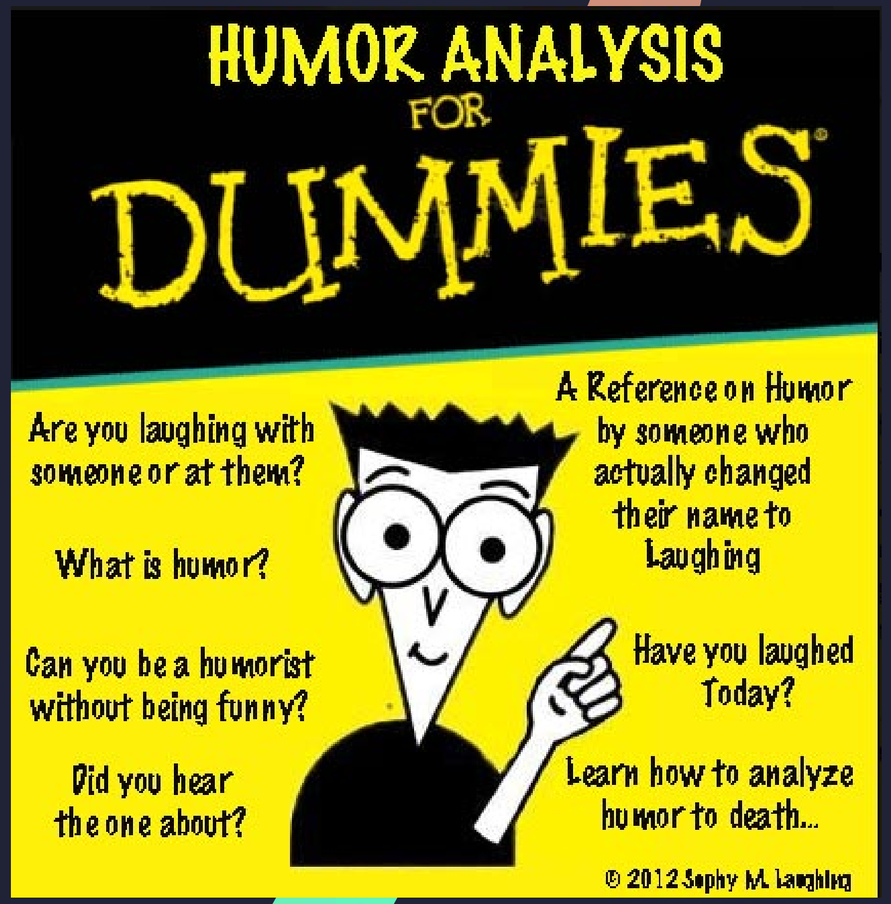
UNIVERSITY OF AMSTERDAM AT THE CLEF 2025 JOKER TRACK

Jaap Kamps, Jan Bakker, Cem Selçuk, Finley Helms, Alecsandru Kreefft-Libiu

University of Amsterdam

CLEF 2025



Motivation

- Understanding context in humour
- Bridging cross-cultural boundaries

- Expanding on the topic
  - Continuing on 2024 participation
  - Human Based Evaluation
  - New Scoring System

Experiments Humor-Aware Information Retrieval and Humor-Aware Machine Translation

