LOL NLP:
an Overview of
Computational Humor

Pavel Braslavski

05.03.2021

About myself

- Research/academia: Ural Federal University, Yekaterinburg/ HSE University, Moscow
- Past industrial experience: Yandex/SKB Kontur
- Recent research interests: question answering, fiction analysis, computational humor

Homepage: http://kansas.ru/pb/

Humor is a promising area for studies of intelligence and its automation: it is hard to imagine a computer passing a rich Turing test without being able to understand and produce humor.

West & Horvitz, AAAI2019

Humor at Alexa Prize competition

…it’s amazing to see that now humor is coming in… Good sense of humor is a sign of intelligence in my mind and something very hard to do.

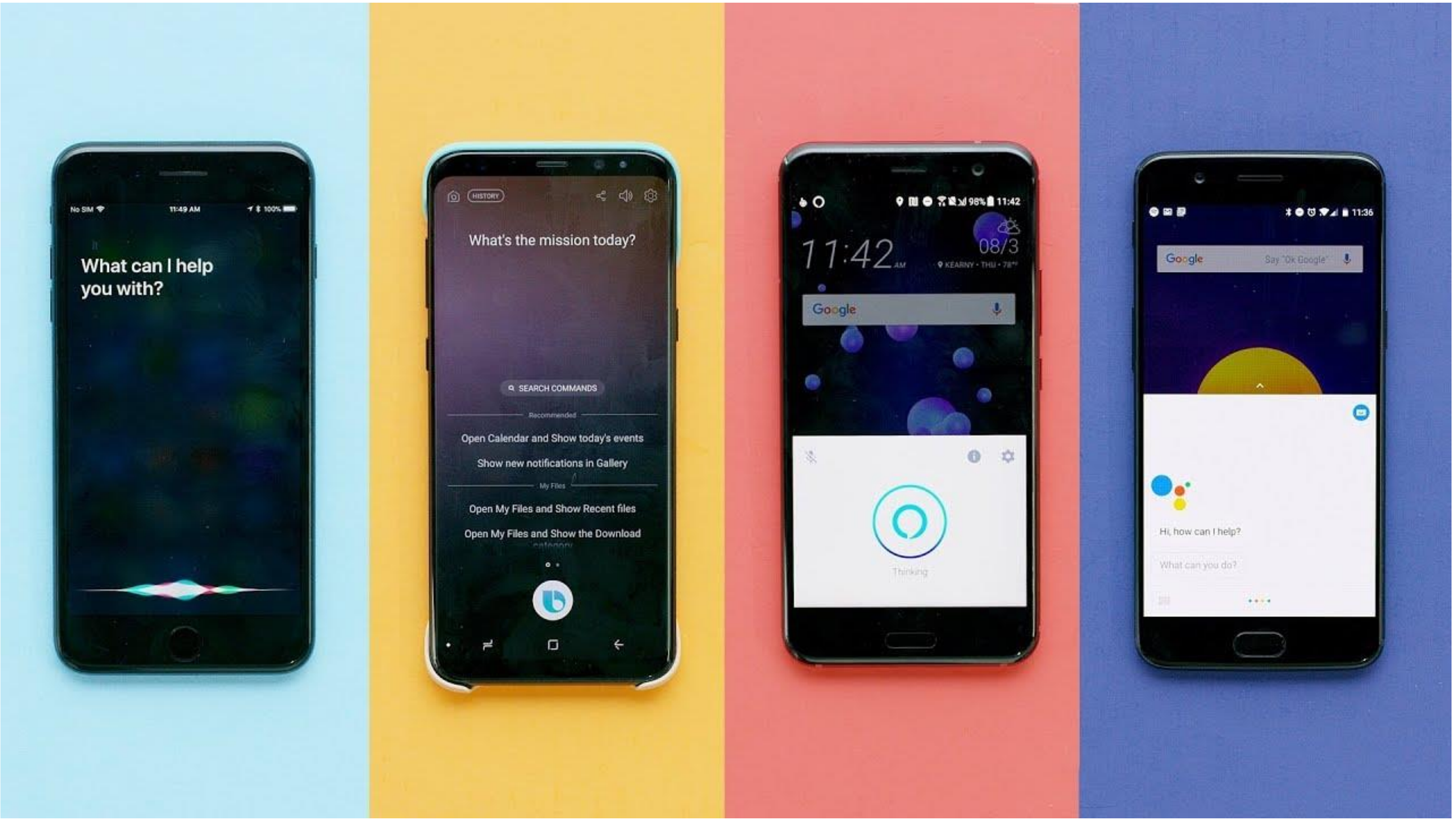
Rohit Prasad
vice president and head scientist of Amazon Alexa
in conversation with Lex Fridman, 14 December 2019

https://youtu.be/Ad89JYS-uZM

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

Motivation for Computational Humor


What can I help you with?
What's the mission today?
SEARCH COMMANDS
Open Calendar and Show today's events
Show new notifications in Gallery
Open My Files and Show Recent files
Open My Files and Show the Download
11:42
08/3
Google
Thinking
Google
Hi, how can I help?
What can you do?

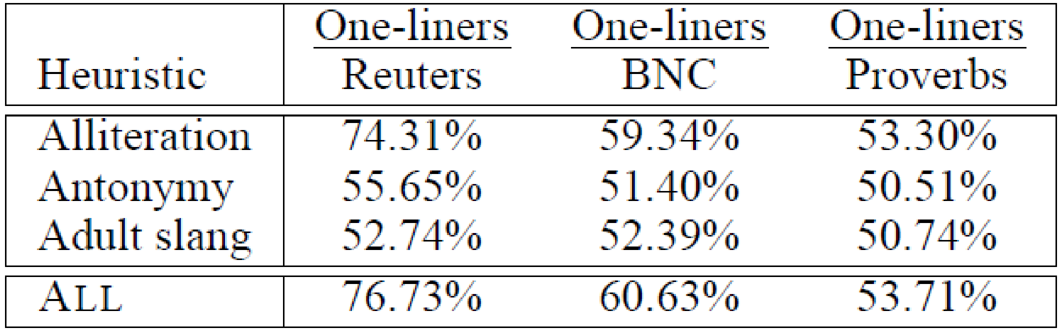
Humor…

- … detection
- … datasets
- … generation
- … evaluation

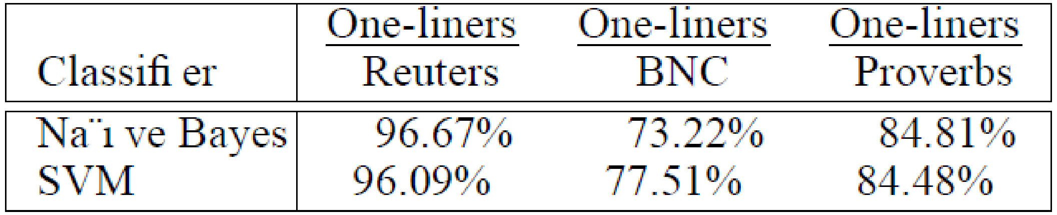
Humor Detection

Humor classifier [Mihalcea & Strapparava, 2005]

- 16K one-liners/16K non-funny sentences
- Features: alliteration/rhyme, antonymy (WordNet), adult slang, content words
- Classifiers: NB and SVM



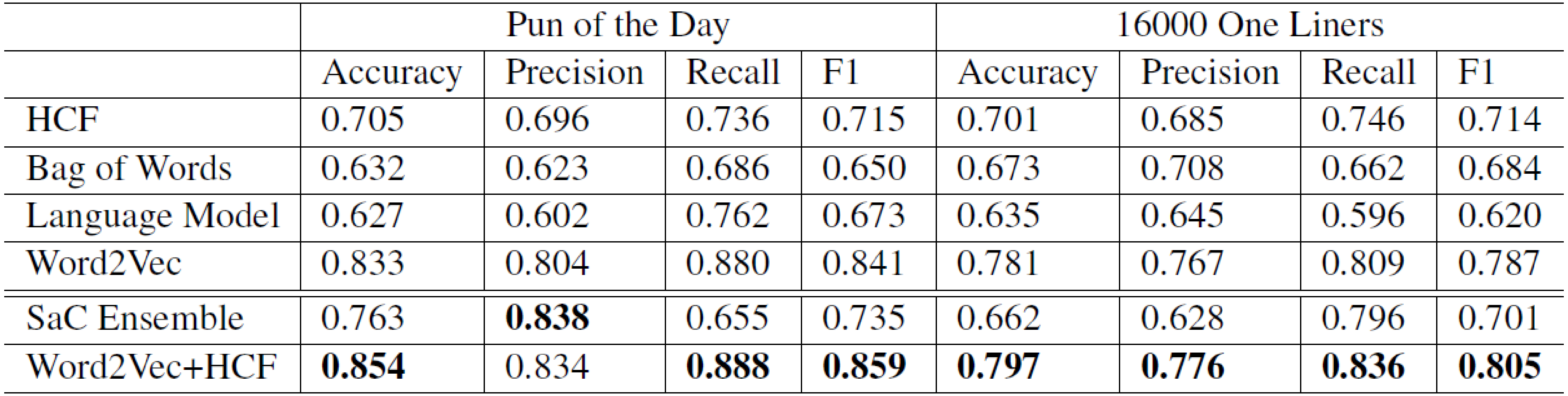
	One-liners	One-liners	One-liners
Heuristic	Reuters	BNC	Proverbs
Alliteration	74.31%	59.34%	53.30%
Antonymy	55.65%	51.40%	50.51%
Adult slang	52.74%	52.39%	50.74%
ALL	76.73%	60.63%	53.71%



