.58	.30	.20
UAms_rm3_CE100	.07	.22	.01	.03	.45	.45	.09	.12	.00	.08	.09
UAms_bm25_CE100	.07	.22	.01	.03	.46	.46	.09	.12	.00	.08	.08
LIS_MiniLM-T5	.00	.01	.00	.00	.01	.01	.01	.01	.00	.00	.00

Topical relevance results on TRAIN

run ID	map	ndcg	R5	R10	R100	R1000	bpref	MRR	P1	P5	P10
Arampatzis_DecisionTree	.40	.55	.24	.30	.44	.45	.42	.92	.92	.68	.53
AB&DPV_TFIDF	.38	.56	.08	.13	.36	.58	.58	.72	.50	.67	.65
Arampatzis_SVM	.36	.52	.25	.28	.44	.45	.39	.83	.75	.68	.52
Arampatzis_kNN	.36	.50	.23	.28	.44	.45	.38	.71	.50	.60	.51
UAms_Anserini_rm3	.35	.58	.05	.09	.37	.67	.67	.73	.58	.58	.52
UAms_Anserini_bm25	.35	.57	.06	.11	.37	.65	.65	.66	.50	.55	.53
Arampatzis_GaussianNB	.35	.50	.24	.28	.44	.45	.38	.72	.58	.63	.51
UAms_bm25_BERT_Filter	.30	.50	.06	.12	.34	.52	.52	.66	.50	.57	.58
UAms_rm3_T5_Filter1	.25	.42	.06	.11	.28	.39	.39	.73	.67	.58	.62
UAms_rm3_T5_Filter2	.23	.39	.14	.25	.44	.52	.35	.34	.17	.28	.28
UAms_rm3_BERT_Filter	.23	.42	.12	.23	.50	.60	.36	.37	.17	.23	.23
UAms_rm3_CE100	.20	.37	.05	.08	.37	.37	.37	.81	.67	.52	.52
UAms_bm25_CE100	.20	.37	.05	.08	.37	.37	.37	.81	.67	.52	.50
Arampatzis_NeuralNetwork	.17	.34	.09	.17	.43	.45	.14	.41	.33	.28	.25
Arampatzis_LSTM	.17	.33	.09	.19	.44	.45	.11	.20	.08	.18	.19
jokester_TFIDF_LogRegr	.06	.17	.03	.03	.04	.17	.22	.59	.58	.30	.21
LIS_MiniLM-T5	.00	.02	.01	.01	.01	.01	.01	.23	.08	.08	.09

Task 1: Observations (1)

- Low precision due to the presence of the query terms in the non-humorous texts
- Low recall (both train and test): length of the text + the query terms do not appear in many humorous and topically relevant texts
- The runs based on pseudo-relevance feedback RM3 query expansi outperform the BM25 baselines
- Cross-encoder rerankers do NOT exhibit better performance than th baseline models
- Simple solutions such as ones with TF IDF and Logistic Regression remain competitive
- Filtering trained on the wordplay detection task improved systems' results
- Using T5 and BERT language models with RM3 is one of best approaches both in terms of precision and recall

Task 1: Observations (2)

Similar trends for TOPICAL relevance ONLY on train and test

Un ltered runs tend to have higher topical relevance alone but a signi cant drop according to the o cial ranking

The topical relevance scores on train and test are similar, but the ranking on both topical relevance and humor is twice as low on test ! potential over tting

Unique reusable test collection for wordplay retrieval in English

Task 2: Humour classi cation according to genre and technique

The main task is to classify short humorous texts (Multiclass classi cation): Irony, Sarcasm, Exaggeration, Self-deprecating humour, Wit, Incongruity-Absurdity, and Wit-Surprise

Mix of existing datasets + internet collections: JOKER, COVID-19 Humor, iSarcasm, Wallace, Web

Evaluation: Traditional metrics for classi cation tasks

Class	# texts		
	test	train	total
Irony (IR)	147	356	503
Sarcasm (SC)	59	162	221
Exaggeration (EX)	106	210	316
Incongruity/Absurdity (AID)	270	634	904
Self-deprecating (SD)	91	228	319
Wit/Surprise (WS)	49	125	174
Total	722	1,715	2,437

Task 2: Results

18 teams, 54 runs

We report the best result per team

		macro average			weighted average			
Run ID	A	P	R	F1	P	R	F1	#
ORPAILLEUR_mistral-7b-ens	76	71	70	70	75	76	75	722
Code Rangers_roberta	70	75	63	59	78	70	66	509
CYUT_llama3- ne-tuning	70	64	65	64	70	70	70	718
PunDerstand_DeBERTa	69	59	65	60	68	69	67	722
Arampatzis_BERT	68	60	60	59	67	68	67	722
DadJokers_bert_base_uncased	67	60	60	60	67	67	67	722
NLPalma_BERTd	67	60	60	59	67	67	67	722
Demonteam_BERTM	66	58	58	58	65	66	65	722
UAms_BERT_ft	63	57	58	52	66	63	60	722
VayamSolveKurmaha_BERT	60	54	53	51	59	60	58	722
NaiveNeuron_fastText	59	51	51	51	58	59	58	722
DadJokers_RandomForest_MLP_Ensemble	56	49	48	47	54	56	53	722
HumourInsights_Random Forest	55	50	45	45	53	55	52	722
RubyAiYoungTeam	53	53	39	40	52	53	48	722
Petra_and_Regina_LogisticRegression	53	53	39	40	52	53	48	722
Tomislav&Rowan_SVM	51	44	37	38	48	51	47	722
AB&DPV_MLP3000params	48	41	38	38	45	48	44	722