Table 1
CLEF 2025 JOKER Track Submissions

	Run	Description
1	UAms_en_bm25	BM25 baseline (Anserini, stemming)
1	UAms_en_rm3	RM3 baseline (Anserini, stemming)
1	UAms_en_bm25_CE1K	BM25 + Crossencoder top 1,000
1	UAms_en_rm3_CE1K	BM25/RM3 + Crossencoder top 1,000
1	UAms_RM3	Okapi BM25/RM3
1	UAms_RM3RoBERTa	BM25/RM3 + Filter on RoBERTa Pun classifier (keeps 90%)
1	UAms_RM3RoBERTa_drop60	BM25/RM3 + Filter on RoBERTa Pun classifier (keeps 40%)
1	UAms_pt_bm25	BM25 baseline (Anserini, stemming)
1	UAms_pt_rm3	RM3 baseline (Anserini, stemming)
1	UAms_pt_bm25_CE1K	BM25 + Crossencoder top 1,000
1	UAms_pt_rm3_CE1K	BM25/RM3 + Crossencoder top 1,000
2	UvA_finetunedmBARTcc25	mBART-large-cc25 Finetuned
2	UvA_mBARTcc25&finetunedroBERTa	mBART-large-cc25 + humour detector roBERTa-large filter
2	UvA_finetunedT5-base	T5-base Finetuned
2	UvA_T5-base&finetunedroBERTa	T5-base + humour detector roBERTa-large filter
2	UvA_finetunedNLLB-1.3B	NLLB-200-1.3B Finetuned
2	UvA_finetunedNLLB-1.3B&finetunedroBERTa	NLLB-200-1.3B + humour detector roBERTa-large filter
2	UvA_finetunedMarianMT	MarianMT Finetuned
2	UvA_finetunedMarianMT&finetunedroBERTa	MarianMT + humour detector roBERTa-large

Finding Humour

Using Information Retrieval techniques

Evaluate on Humour

- Standard lexical and neural rankers retrieved topically relevant but often non-humorous texts
- To address this, we trained and applied a pun classifier to filter out non-humorous results
- Filtering improved performance across metrics (precision, recall, MRR, NDCG, MAP)
- The more aggressive filter (40%) performed better than the moderate one (90%).

Table 2
Evaluation of JOKER Task 1 (test data).

Run	MAP	GMAP	P@R	MRR	Precision				NDCG
					5	10	100	1,000	5
UAms_en_bm25	11.91	5.64	12.23	26.28	12.95	12.71	5.94	0.91	14.00
UAms_en_rm3	10.76	5.35	10.56	25.03	11.88	12.22	6.14	1.03	12.37
UAms_en_bm25_CE1K	4.88	2.60	2.47	5.68	0.48	2.75	4.30	0.91	0.38
UAms_en_rm3_CE1K	4.78	2.67	2.32	5.48	0.39	2.51	4.32	1.03	0.27
UAms_RM3	15.02	7.22	19.53	40.87	24.35	20.00	6.22	0.90	25.66
UAms_RM3RoBERTa	14.94	7.16	19.56	42.47	25.51	19.61	6.07	0.87	26.76
UAms_RM3RoBERTa_drop60	16.72	7.04	23.05	54.46	30.82	23.09	5.98	0.70	1.52
UAms_pt_bm25	7.89	0.19	5.96	9.83	5.22	6.09	3.03	0.33	5.13
UAms_pt_rm3	6.54	0.25	5.91	9.51	4.64	5.65	2.78	0.38	4.47
UAms_pt_bm25_CE1K	3.84	0.12	1.99	4.77	1.16	3.04	2.35	0.33	0.91
UAms_pt_rm3_CE1K	4.16	0.19	2.47	5.20	1.45	3.19	2.41	0.38	1.34

Translating Humour

Humour-aware Machine Translation (based on IR)

Humour Aware Machine Translation

- Building upon the foundation laid in 2024 by our peers at the UvA
- Modular scoring system evaluates multiple factors and selects the best translation based on balanced combined score

Table 4
CLEF 2025 JOKER Task 2: Example (id_en: en_1007 in train, and en_1 in test)