	One-liners	One-liners	One-liners
Classifi er	Reuters	BNC	Proverbs
Na¨ı ve Bayes	96.67%	73.22%	84.81%
SVM	96.09%	77.51%	84.48%

Humor anchors [Yang et al., 2015]

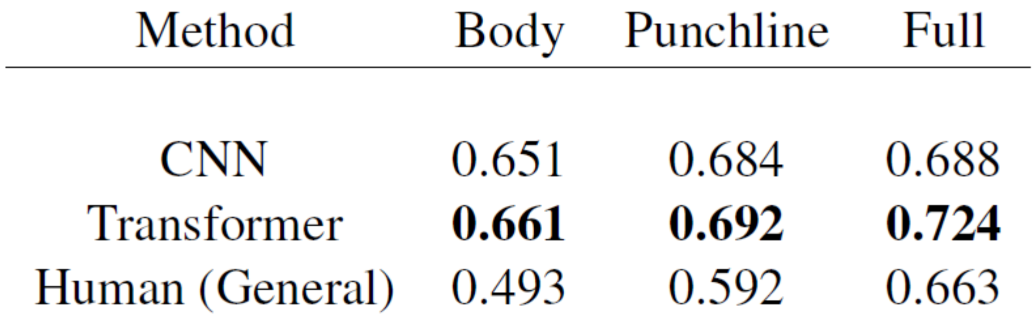
- Humor features:
  - incongruity,
  - ambiguity,
  - interpersonal effect (sentiment/subjectivity),
  - phonetic style.
- ‘Humor anchors’ – structures enabling humorous effect



	Pun of the Day				16000 One Liners			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
HCF	0.705	0.696	0.736	0.715	0.701	0.685	0.746	0.714
Bag of Words	0.632	0.623	0.686	0.650	0.673	0.708	0.662	0.684
Language Model	0.627	0.602	0.762	0.673	0.635	0.645	0.596	0.620
Word2Vec	0.833	0.804	0.880	0.841	0.781	0.767	0.809	0.787
SaC Ensemble	0.763	**0.838**	0.655	0.735	0.662	0.628	0.796	0.701
Word2Vec+HCF	**0.854**	0.834	**0.888**	**0.859**	**0.797**	**0.776**	**0.836**	**0.805**

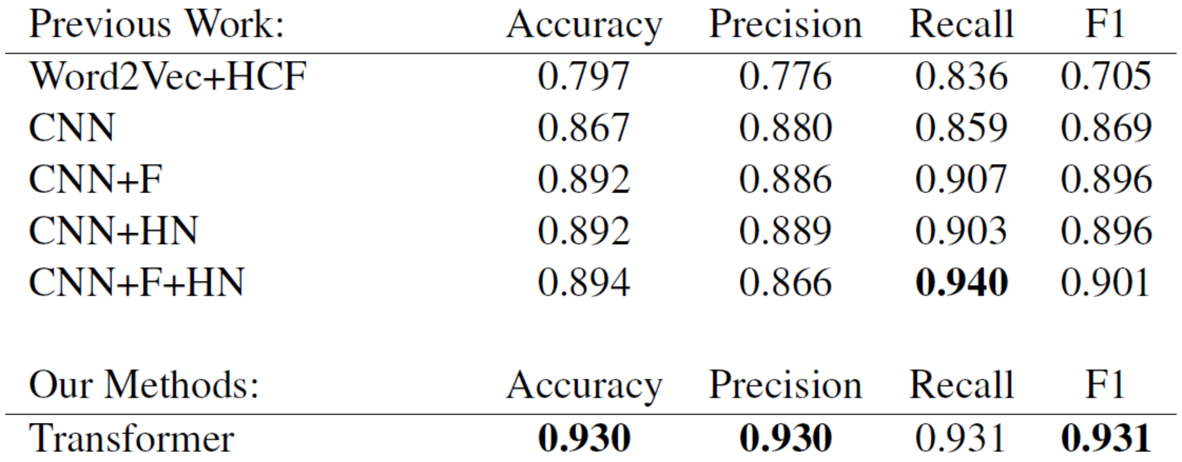
Transformer Gets the Last Laugh [Weller and Seppi, 2019]

- 14K jokes from reddit:



Method	Body	Punchline	Full
CNN	0.651	0.684	0.688
Transformer	**0.661**	**0.692**	**0.724**
Human (General)	0.493	0.592	0.663

- 16K one-liners:



Previous Work:	Accuracy	Precision	Recall	F1
Word2Vec+HCF	0.797	0.776	0.836	0.705
CNN	0.867	0.880	0.859	0.869
CNN+F	0.892	0.886	0.907	0.896
CNN+HN	0.892	0.889	0.903	0.896
CNN+F+HN	0.894	0.866	**0.940**	0.901

Our Methods:	Accuracy	Precision	Recall	F1
Transformer	**0.930**	**0.930**	0.931	**0.931**

Shared Tasks & Datasets

Detection and Interpretation of English Puns SemEval-2017 Task 7 [Miller et al., 2017]

Two types of puns in the data:

- homophonic
- homographic

Tasks:

- Pun detection: binary classification of contexts
- Pun location: locate the punning word
- Pun interpretation: indicate WordNet senses of the punning words

Ten participating systems

#HashtagWars
SemEval-2017 Task [Potash et al., 2017]

From classification to ranking
Data: tweets related to a Comedy TV show with funniness scores (0, 1, 2); 112 hashtags ‘topics’, 20-180 tweets per hashtag

Tasks:

- Pair-wise comparison of tweets
- Tweets ranking

Eight participating teams

Humor in Edited News Headlines SemEval-2020 Task 7 [Hossain, 2020]

Data: 15k presumably funny news headlines

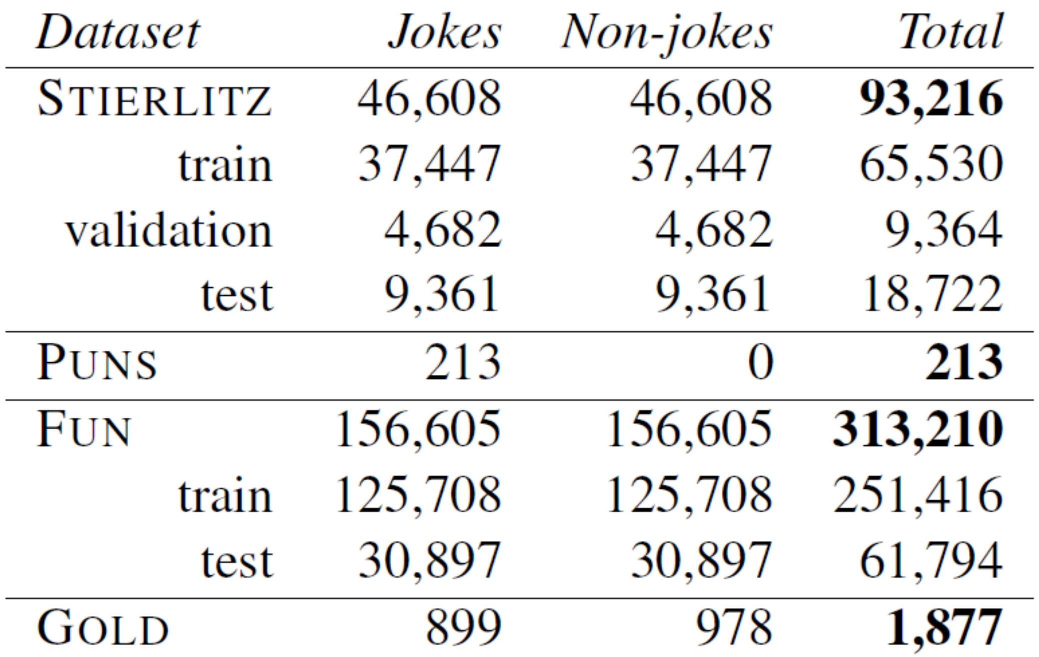
Tasks:

- Estimate the funniness of headlines (0..3)
- Pair-wise comparison of headlines

English Datasets

Dataset	Description	Reference
One-liners	16K one-liners / 16K headlines/proverbs/BNC	[Mihalcea & Strapparava, 2005]
Pun of the Day	2,400 puns/ 2,400 headlines	[Yang et al., 2015]
#HashTagWars	12K tweets for 112 hashtags, graded scores	[Potash et al., 2017]
English Puns	4K (71% puns) + WN annotations	[Miller et al., 2019]
Unfun.me	2.8K headline pairs (1.2K seeds), funny → serious edits	[West & Horvitz, 2019]
Humicroedit	15K headlines, serious → funny edits	[Hossain et al., 2019]
FunLines	8K headlines, serious → funny edits	[Hossain et al., 2020]
Reddit	13,884 not-funny / 2,025 funny jokes	[Weller & Seppi, 2019]

Russian Humor Datasets [Blinov et al., 2019]

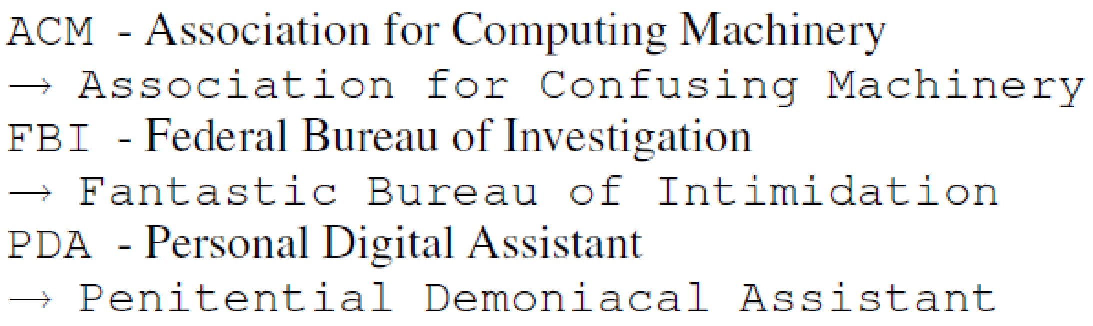


Dataset	*Jokes*	*Non-jokes*	*Total*
Stierlitz	46,608	46,608	**93,216**
train	37,447	37,447	65,530
validation	4,682	4,682	9,364
test	9,361	9,361	18,722
Puns	213	0	**213**
Fun	156,605	156,605	**313,210**
train	125,708	125,708	251,416
test	30,897	30,897	61,794
Gold	899	978	**1,877**

http://bit.ly/fun_data

Humor Generation

HAHAcronym

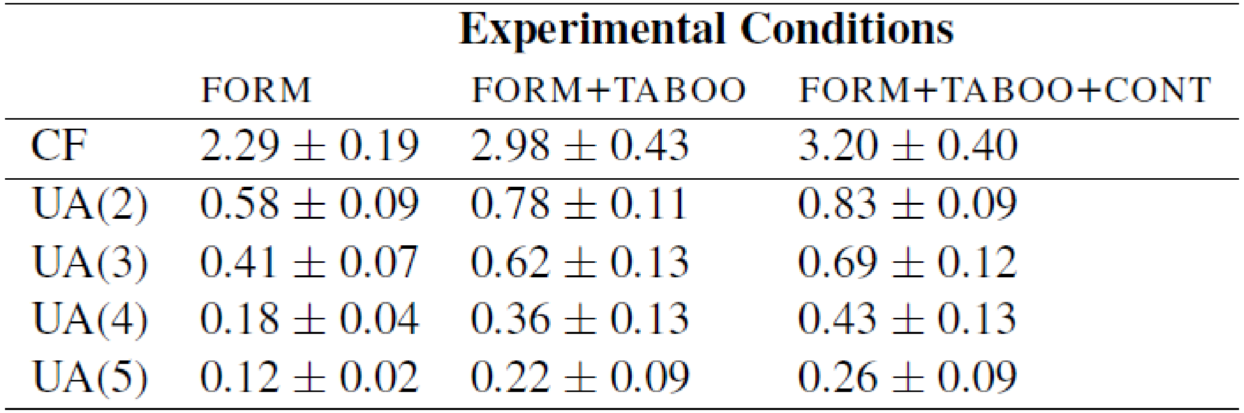


ACM - Association for Computing Machinery
→ Association for Confusing Machinery
FBI - Federal Bureau of Investigation
→ Fantastic Bureau of Intimidation
PDA - Personal Digital Assistant
→ Penitential Demoniacal Assistant

- Substitutions:
  - semantic field oppositions;
  - rhyme and rhythm;
  - for adjectives: antonym clustering and other semantic relations in WORDNET.
- Stock & Strapparava. HAHAcronym: A Computational Humor System, 2005.

Adult humor via lexical substitution

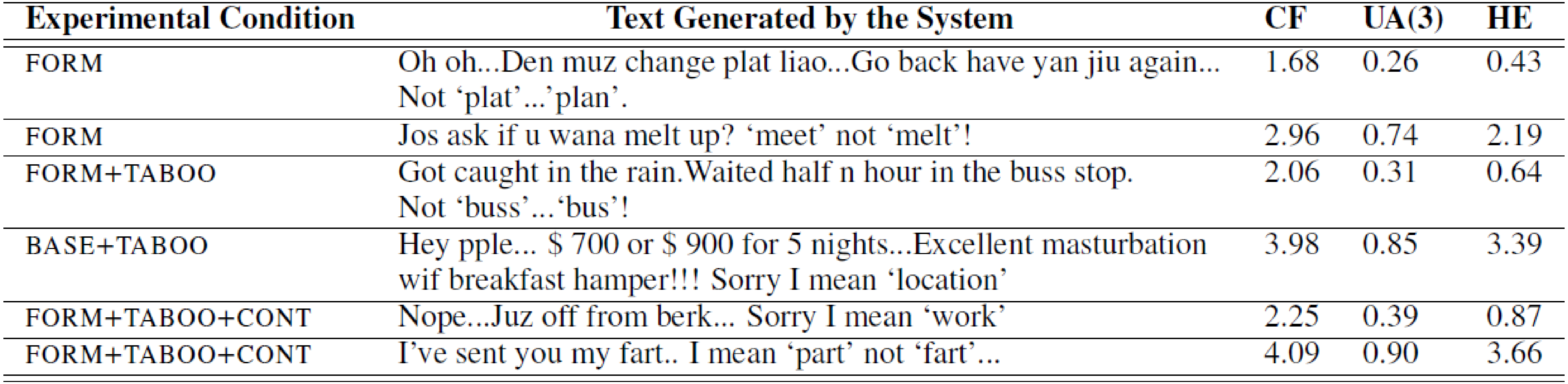
- Sms corpus
- Lexical constraints:
  - Similarly sounding/spelled words, the same POS
  - Taboo words (700)
  - End of the text, coherence (based on google book n-grams)



	Experimental Conditions		
	FORM	FORM+TABOO	FORM+TABOO+CONT
CF	2.29 ± 0.19	2.98 ± 0.43	3.20 ± 0.40
UA(2)	0.58 ± 0.09	0.78 ± 0.11	0.83 ± 0.09
UA(3)	0.41 ± 0.07	0.62 ± 0.13	0.69 ± 0.12
UA(4)	0.18 ± 0.04	0.36 ± 0.13	0.43 ± 0.13
UA(5)	0.12 ± 0.02	0.22 ± 0.09	0.26 ± 0.09

- Evaluation: crowdsourcing (100*3), 1..5 scale
- Valitutti et al. “Let Everything Turn Well in Your Wife”: Generation of Adult Humor Using Lexical Constraints, 2013.

