

Computational Humor: Approaches, Data, Evaluation

Pavel Braslavski

27.10.2025

Tasks and Scope

Computational Humor

- Humor detection / generation
- Theoretical / practical aspects

Tell me which are funny, which are not -- and which get a giggle first time but are cold pancakes without honey to hear twice.

–Robert Heinlein, *The Moon Is a Harsh Mistress*

You Told Me That Joke Twice, 2000

Source: <https://banksyexplained.com/you-told-me-that-joke-twice-2000/>



THE WEEKEND ESSAY

WHY A.I. ISN'T GOING TO MAKE ART

*To create a novel or a painting, an artist makes choices that are
fundamentally alien to artificial intelligence.*

By Ted Chiang

August 31, 2024



Humor Detection

Russian FUN Dataset [Blinov+, 2019]

- Funny part: VK.com & anekdot.ru
- Unfunny counterpart for each joke: top-ranked (BM25) non-identical forum post
- KL divergence
 - STIERLITZ: 0.50
 - FUN: 0.18

FUN Gold

- 94% correct

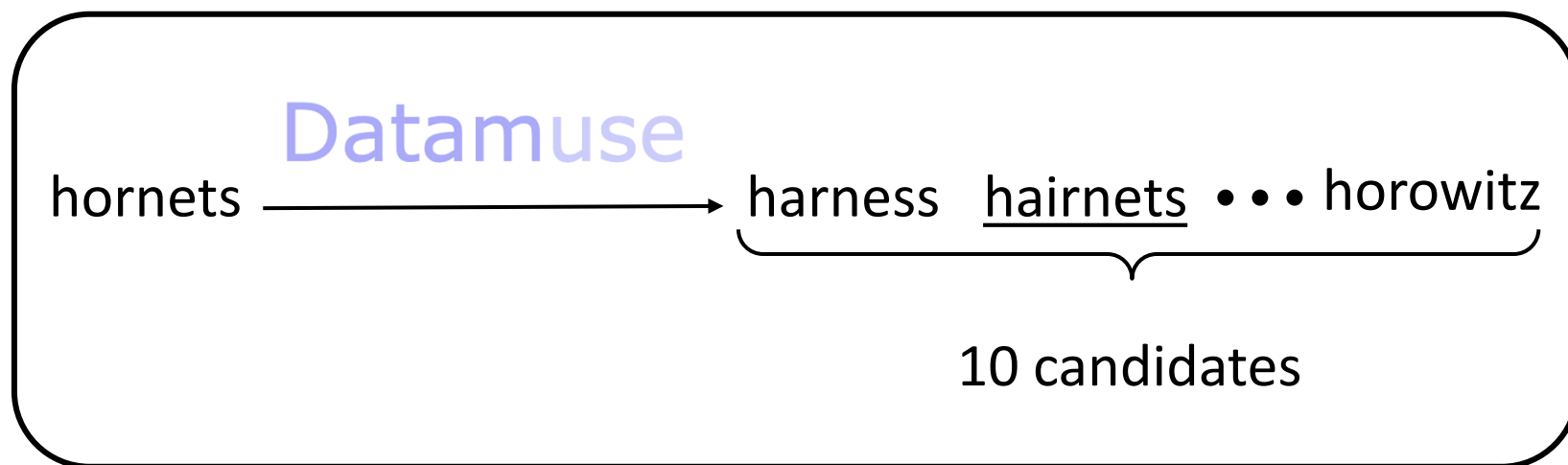
<i>Dataset</i>	<i>Jokes</i>	<i>Non-jokes</i>	<i>Total</i>
STIERLITZ	46,608	46,608	93,216
train	37,447	37,447	65,530
validation	4,682	4,682	9,364
test	9,361	9,361	18,722
PUNS	213	0	213
FUN	156,605	156,605	313,210
train	125,708	125,708	251,416
test	30,897	30,897	61,794
GOLD	899	978	1,877

Transferability [Baranov+, 2023]