Run	Text
Source	Save the whales, spouted Tom.
Reference(s)	"Il faut sauver les baleines," jeta Tom avant de se tasser. "Il faut sauver les baleines," interjeta Tom. Moi je sauve les baleines, Tom s'en venta. Louis évent-a le projet de sauvetage des baleines. "Sauvez les baleines," proclama Tom à tout évent. "Sauvez les baleines, cracha Toto, Cétacé!"
UvA_finetunedmBARTcc25	"Les baleines, ..., les baleines, les baleines."
UvA_mBARTcc25&finetunedroBERTa	"Les baleines, les baleines, s'exclama Tom"
UvA_finetunedT5-base	"Sauver les baleines, dit Tom."
UvA_T5-base&finetunedroBERTa	"Sauver les baleines, dit Tom."
UvA_finetunedNLLB-1.3B	"Sauvez les baleines, s'exclama Tom."
UvA_finetunedNLLB-1.3B&finetunedroBERTa	"Sauvez les baleines, s'exclama Toto."
UvA_finetuneMarianMT	"Sauver les baleines, a filé Tom."
UvA_finetuneMarianMT&finetunedroBERTa	"Sauver les baleines, a lancé Tom."

Table 3
CLEF 2025 JOKER Task 2: Test results

Run	Location	BLEU	Precisions				BERTScore		
			1	2	3	4	P	R	F1
UvA_finetunedmBARTcc25	3.80	16.55	39.64	19.49	12.22	7.95	79.48	79.79	79.59
UvA_mBARTcc25&finetunedroBERTa	5.29	18.20	40.86	21.09	13.74	9.26	80.07	80.00	80.00
UvA_finetunedT5-base	6.30	36.77	60.29	40.58	30.94	24.15	86.69	86.24	86.44
UvA_T5-base&finetunedroBERTa	6.36	36.14	59.75	39.92	30.30	23.61	86.42	86.12	86.24
UvA_finetunedNLLB-1.3B	6.36	42.55	64.74	46.26	36.70	29.83	87.85	87.04	87.42
UvA_finetunedNLLB-1.3B&finetunedroBERTa	6.60	41.80	63.86	45.49	36.01	29.17	87.55	86.96	87.23
UvA_finetunedMarianMT	6.78	41.19	63.37	44.74	35.31	28.76	87.72	86.92	87.24
UvA_finetunedMarianMT&finetunedroBERTa	6.78	40.85	62.90	44.38	35.03	28.49	87.50	86.78	87.11

Re-Evaluated Scoring System

- Building upon the foundation laid in 2024 by our peers at the UvA
- Modular scoring system evaluates multiple factors and selects the best translation based on balanced combined score

```
scored_translations = []
for translation, gen_score in zip(translations, sequence_scores):
    try:
        translation_has_pun, translation_pun_prob = self.detect_pun(translation)

        pun_match_score = 1.0 if source_has_pun == translation_has_pun else 0.0
        combined_score = (0.6 * gen_score +
                          0.3 * pun_match_score +
                          0.1 * translation_pun_prob if source_has_pun else 0.1)

        scored_translations.append({
            'text': translation,
            'gen_score': float(gen_score),
            'has_pun': translation_has_pun,
            'pun_prob': float(translation_pun_prob),
            'pun_match': bool(pun_match_score),
            'combined_score': float(combined_score)
        })
    except Exception as e:
        logger.warning(f"Error scoring translation: {translation}. Error: {str(e)}")
        continue

if not scored_translations:
    return None, []

# Sort and get top 3
scored_translations.sort(key=lambda x: x['combined_score'], reverse=True)
top_translation = scored_translations[0]['text'] if scored_translations else None
top_3 = scored_translations[:3]

return top_translation, top_3
```

Human Based Evaluation

- Asked two human evaluators to rank machine translated sentences based on grammar, pun retention and context preservaton
- Ranking 1 through 8 for 15 English sentences, hence evaluating 120 machine translated sentences.

id_en	Translation	Grammar Error?	Pun sentence? Y/N	Meaning preserved? Y/N
en_XXXX	English sentence			
	MarianMT			
	MarianMT+RoBERTa			
	mBARTcc25			
	mBARTcc2+RoBERTa			
	NLLB-200			
	NLLB-200+RoBERTa			
	T5-base			
	T5-base+RoBERTa			