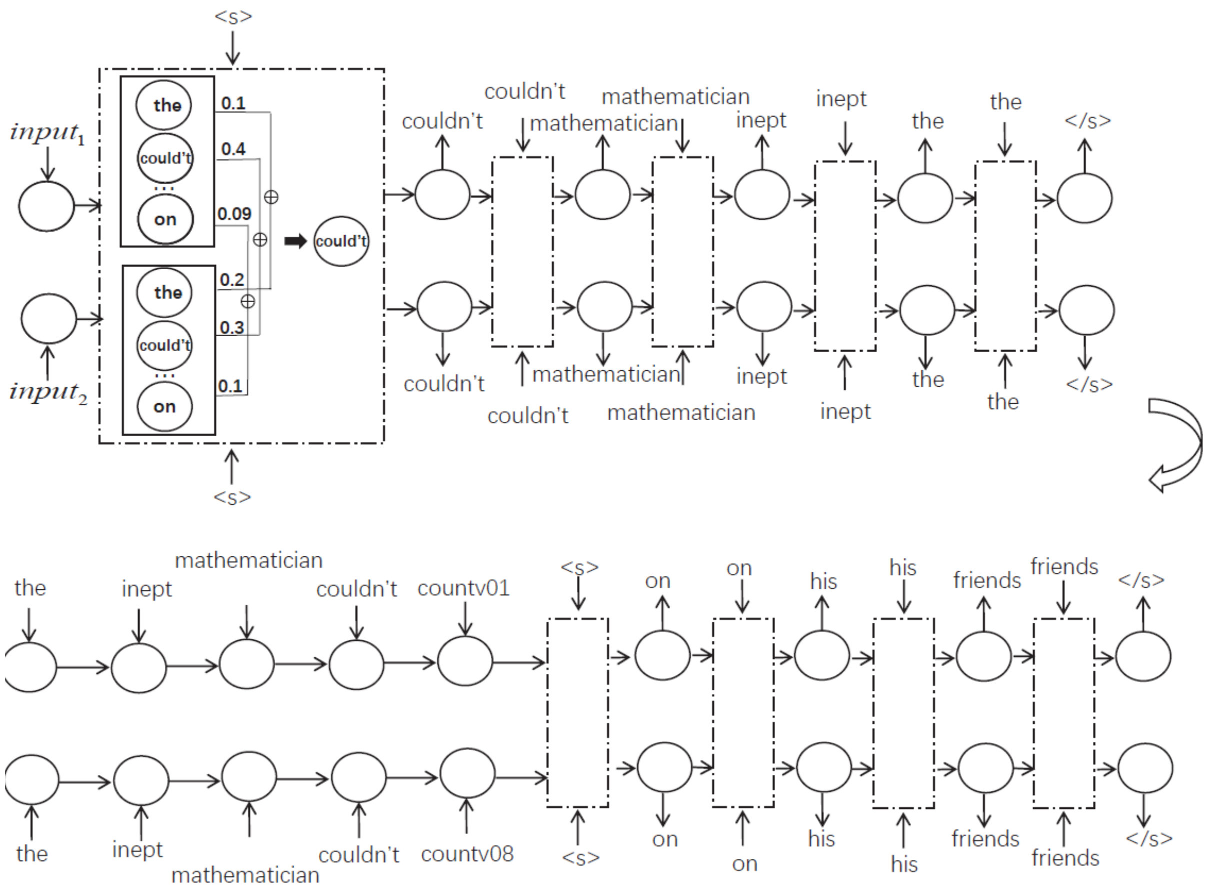
Examples



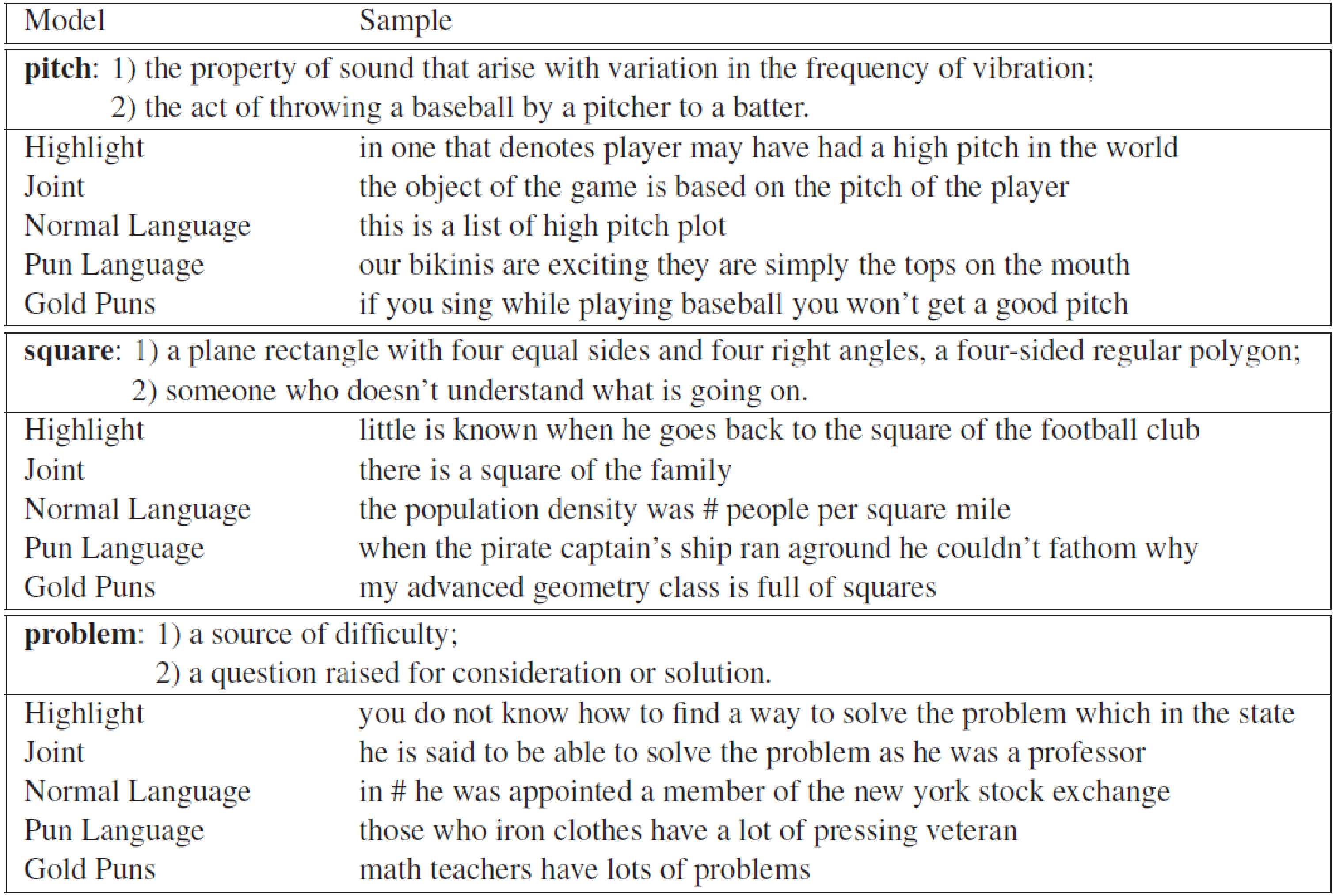
Experimental Condition	Text Generated by the System	CF	UA(3)	HE
FORM	Oh oh...Den muz change plat liao...Go back have yan jiu again... Not ‘plat’...’plan’.	1.68	0.26	0.43
FORM	Jos ask if u wana melt up? ‘meet’ not ‘melt’!	2.96	0.74	2.19
FORM+TABOO	Got caught in the rain.Waited half n hour in the buss stop. Not ‘buss’...‘bus’!	2.06	0.31	0.64
BASE+TABOO	Hey pple... $ 700 or $ 900 for 5 nights...Excellent masturbation wif breakfast hamper!!! Sorry I mean ‘location’	3.98	0.85	3.39
FORM+TABOO+CONT	Nope...Juz off from berk... Sorry I mean ‘work’	2.25	0.39	0.87
FORM+TABOO+CONT	I’ve sent you my fart.. I mean ‘part’ not ‘fart’...	4.09	0.90	3.66

Seq2pun [Yu et al., 2018]

- *Homographic* puns
- LSTM trained on a corpus with *explicitly disambiguated words*



Examples



Model	Sample
pitch: 1) the property of sound that arise with variation in the frequency of vibration; 2) the act of throwing a baseball by a pitcher to a batter.	
Highlight	in one that denotes player may have had a high pitch in the world
Joint	the object of the game is based on the pitch of the player
Normal Language	this is a list of high pitch plot
Pun Language	our bikinis are exciting they are simply the tops on the mouth
Gold Puns	if you sing while playing baseball you won't get a good pitch
square: 1) a plane rectangle with four equal sides and four right angles, a four-sided regular polygon; 2) someone who doesn't understand what is going on.	
Highlight	little is known when he goes back to the square of the football club
Joint	there is a square of the family
Normal Language	the population density was # people per square mile
Pun Language	when the pirate captain's ship ran aground he couldn't fathom why
Gold Puns	my advanced geometry class is full of squares
problem: 1) a source of difficulty; 2) a question raised for consideration or solution.	
Highlight	you do not know how to find a way to solve the problem which in the state
Joint	he is said to be able to solve the problem as he was a professor
Normal Language	in # he was appointed a member of the new york stock exchange
Pun Language	those who iron clothes have a lot of pressing veteran
Gold Puns	math teachers have lots of problems