Task 3: Pun Translation

- The goal of this task is to translate English punning jokes into French and preserve:
  - wordplay form
  - wordplay meaning
- Train data: 5,838 manual FR translations of 1,405 EN puns
- Test data: 832 manual FR translations of 376 EN puns
- In 2023, success rate of wordplay translation was extremely low for both language pairs (EN$\to$FR, EN$\to$ES)

Frame Title

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

Figure 3: Histogram of translation references in French per English pun (TEST)

Top easiest punning words

EN	FR
Martians welcome. We have space for everyone.	Bienvenue les extraterrestres ! Installez vous, on a créé ces espaces détente pour vous.
A lot of trees were dying, but they needed to figure out the root of the problem.	De nombreux arbres mouraient mais personne ne trouvait la racine du mal qui les rongeait.
She was suspected of stealing a brooch but they couldn't pin it on her.	Elle s'est fait épingler pour une histoire de broche volée.
Well drilling is a deep subject.	Le forage de puits est un sujet profond.
The inept mathematician couldn't count on his friends.	Un mathématicien qui ne peut compter sur ses amis n'est pas un mathématicien...

Task 3: O cial Results

11 teams, 23 runs

1 team, 3 runs EN ES

	BLEU						BERT_Score			
run_id	count	Score	n_1	n_2	n_3	n_4	count	P	R	F1
Arampatzis_GoogleTranslate	376	65.23	78.96	67.48	61.59	57.52	832	91.93	91.82	91.85
Frane_TranslationModel	92	57.13	64.33	58.41	54.66	51.85	279	92.06	91.53	91.77
Dajana&Kathy	376	58.45	71.94	60.27	54.11	49.73	832	91.35	91.00	91.15
UBO_SDL	312	13.17	71.90	57.17	49.13	43.24	598	90.13	90.21	90.15
Tomislav&Rowan_MarianMT	376	58.85	77.11	63.66	56.06	50.45	832	90.82	89.19	89.95
Arampatzis_MarianMT	376	58.85	77.11	63.66	56.06	50.45	832	90.82	89.19	89.95
UBO_ChatGPT	312	13.09	69.90	54.08	46.07	40.31	598	89.12	89.34	89.21
UBO_DeepL	312	11.97	68.53	50.32	41.38	35.11	598	89.06	89.31	89.16
UAms_T5-base_ft	376	48.74	71.75	54.57	45.18	38.05	832	89.53	88.52	89.00
Arampatzis_mBART	376	48.71	70.95	54.40	45.29	38.67	832	88.95	87.41	88.13
Arampatzis_M2M100	376	42.37	68.46	48.73	37.72	29.93	832	88.23	87.23	87.70
UAms_Marian_ft	376	25.69	47.05	28.47	20.74	15.69	832	81.06	82.53	81.74
Farhan_2	376	14.33	23.68	15.84	12.05	9.32	832	69.38	77.14	72.96
Farhan_1	376	9.21	15.92	9.97	7.65	5.92	832	64.30	73.18	68.41
jokester_MarianMT	49	0.29	15.34	0.14	0.08	0.04	112	67.30	66.38	66.80
Arampatzis_opus_mt	63	0.29	15.04	0.23	0.06	0.03	157	66.98	66.05	66.47
Arampatzis_T5	63	0.32	11.35	0.17	0.10	0.06	157	65.91	64.79	65.31

BLEU scores (train)

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

Presence of identi ed 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%

Histogram of distinct pun locations in FR per EN pun (train)

BLEU scores EN ! ES (train)

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- netuned	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

BERT scores EN ! ES (train)

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- netuned	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%

Task 3: Observations

- Participants mainly used LLMs, commercial machine translation engines, and out-of-the-box translation models
- Only a small percentage of translations contain at least one word identified as carrying multiple meanings in references despite high BLEU and BERT Scores
- 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
- The success rate of wordplay translation remains low

JOKER Sessions at CLEF 2024

Date	Event
Sep 10 16:40-18:10	*Participant's talks (1x) (w/ SimpleText)*
Sep 11 14:00-15:30	Overview Talks JOKER Task 1-3 *Participant's talks (4x)*
Sep 11 16:00-18:00	*Participant's talks (7x)*
Sep 12 11:15-12:45	Keynote Pavel Braslavski (Nazarbayeb) on *What will we be laughing about tomorrow?* *Participant's talks (1x)* Planning Session JOKER 2025 *Humor-aware IR, Wordplay translation, Funny names, Controlled creativity?*

- Please join the JOKER sessions in Room 2!

Thank you !

See you at our track !

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

E-mail : contact@joker-project.com

Twitter : https://twitter.com/joker_research

Google group : https://groups.google.com/g/joker-project

[1]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. JOKER is supported by *La Maison des sciences de l'homme en Bretagne*.