Table 4.3: Human evaluation table format used in experiment 3

en_38	The mortician was late for dinner because he was buried in his work.	ranking 1-8	Grammar error	Pun? Y/N	Preserved meaning? Y/N
	"Le morticien était en retard pour dîner parce qu'il était enterré dans son travail."	1		Y	Y
	"Le morticien était en retard pour dîner car il était enterré dans son travail."	2		Y	Y
	"Les morticiens étaient en retard pour le dîner parce qu'ils étaient enfermés dans leur travail."	6	plural	N	N
	"Un mortician était en retard pour le dîner parce qu'il était enterré dans son travail"	3	mortician is English	Y	Y
	"Le pompier était en retard pour le dîner parce qu'il était enterré dans son travail."	8	pompier?	N	N
	"Le pompier était en retard pour le dîner parce qu'il était enterré dans son travail."	7	pompier?	N	N
	"Le mortician était en retard pour le dîner parce qu'il était enterré dans son travail."	4	mortician is English	Y	Y
	"Le mortician était en retard pour le dîner parce qu'il était enterré dans son travail."	5	mortician is English	Y	Y
en_42	Use your own toothbrush!', Tom bristled.	ranking 1-8	Grammar error	Pun? Y/N	Preserved meaning? Y/N
	""Utilisez votre propre brosse à dents !", dit Tom."	1		Y	Y
	""Utilisez votre propre brosse à dent !", dit Tom."	3	dent not plural	Y	Y
	"C'était une bonne idée de prendre son propre brosse à dents."	8	wrong	N	N
	"La brosse à dents est une brosse à dents", dit Tom en s'agrippant à sa brosse."	7		N	N
	""Utilise ta propre brosse à dents !" s'écria Tom, poilu."	5		Y	Y
	""Utilise ta propre brosse à dents !" s'écria Tom, poilu."	4		Y	Y
	""Utilisez votre propre brosse à dents ! ", dit Tom."	2		Y	Y
	""Utilisez votre propre brosse à dents! ", disait Tom."	6		Y	Y

BLEU-based Evaluation

- Automatic ranking of sentences based on their respective BLEU-score relative to the correct French translation
- Kendall-Tau correlation computation for statistical analysis regarding human & machine agreement on rankings

en_31	The trucker explained that he was early because he had no breaks.	ranking 1-8	BLEU
	"Le camionneur a expliqué qu'il était en avance parce qu'il n'avait pas eu de pause."	4	2.67
	"Le camionneur a expliqué qu'il était en avance car il n'avait pas eu de pauses."	3	4.82
	"Les employés du camion ont dit que le chauffeur était tard parce qu'il n'avait pas eu de problèmes."	8	2.41
	"Un camionnier a dit que c'était tôt parce qu'il n'avait pas eu de retard."	5	2.65
	"Le chauffeur de camion a expliqué qu'il était en avance car il n'avait pas eu de pause."	1	4.83
	"Le chauffeur de camion a expliqué qu'il était en avance car il n'avait pas eu de pause."	1	4.83
	"Le camionneur a expliqué qu'il était tôt parce qu'il n'avait pas eu de pause."	5	2.65
	"Le camionneur a expliqué qu'il était tôt parce qu'il n'avait pas eu de pause."	5	2.65
en_38	The mortician was late for dinner because he was buried in his work.	ranking 1-8	BLEU
	"Le morticien était en retard pour dîner parce qu'il était enterré dans son travail."	1	56.59
	"Le morticien était en retard pour dîner car il était enterré dans son travail."	7	34.79
	"Les morticiens étaient en retard pour le dîner parce qu'ils étaient enfermés dans leur travail."	8	7.31
	"Un mortician était en retard pour le dîner parce qu'il était enterré dans son travail"	6	49.01
	"Le pompier était en retard pour le dîner parce qu'il était enterré dans son travail."	4	52.66
	"Le pompier était en retard pour le dîner parce qu'il était enterré dans son travail."	4	52.66
	"Le mortician était en retard pour le dîner parce qu'il était enterré dans son travail."	2	55.55
	"Le mortician était en retard pour le dîner parce qu'il était enterré dans son travail."	2	55.55