Pun generation with surprise [He & Liang, 2019]

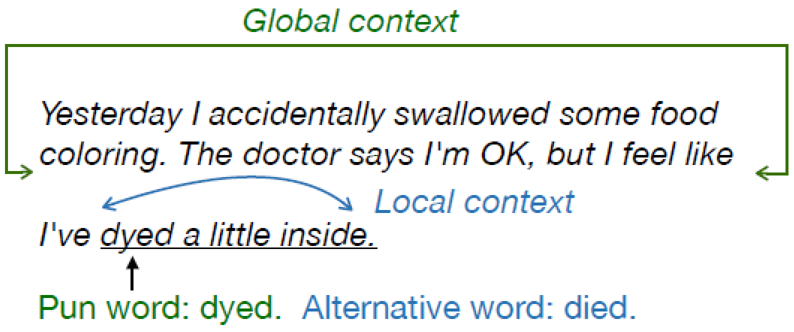
- Puns based on *homophones*
- Local vs global context

$$S(c) \overset{\text{def}}{=} -\log \frac{p(w^{\mathrm{p}} \mid c)}{p(w^{\mathrm{a}} \mid c)} = -\log \frac{p(w^{\mathrm{p}}, c)}{p(w^{\mathrm{a}}, c)}. \quad (1)$$

$$S_{\text{local}} \overset{\text{def}}{=} S(x_{p-d:p-1}, x_{p+1:p+d}), \quad (2)$$

$$S_{\text{global}} \overset{\text{def}}{=} S(x_{1:p-1}, x_{p+1:n}), \quad (3)$$

$$S_{\text{ratio}} \overset{\text{def}}{=} \begin{cases} -1 & S_{\text{local}} < 0 \text{ or } S_{\text{global}} < 0, \\ S_{\text{local}}/S_{\text{global}} & \text{otherwise.} \end{cases} \quad (4)$$



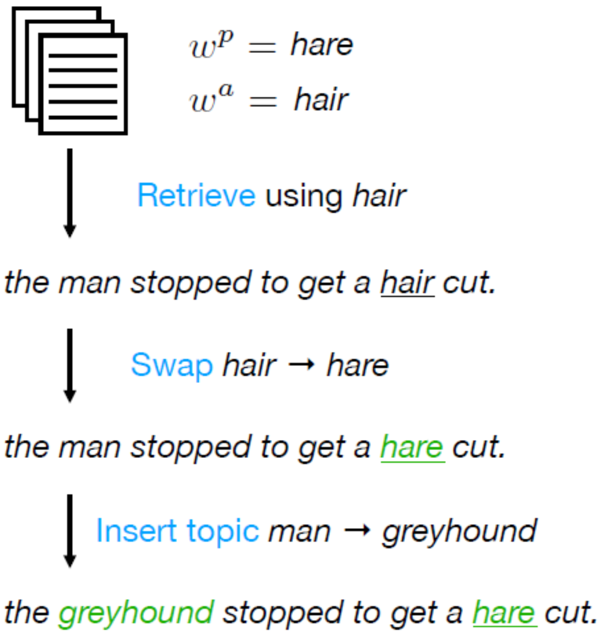
Global context

Yesterday I accidentally swallowed some food coloring. The doctor says I'm OK, but I feel like

Local context

I've dyed a little inside.

Pun word: dyed. Alternative word: died.



w^p = *hare*

w^a = *hair*

Retrieve using *hair*

the man stopped to get a hair cut.

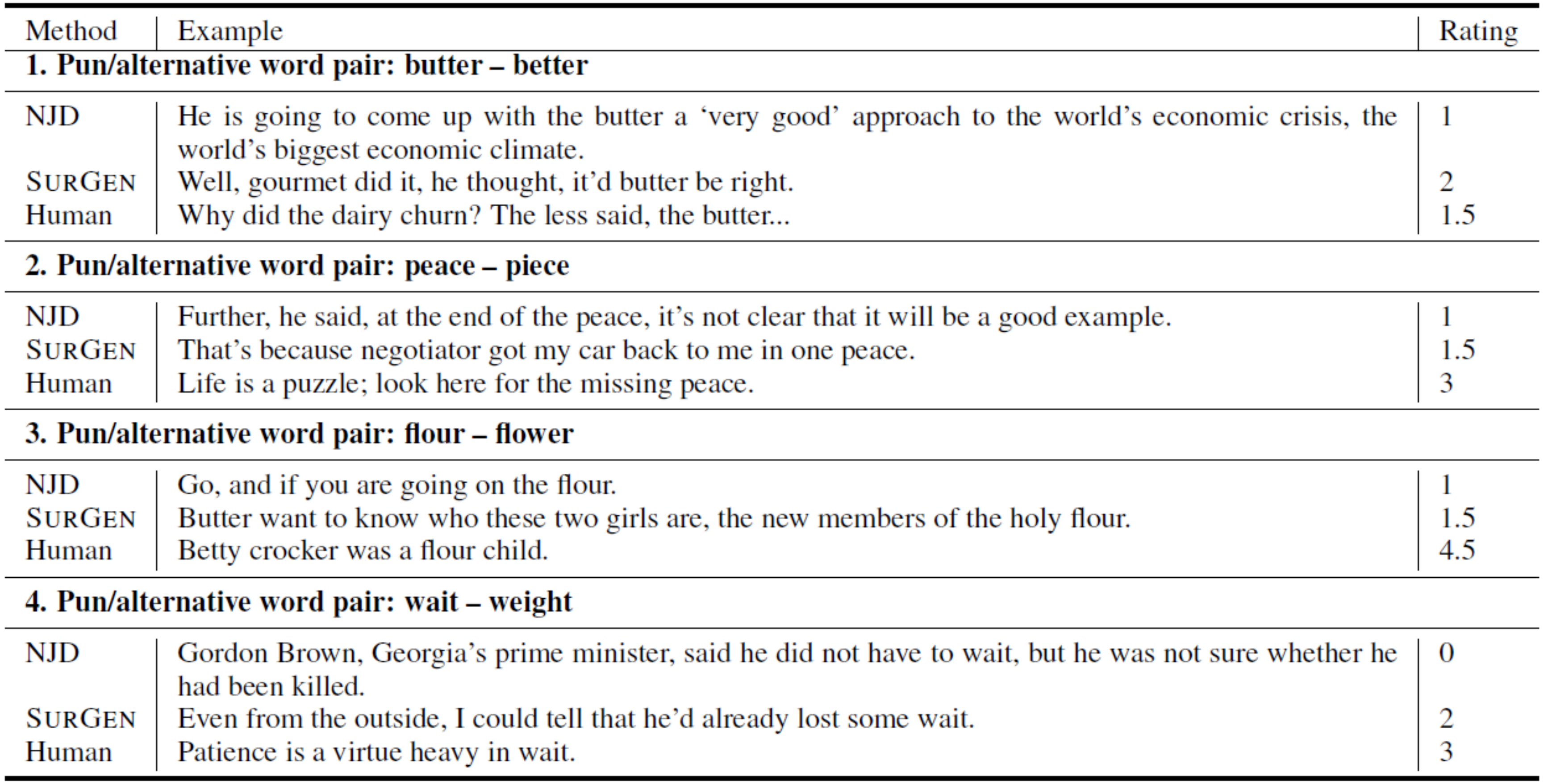
Swap *hair* → *hare*

the man stopped to get a hare cut.

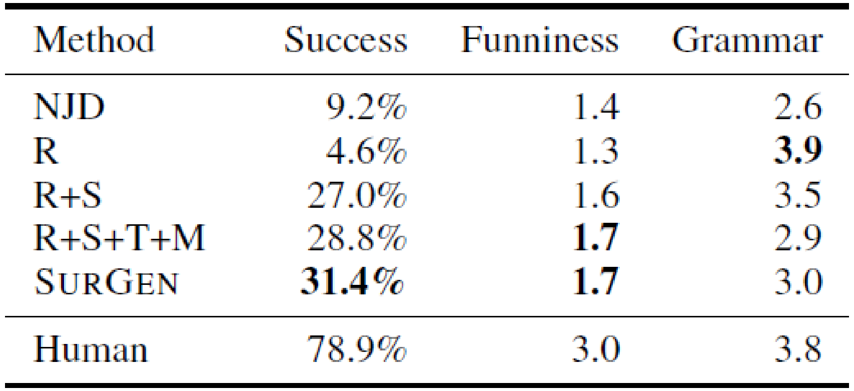
Insert topic *man* → *greyhound*

the greyhound stopped to get a hare cut.