	Model	16kOL	PotD	EnPuns	ShJ	ReJ	Haha	FL+HME	Unfun.me	NF	COMB	avg F1	TheO
RoBERTa fine-tuned on...	16kOL	0.92	0.30	0.67	0.63	0.68	0.77	0.28	0.38	0.91	0.65	0.62	0.57
	PotD	0.01	0.98	0.09	0.64	0.04	0.23	0.12	0.01	0.09	0.52	0.27	0.01
	EnPuns	0.53	0.74	0.92	0.53	0.67	0.76	0.70	0.61	0.85	0.69	0.70	0.91
	ShJ	0.73	0.72	0.62	0.93	0.57	0.82	0.37	0.28	0.92	0.78	0.67	0.38
	ReJ	0.36	0.13	0.18	0.21	0.70	0.56	0.22	0.16	0.44	0.29	0.33	0.25
	Haha	0.62	0.71	0.70	0.84	0.66	0.94	0.36	0.40	0.95	0.79	0.70	0.47
	FL+HME	0.03	0.50	0.03	0.51	0.26	0.17	0.98	0.31	0.33	0.40	0.35	0.13
	Uf.me	0.68	0.54	0.79	0.58	0.66	0.74	0.72	0.92	0.92	0.73	0.73	0.99
	NF	0.67	0.68	0.85	0.65	0.67	0.76	0.72	0.61	1.00	0.75	0.74	1.00
	COMB	0.64	0.97	0.54	0.85	0.65	0.93	0.57	0.97	0.96	0.92	0.80	0.99
	chatGPT	0.77	0.85	0.78	0.85	0.67	0.91	0.42	0.56	0.70	0.82	0.73	0.59
	FLAN	0.75	0.74	0.66	0.86	0.67	0.91	0.65	0.69	0.96	0.83	0.77	0.67

Adversarial attack [Baranov+, 2023]

Original: Law catches flies, but let's hornets go free.



Modified: Law catches flies, but let's hairnets go free.

Model	% of labels changed
16kOL	0.4
PotD	10.8
EnPuns	9.5
ShJ	3.5
ReJ	0.7
Haha	4.5
FL+HME	49.0
Unfun.me	2.3
NF	4.9
COMB	1.3

External datasets [Baranov+, 2023]

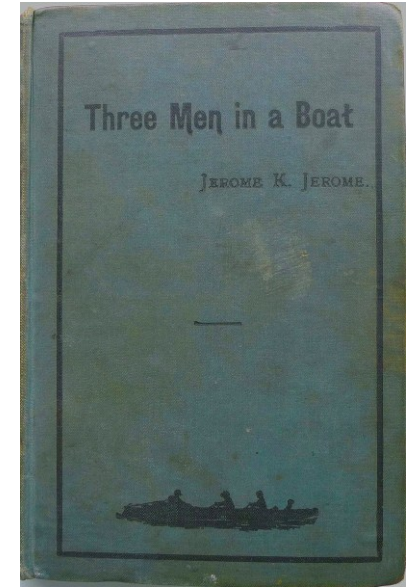
The Old Curiosity Shop (1841)
by Charles Dickens



Alice's Adventures in Wonderland
(1865) by Lewis Carroll



Three Men in a Boat (1889)
by Jerome K. Jerome



THE
WALKING DEAD

(2010-2022)

F•R•I•E•N•D•S

(1994-2004)

Top 5 models

Model	COMB _(ours)	FLAN	NF	Uf.me	chatGPT
macro avg F1	0.8	0.77	0.74	0.73	0.73
Behavior on external data					

KoWit-24 [Baranov+,2025]

	<i>Wordplay type</i>	<i>#</i>	<i>AAL</i>	<i>Links</i>
Puns	Polysemy	190	1.51	
	Homonymy	26	1.57	
	Phonetic similarity	98	1.80	
Trans.	Collocation	423	2.64	126
	Idiom	177	3.43	118
	Reference	353	3.73	214
	Nonce word	185	1.44	
	Oxymoron	48	2.02	

[Культура](#)

24.10.2025, 19:55



Шумы на оба ваши дома

Вышел в прокат корейский хоррор «Паранормальное явление. Шум»

