University of Amsterdam at the CLEF 2024 JOKER Track

Jaap Kamps, Emma Schuurman, Mick Cazemier, Luc Buijs
University of Amsterdam

CLEF 2024 SimpleText Track, September 11, 2024, Grenoble, France

Motivation
Is this a joke?

- State of the art AI, NLP AI models cannot cope with **humor or other non-literal meaning** in text
- Impossible to learn from **text usage alone!**
- Related to the **surface structure** of the utterance (orthography) and not the **deeper semantics**
- Important for **jokes**, but also to understand **cultural references**, or detect **harassment** and **bullying**

What Happens When Searching, Classifying and Translating Humor?

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

Task	Run	Description
1	UAms_Task1_Anserini_bm25	BM25 baseline (Anserini, stemming)
1	UAms_Task1_Anserini_rm3	RM3 baseline (Anserini, stemming)
1	UAms_Task1_bm25_CE100	BM25 + Crossencoder top 100
1	UAms_Task1_rm3_CE100	BM25/RM3 + Crossencoder top 100
1	UAms_Task1_bm25_BERT_Filter	BM25 + Filter on BERT WordPlay classifier (keeps 76%)
1	UAms_Task1_rm3_BERT_Filter	BM25/RM3 + Filter on BERT WordPlay classifier (keeps 46%)
1	UAms_Task1_rm3_T5_Filter1	BM25/RM3 + Filter on WordPlay classifier (keeps 53%)
1	UAms_Task1_rm3_T5_Filter2	BM25/RM3 + Filter on WordPlay classifier (keeps 43%)
2	UAms_Task2_BERT_ft	BERT classifier (fine-tuned)
3	UAms_Task3_Marian_ft	Marian Finetuned
3	UAms_Task3_T5-base_ft	T5-base Finetuned

Finding Humor?

#1 Topically relevant versus humorous text

Evaluate on Humor (top) or Relevant (bottom)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
UAms_Task1_Anserini_bm25	0.1906	0.1167	0.1583	0.1361	0.1008	0.1598	0.2272	0.2376	0.1582
UAms_Task1_bm25_CE50	0.1248	0.0833	0.0750	0.1028	0.0697	0.0683	0.1498	0.1155	0.0668
UAms_Task1_bm25_CE100	0.1233	0.0833	0.0750	0.0889	0.0685	0.0682	0.1300	0.0922	0.0702
UAms_Task1_bm25_CE1000	0.1039	0.0833	0.0750	0.0806	0.0660	0.0666	0.1188	0.0687	0.0898
UAms_Task1_Anserini_rm3	0.2407	0.1667	0.1750	0.1250	0.1506	0.1896	0.2339	0.2989	0.1725
UAms_Task1_rm3_CE50	0.1259	0.1000	0.0833	0.1056	0.0806	0.0754	0.1582	0.1233	0.0662
UAms_Task1_rm3_CE100	0.1231	0.0833	0.0917	0.1028	0.0685	0.0801	0.1422	0.0921	0.0712
UAms_Task1_rm3_CE1000	0.1038	0.0833	0.0667	0.0833	0.0660	0.0618	0.1238	0.0837	0.0957
UAms_Task1_Anserini_bm25	0.6597	0.5500	0.5333	0.5111	0.3182	0.3477	0.4125	0.6510	0.3503
UAms_Task1_bm25_CE50	0.8917	0.5833	0.5167	0.5056	0.3453	0.3267	0.3976	0.2897	0.1622
UAms_Task1_bm25_CE100	0.8056	0.5167	0.5000	0.4917	0.3048	0.3076	0.3757	0.3655	0.1959
UAms_Task1_bm25_CE1000	1.0000	0.5500	0.5083	0.5083	0.3435	0.3312	0.3935	0.6510	0.3639
UAms_Task1_Anserini_rm3	0.7282	0.5833	0.5250	0.4944	0.3686	0.3659	0.4105	0.6682	0.3528
UAms_Task1_rm3_CE50	0.8917	0.6000	0.5167	0.4861	0.3562	0.3312	0.3930	0.2847	0.1590
UAms_Task1_rm3_CE100	0.8056	0.5167	0.5167	0.5111	0.3048	0.3198	0.3907	0.3652	0.1972
UAms_Task1_rm3_CE1000	1.0000	0.5500	0.5000	0.5056	0.3435	0.3262	0.3951	0.6682	0.3682

- Neural rankers work on topical relevance, but fail dramatically on humor

#1 Relevant + Humorous

Humor-aware IR is different from topical relevance

Detecting Humor?

#2 Can we detect humorous text?

CLEF 2023 Joker Task 1: Pun Detection Revisited

- General approach to the Joker Track:
  - What if we can **detect humorous text**?
- If successful, we can create:
  - **Humor-aware Information Retrieval** by filtering results of standard IR model
  - **Humor-aware Machine Translation** by selecting from candidate translations
- Problem: Pun Detection proved very hard
  - Best 2023 system **F1 of 53.61%** — on a binary classification problem!