Results

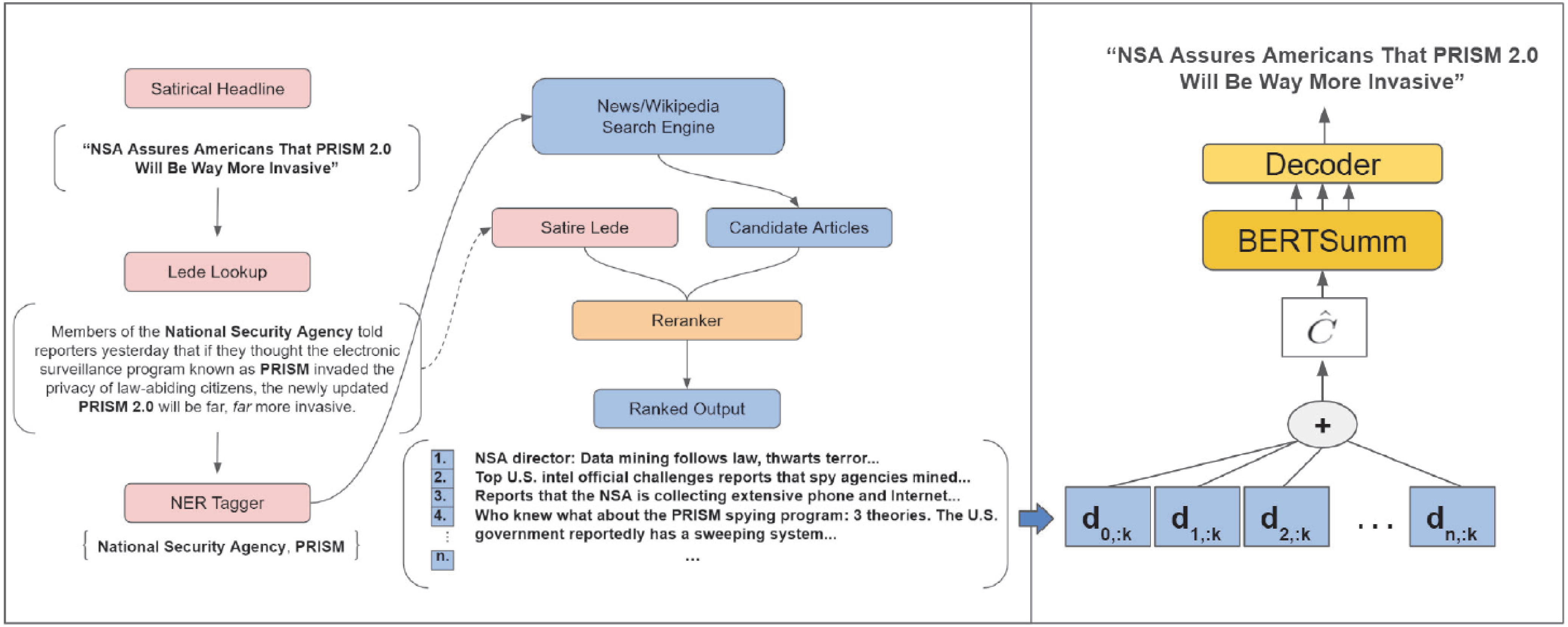


Method	Example	Rating
1. Pun/alternative word pair: butter – better		
NJD	He is going to come up with the butter a ‘very good’ approach to the world’s economic crisis, the world’s biggest economic climate.	1
SurGen	Well, gourmet did it, he thought, it’d butter be right.	2
Human	Why did the dairy churn? The less said, the butter...	1.5
2. Pun/alternative word pair: peace – piece		
NJD	Further, he said, at the end of the peace, it’s not clear that it will be a good example.	1
SurGen	That’s because negotiator got my car back to me in one peace.	1.5
Human	Life is a puzzle; look here for the missing peace.	3
3. Pun/alternative word pair: flour – flower		
NJD	Go, and if you are going on the flour.	1
SurGen	Butter want to know who these two girls are, the new members of the holy flour.	1.5
Human	Betty crocker was a flour child.	4.5
4. Pun/alternative word pair: wait – weight		
NJD	Gordon Brown, Georgia’s prime minister, said he did not have to wait, but he was not sure whether he had been killed.	0
SurGen	Even from the outside, I could tell that he’d already lost some wait.	2
Human	Patience is a virtue heavy in wait.	3



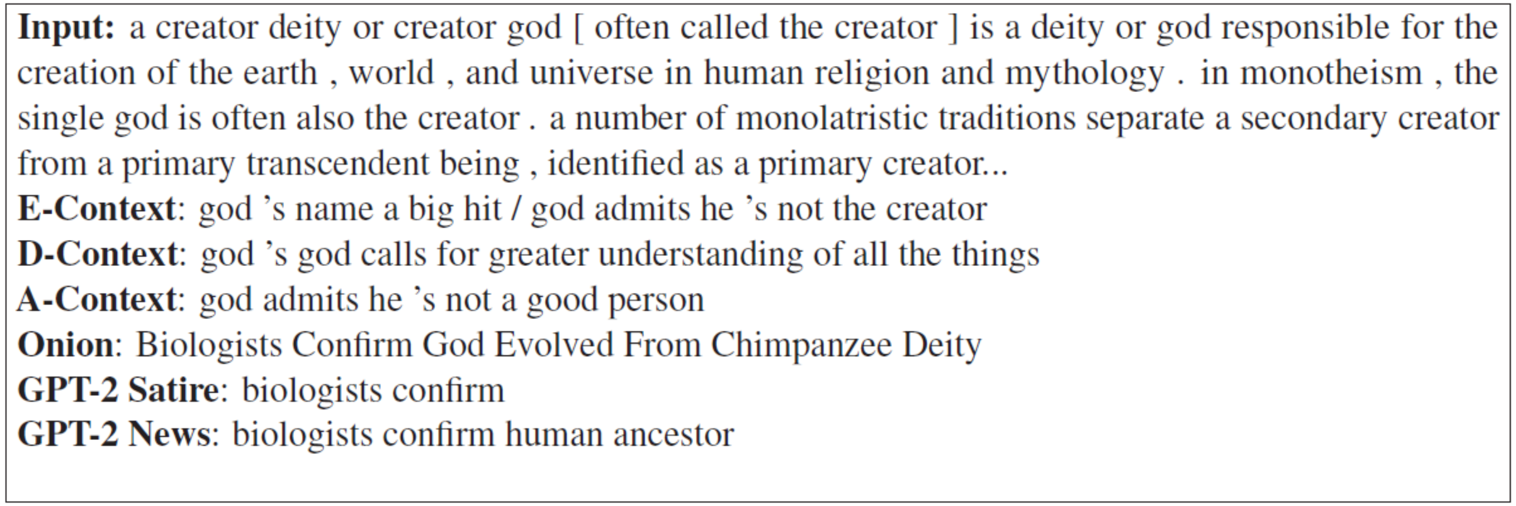
Method	Success	Funniness	Grammar
NJD	9.2%	1.4	2.6
R	4.6%	1.3	**3.9**
R+S	27.0%	1.6	3.5
R+S+T+M	28.8%	**1.7**	2.9
SurGen	**31.4%**	**1.7**	3.0
Human	78.9%	3.0	3.8

Funny headline generation


Satirical Headline
"NSA Assures Americans That PRISM 2.0 Will Be Way More Invasive"
Lede Lookup
Members of the **National Security Agency** told reporters yesterday that if they thought the electronic surveillance program known as **PRISM** invaded the privacy of law-abiding citizens, the newly updated **PRISM 2.0** will be far, *far* more invasive.
NER Tagger
National Security Agency, PRISM
News/Wikipedia Search Engine
Satire Lede
Candidate Articles
Reranker
Ranked Output
1. NSA director: Data mining follows law, thwarts terror...
2. Top U.S. intel official challenges reports that spy agencies mined...
3. Reports that the NSA is collecting extensive phone and Internet...
4. Who knew what about the PRISM spying program: 3 theories. The U.S. government reportedly has a sweeping system...
n. ...
"NSA Assures Americans That PRISM 2.0 Will Be Way More Invasive"
Decoder
BERTSumm
$\hat{C}$
+
$d_{0,:k}$
$d_{1,:k}$
$d_{2,:k}$
...
$d_{n,:k}$

[Horvitz et al., 2020]

Examples



Input: a creator deity or creator god [often called the creator] is a deity or god responsible for the creation of the earth , world , and universe in human religion and mythology . in monotheism , the single god is often also the creator . a number of monolatristic traditions separate a secondary creator from a primary transcendent being , identified as a primary creator...