Interpretation Results

<i>Wordplay type</i>	#	<i>GigaChat Lite</i>		<i>GigaChat Max</i>		<i>YaGPT4</i>		<i>Mistral</i>		<i>GPT-4o</i>	
		manual	auto	manual	auto	manual	auto	manual	auto	manual	auto
Polysemy	12	0.17	0.33	0.33↑	0.25	0.08	0.00	0.17	0.33	0.50	0.50
Homonymy	7	0.14	0.43	0.29	0.29	0.14	0.14	0.00	0.43	0.43 ↑	0.29
Phonetic similarity	85	0.11	0.25	0.28	0.39	0.11	0.21	0.15	0.32	0.52 ↑	0.51
Collocation	393	0.07	0.18	0.27	0.27	0.16	0.20	0.20	0.30	0.44 ↑	0.41
Idiom	164	0.15	0.18	0.37↑	0.30	0.32↑	0.28	0.34↑	0.30	0.55 ↑	0.48
Reference	326	0.10	0.12	0.25↑	0.23	0.20↑	0.16	0.23	0.25	0.46 ↑	0.36
Nonce word	166	0.15	0.46	0.29	0.57	0.28	0.43	0.28	0.57	0.61	0.69
Oxymoron	6	0.67	0.67	0.50	0.50	0.33	0.33	0.83	0.83	0.67↑	0.50
Total	1,159	0.11	0.19	0.28	0.28	0.20	0.22	0.24	0.30	0.48 ↑	0.43

Table 4: Recall on wordplay interpretation by type; manual and automatic evaluation (↑ marks cases, where manual score exceeds the automatic one).

Generation & Evaluation

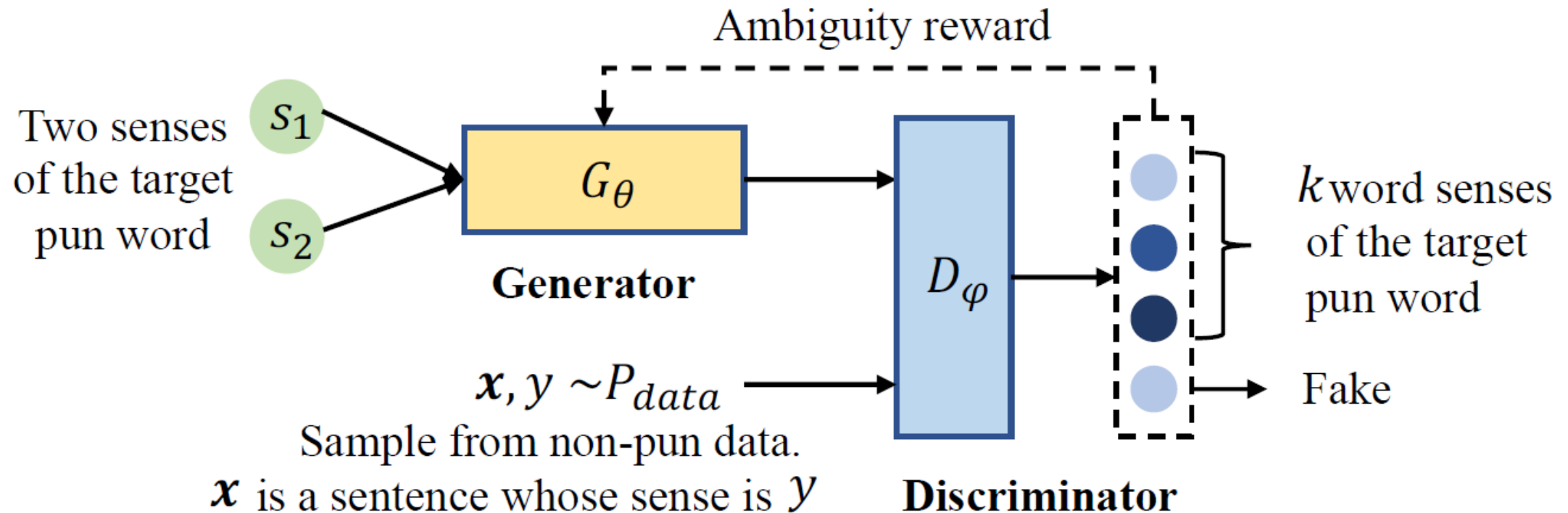
Puns

A pun is a form of wordplay in which one sign (e.g., a word or a phrase) suggests two or more meanings by exploiting polysemy, homonymy, or phonological similarity to another sign, for an intended humorous or rhetorical effect.