CLEF 2023 Joker Task 1: Pun Detection (English)

Evaluation of the CLEF 2023 Joker Pun Detection Task (English)

Model	F1 Score	Precision	Recall	Accuracy
BERT	0.70	–	–	0.72
SimpleT5_V1	0.80	0.72	0.90	0.76
SimpleT5_V2	0.80	0.74	0.87	0.77

- Pun detection is hard but “works”:
    - F1 of 80% on hold out/unseen data
    - Requiring safeguards against overfitting!
    - Best performing 2023 model F1 of 0.5361 for English
    - Majority class prediction F1 of 50% (test) and 58% (train).

CLEF 2023 Joker Task 1: Pun Detection (French)

Evaluation of the CLEF 2023 Joker Pun Detection Task (French) on 10% hold-out (top) and train/test data (bottom)

Model	n	Accuracy	Precision	Recall	F1 Score
Dummy-Model	399	0.49	0.47	0.52	0.50
DistilBERT-base	399	0.71	0.71	0.69	0.68
DistilBERT-FT1	399	0.72	0.71	0.69	0.70
DistilBERT-FT2	399	0.70	0.57	0.75	0.65
DistilBERT (FT1) *train*	3,999	0.9395	0.9475	0.9304	0.9389
DistilBERT (FT1) *test*	17,791	0.7518	0.7189	0.7009	0.7098

- Pun detection is hard but "works":
  - F1 of 71% on hold out/unseen data — majority class prediction F1 of 50%
  - Best performing 2023 model F1 of 0.6645 for French.

#2 We can detect humorous text!

Can we exploit effective humor detection?

Searching for Humor?

#3 Humor-aware IR based on humor detection

Humor-Aware Information Retrieval

Run	MRR	Precision		Recall			NDCG	Bpref	MAP
		5	10	5	10	20			
UAms_Task1_Anserini_bm25	0.1873	0.0489	0.0556	0.0564	0.0819	0.1624	0.2417	0.0928	0.0800
UAms_Task1_Anserini_rm3	0.1977	0.0578	0.0622	0.0611	0.0830	0.1511	0.2677	0.0921	0.0845
UAms_Task1_bm25_CE100	0.0762	0.0356	0.0267	0.0332	0.0388	0.0964	0.1749	0.0610	0.0416
UAms_Task1_rm3_CE100	0.0749	0.0356	0.0267	0.0332	0.0388	0.0967	0.1769	0.0602	0.0410
UAms_Task1_bm25_BERT_Filter	0.1883	0.0489	0.0844	0.0590	0.1165	0.1822	0.2430	0.1173	0.0878
UAms_Task1_rm3_BERT_Filter	0.2668	0.1111	0.1156	0.0882	0.1436	0.2079	0.2739	0.1608	0.1156
UAms_Task1_rm3_T5_Filter1	0.2283	0.0933	0.1111	0.0861	0.1478	0.1943	0.2651	0.1628	0.1077
UAms_Task1_rm3_T5_Filter2	0.2604	0.1067	0.1289	0.0882	0.1508	0.2261	0.2820	0.1841	0.1207

- Lexical rankers work OK'ish, but neural zero-shot rerankers fail
  - Filtering using pun detection leads to significant improvement on all measures and all topics!

#3 Humor-aware IR works!

Humor-aware IR based on effective humor detection

Translating Humor?

#4 Humor-aware MT based on humor detection

Humor-Aware Machine Translation

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!"
UAms_Task3_Marian_ft	"Sauvez les baleines," proclama Tom à tout évent.
UAms_Task3_T5-base_ft	"Sauvez les baleines," dit Tom.

- Translating wordplay very hard, for humans and machines!
  - Some MT candidates match reference translations
  - But references differ quite a lot, and share many words with literal translations
  - So careful to interpret text overlap measures...

Humor-Aware Machine Translation

Run	n	BLEU	Precisions				Length		BERTScore			
			1	2	3	4	Rat.	Tok.	n	P	R	F1
Reference (test)	376	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5,774	834	1.0000	1.0000	1.0000
MarianMT (optimized)	376	0.5100	0.7169	0.5480	0.4520	0.3810	1.042	6,015	834	0.8985	0.8965	0.8973
MarianMT/Pun Detector	376	0.4663	0.6902	0.5085	0.4061	0.3318	1.050	6,061	834	0.8853	0.8849	0.8849

- Humor-aware machine translation
  - Careful to generate 5 candidate translations with sufficient variation
    - otherwise hit or miss (all are puns, or none are puns)
  - We filter out the single candidate with the highest expected pun detection score
    - we pick a candidate with a lower translation score in 50% of the cases
  - Evaluation on BLEU and BERTScore looks only at word overlap
    - Scores go down a little, but we score much higher on the pun detector
    - Human inspection of small sample where we pick a next candidate supports this: many are puns.

#4 Humor-aware MT works?

Humor-aware MT based on effective humor detection

Classifying Humor?

#5 Is it irony, sarcasm, exaggeration, or not funny at all...