E-Context: god 's name a big hit / god admits he 's not the creator

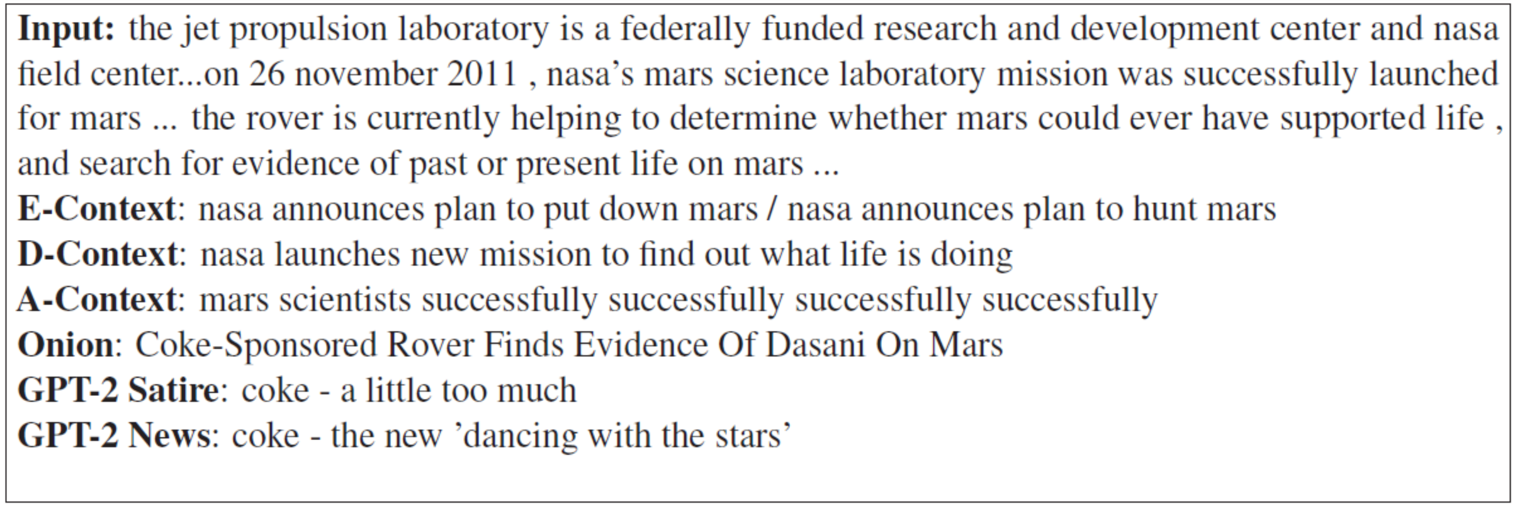
D-Context: god 's god calls for greater understanding of all the things

A-Context: god admits he 's not a good person

Onion: Biologists Confirm God Evolved From Chimpanzee Deity

GPT-2 Satire: biologists confirm

GPT-2 News: biologists confirm human ancestor



Input: the jet propulsion laboratory is a federally funded research and development center and nasa field center...on 26 november 2011 , nasa's mars science laboratory mission was successfully launched for mars ... the rover is currently helping to determine whether mars could ever have supported life , and search for evidence of past or present life on mars ...

E-Context: nasa announces plan to put down mars / nasa announces plan to hunt mars

D-Context: nasa launches new mission to find out what life is doing

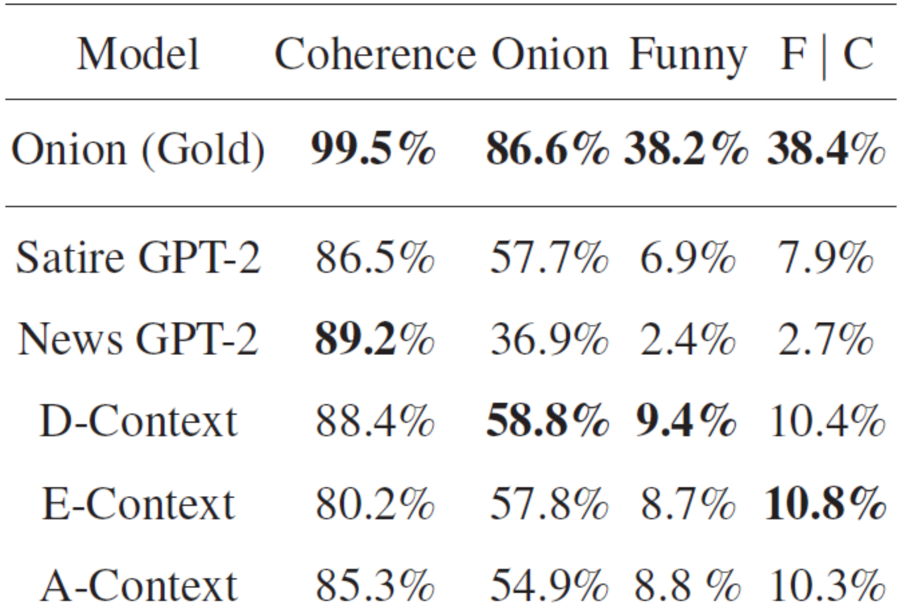
A-Context: mars scientists successfully successfully successfully successfully

Onion: Coke-Sponsored Rover Finds Evidence Of Dasani On Mars

GPT-2 Satire: coke - a little too much

GPT-2 News: coke - the new 'dancing with the stars'

Evaluation



Model	Coherence	Onion	Funny	F | C
Onion (Gold)	**99.5%**	**86.6%**	**38.2%**	**38.4%**
Satire GPT-2	86.5%	57.7%	6.9%	7.9%
News GPT-2	**89.2%**	36.9%	2.4%	2.7%
D-Context	88.4%	**58.8%**	**9.4%**	10.4%
E-Context	80.2%	57.8%	8.7%	**10.8%**
A-Context	85.3%	54.9%	8.8 %	10.3%

Humorous Response Generation using IR [Blinov et al., 2017]

Data: 103 “funny Twitter accounts”; 300K tweets in total

Three Models (BM25, Query-Term Reweighting, doc2vec)

Evaluation: community question answering (CQA), lab settings.

Responses & Evaluation Scores

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. :(

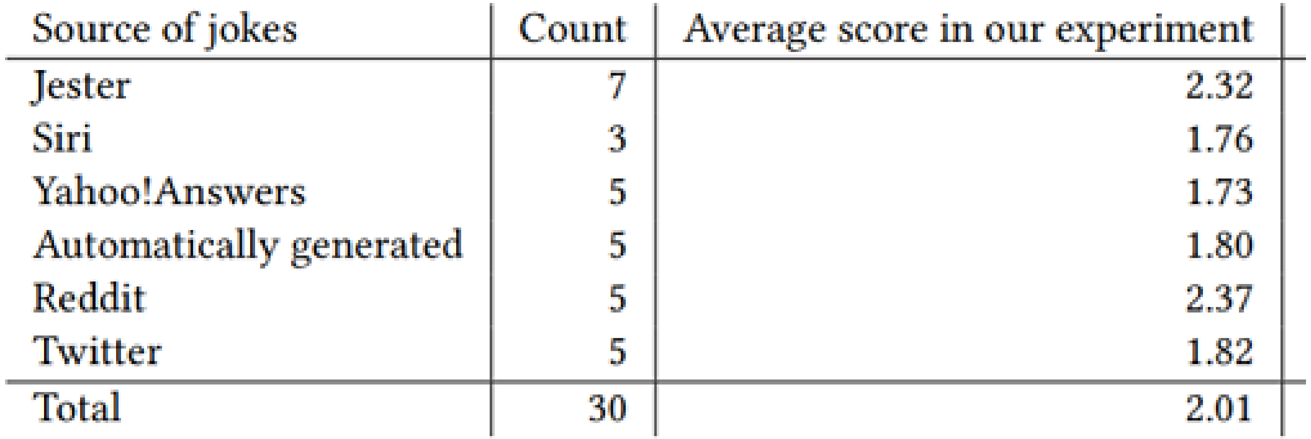
Humor Evaluation

Why Humor Evaluation?

- Massive generation ‘in the wild’ vs. few handcrafted rules.
- Evaluation is crucial for measuring progress.
- Can we get rid of subjectivity of crowdworkers?