Comparison	Mean τ	Variance	Std Dev
Evaluator A vs B	0.672	0.056	0.237
Evaluator A vs Auto	0.392	0.160	0.400
Evaluator B vs Auto	0.420	0.159	0.399

Table 5.9: Kendall-Tau correlation statistics between evaluation methods

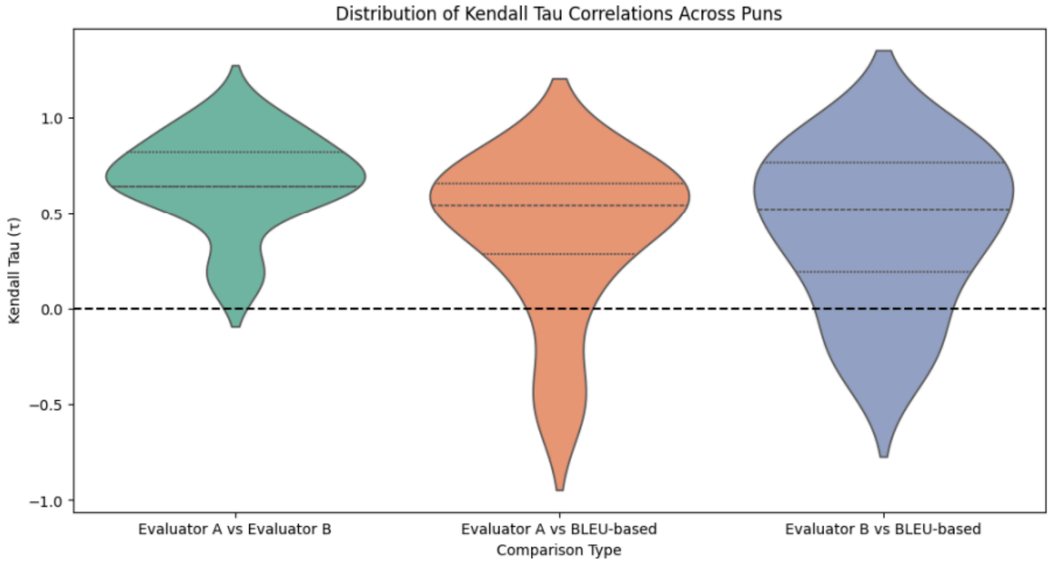


Figure 5.1: Kendall-Tau Correlation violin plot showing the agreement between different evaluators.

The happenings of Searching & Translating Humour

- Developed effective humor detection classifiers for English and French.
- Humor-aware filtering improved retrieval effectiveness.
- Pun detection combined with translation models increased pun preservation (up to +94% higher pun scores in lower-ranked outputs).
- Fine-tuning on the CLEF JOKER corpus strongly boosted performance (lower losses, higher BLEU, better grammaticality).
- MarianMT and NLLB emerged as best-performing models, balancing grammar, meaning, and humor.
- Automatic metrics like BLEU correlated only moderately with human judgment ($\tau = 0.39–0.42$), while human agreement was high ($\tau = 0.67$).

- Trade-off between humor preservation and grammatical correctness (e.g., T5 creative but less grammatical).
- Some architectures, like mBARTcc25, are fundamentally less suitable for humor translation.
- Automatic metrics fail to capture creativity and humor, highlighting the need for humor-specific evaluation methods.
- Humor-focused fine-tuning may reduce generalizability to broader domains.
- Limited human evaluation scale (2 annotators, 120 sentences).

Q&A

HUGE THANKS

To Jaap Kamps, Jan Bakker, Cem Selçuk, Finley Helms, Alecsandru Kreefft-Libiu and the CLEF organization!