[Miller+, 2017]

- Most humor generation experiments deal with puns [Binsted & Ritchie, 1994]
- Approach: 1) devise templates, 2) generate jokes

Can Detectors Improve Generation? [Luo+, 2019]



Crowdsourced Evaluation [Braslavski+, 2018]

	Group	MT	# MT	V	# V	All	# All
Age Group	18–30	2.05	46	2.07	77	2.06	123
	31–40	1.89	37	2.04	54	1.98	91
	41–50	1.80	18	2.02	29	1.94	47
	51–60	1.84	10	1.97	6	1.89	16
	61+	1.73	1	3.33	1	2.53	2
Sex	Male	1.92	52	2.01	82	1.97	134
	Female	1.95	60	2.10	85	2.04	145
Language	Average	–	–	2.16	15	2.16	15
	Good	2.25	5	2.10	69	2.11	74
	Bilingual	2.11	3	2.06	39	2.07	42
	Native	1.91	104	1.95	44	1.92	148
	Global	1.93	112	2.06	167	2.01	279

Highest variation native/no-native:

- **Q:** *Why did 10 die?*
- **A:** *He was in the middle of 9/11*

Highest variation in male/female:

- **Q:** *What is the meaning of life?*
- **A:** *All evidence to date suggests it is chocolate.*

```
graph LR; A[(jokes dataset)] --> B[/LLM/]; B --> C{humor policy prompt}; C --> D["• Wordplay or pun  
• Misdirection  
• Exaggeration  
• Stereotyping  
• Satire  
• ..."]
```

The diagram illustrates a process flow for generating humor. It starts with a green cylinder labeled "jokes dataset". An arrow points from this cylinder to an orange chevron shape labeled "LLM". Another arrow points from the "LLM" shape to a blue diamond shape labeled "humor policy prompt". To the right of the diamond, there is a bulleted list: "• Wordplay or pun", "• Misdirection", "• Exaggeration", "• Stereotyping", "• Satire", and "• ...".

*There Is Nothing More
Practical Than a Good
Theory*
– Kurt Lewin

The diagram illustrates a feedback loop for generating jokes using Large Language Models (LLMs). It features the following components and flow:

- seed topic** (green cylinder): The starting point, labeled "event horizon".
- LLM** (orange chevron): Processes the seed topic to generate **Associations** (yellow rounded rectangle).
- Associations**: Provides context back to the LLM and informs the **humor policy prompt**.
- humor policy prompt** (blue diamond): A prompt that guides the LLM's output.
- LLM** (orange chevron): Takes the humor policy prompt and associations to generate **generated jokes** (blue cylinder).
- generated jokes**: The final output, shown as a sample joke: "- I told my friend that getting into a black hole is easy, but escaping is where things really start to suck."

Below the diagram, two lists provide context for the "event horizon" seed topic:

- 1. Black hole boundary
- 2. Event Horizon (1997 film)
- ...
- 1. The event horizon marks the delineation around a black hole beyond which nothing can escape.
- 2. The term "event horizon" gained popular culture prominence through the 1997 science fiction horror film titled "Event Horizon."
-

Conversational jokes [Blinov+, 2017]

Score	Stimulus	Response
3.00	Does evolution being a theory make it subjective?	There is no theory of evolution, just a list of creatures Chuck Norris allows to live.
2.67	Can you find oil by digging holes in your backyard?	Things to do today: 1.Dig a hole 2. Name it love 3. Watch people fall in it.
1.33	Why don't they put zippers on car doors?	Sick of doors that aren't trap doors.
0.67	What if you're just allergic to working hard?	You're not allergic to gluten.
0.33	What test do all mosquitoes pass?	My internal monologue doesn't pass the Bechdel test. :(

Easier to generate?

MWAHAHA: A Competition on Humor Generation

Can computers be funny?

[About](#)[Quick Links](#)[News](#)[Subtasks](#)[Dates](#)[Data](#)[Evaluation](#)[Organizers](#)[Contact](#)

While Humor Understanding has been the focus of many shared tasks, Humor Generation remains an even more challenging and largely unexplored frontier. **MWAHAHA**, which stands for Models Write Automatic Humor And Humans Annotate, is [SemEval 2026's Task 1](#) and is the first task dedicated to advancing the state of the art in Computational Humor Generation. We invite participants to develop systems capable of generating genuinely humorous content under various constraints.

Our goal is to push models beyond memorization and towards true humorous creativity. By using carefully designed constraints, we aim to ensure fairness in evaluation and encourage the generation of novel jokes. This task has significant implications for more engaging conversational AI, creative writing tools, and a deeper understanding of the complex nature of humor itself.

To participate (and download the input data), please visit [our CodaBench page](#).