Task 2: Classifying Humor

Run	Accuracy	Macro			Weighted		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
UAms_Task2_BERT_ft	0.6561	0.6286	0.6090	0.5672	0.6752	0.6561	0.6254
UAms_Task2_BERT_ft	0.6330	0.5724	0.5845	0.5221	0.6605	0.6330	0.6021

- We also participated in CLEF 2024 Joker Task 2
  - Multi-class prediction problem: incongruity-absurdity (AID), exaggeration (EX), irony (IR), sarcasm (SC), self-deprecating (SD), and wit-surprise (WS).
  - Luke-warm results, OK'ish diagonal in confusion matrix
  - Our model systematically miss-classifies sentences labeled as “irony” with “sarcasm” and “exaggeration”
  - Examples seem to contain elements of irony (typically about a situation and an opposite expectation) and of sarcasm (a form of expression, assuming the utterance appeared in some conversational context), or elements of exaggeration in some sense

#5 What is (not) humor?

Need a rigorous taxonomy of humor

What Happens When Searching, Classifying, and Translating Humor?

#1 Humor-aware IR is different from topical relevance

#2 Can we exploit effective humor detection?

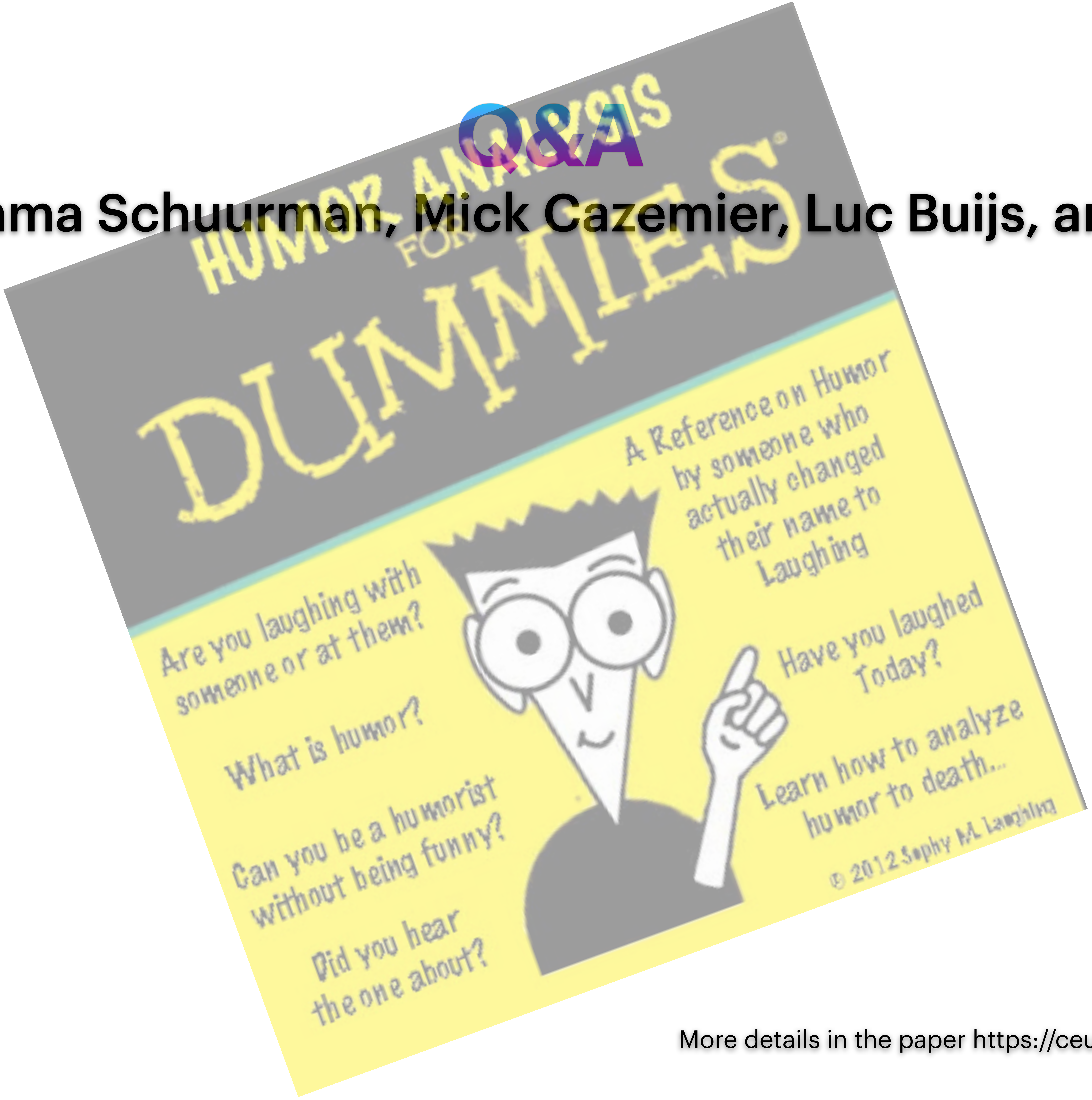
#3 Humor-aware IR based on humor detection

#4 Humor-aware MT based on humor detection

#5 Need a rigorous taxonomy of humor

Q&A

Thanks to Emma Schuurman, Mick Cazemier, Luc Buijs, and David Rau!



More details in the paper https://ceur-ws.org/Vol-3740/paper-181.pdf