Humor Evaluation [Braslavski et al., 2018]

- 30 dialog jokes from different sources

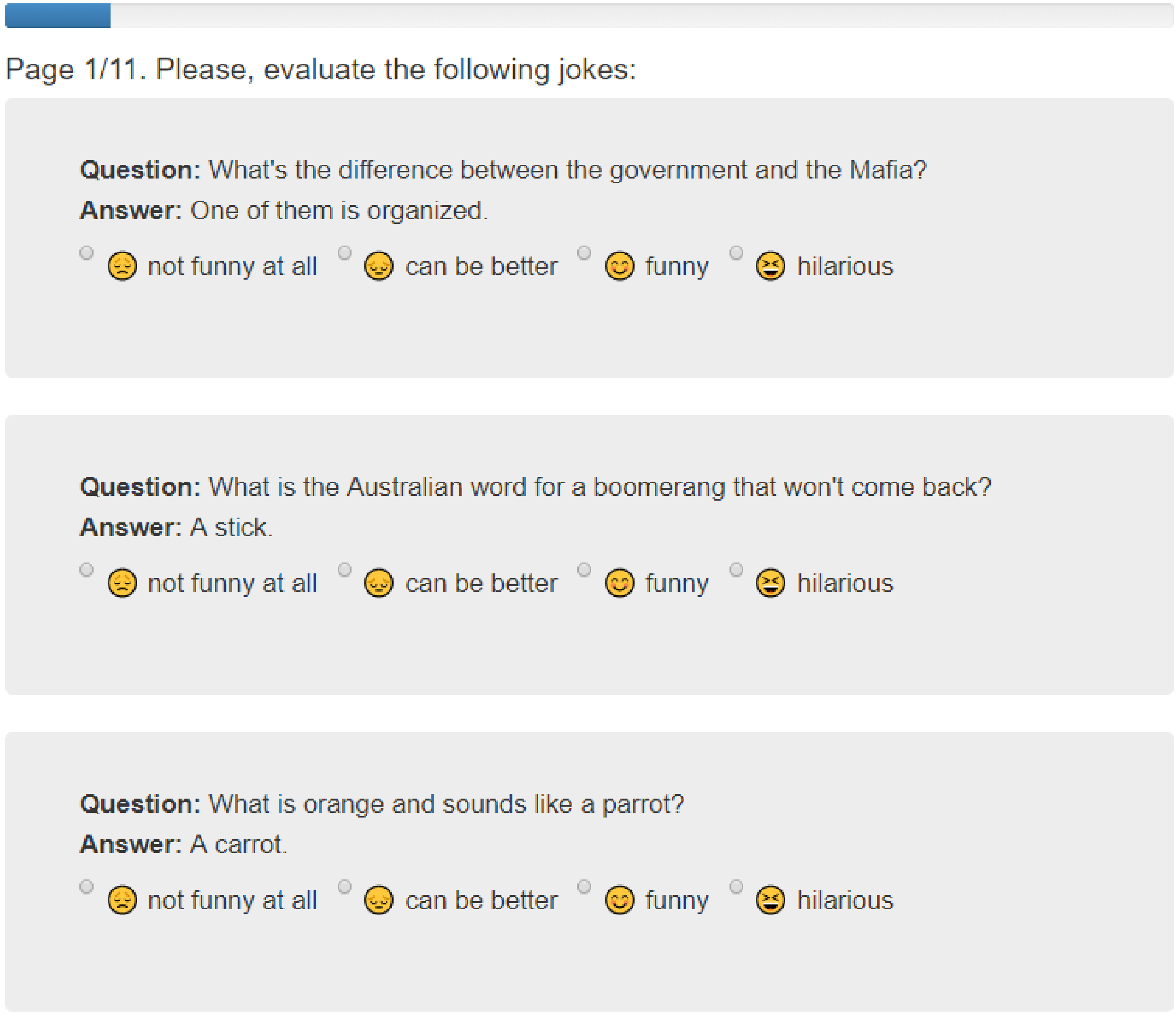


Source of jokes	Count	Average score in our experiment
Jester	7	2.32
Siri	3	1.76
Yahoo!Answers	5	1.73
Automatically generated	5	1.80
Reddit	5	2.37
Twitter	5	1.82
Total	30	2.01

Q: *Am I the coolest person in the world?*
A: *Nope. That person lives in Antarctica.*

Q: *How did the hipster burn his mouth?*
A: *He ate a cookie BEFORE they were cool!*

Evaluation Interface



Page 1/11. Please, evaluate the following jokes:

Question: What's the difference between the government and the Mafia?
Answer: One of them is organized.

○ 😣 not funny at all ○ 😔 can be better ○ 😊 funny ○ 😆 hilarious

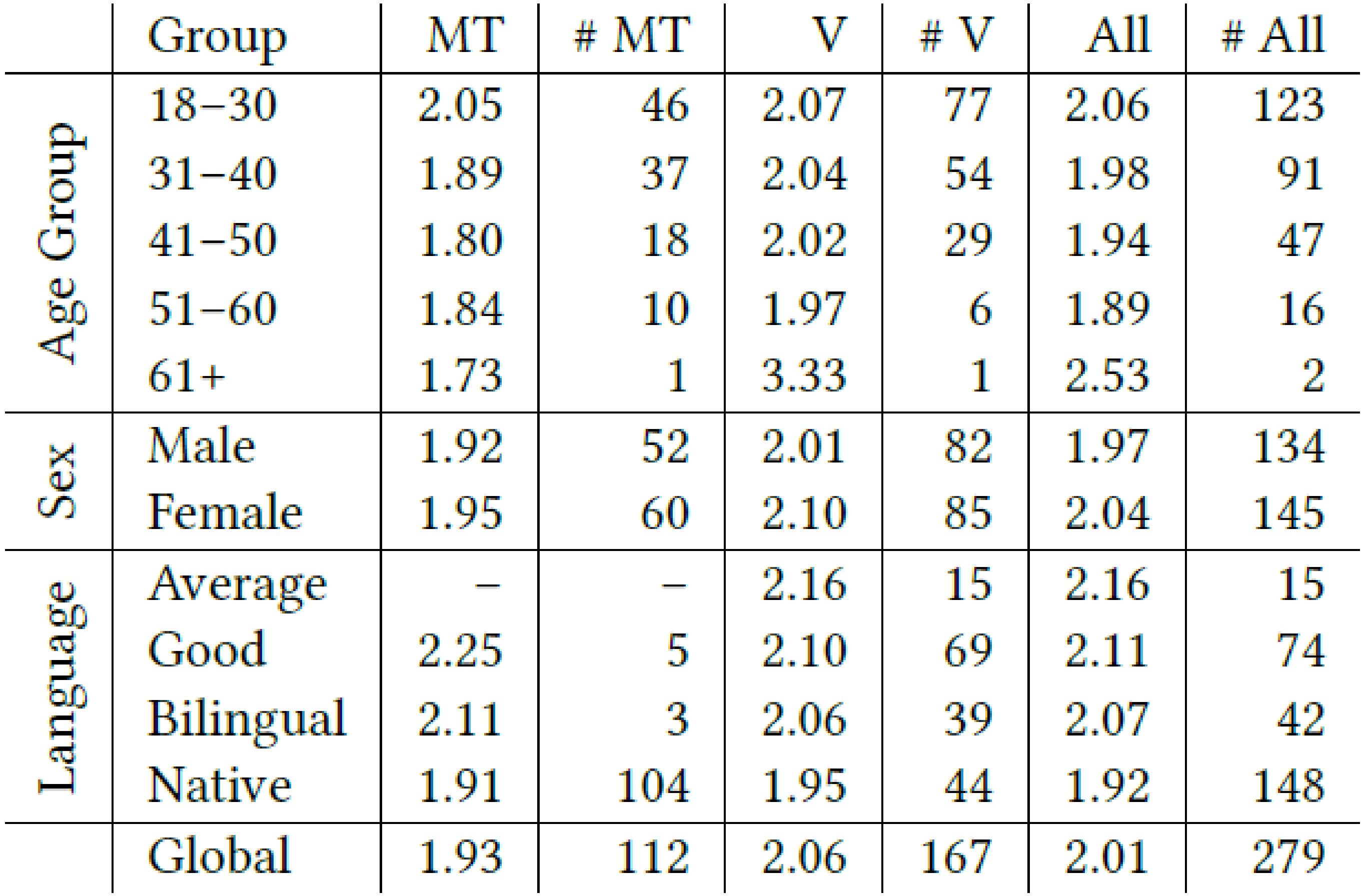
Question: What is the Australian word for a boomerang that won't come back?
Answer: A stick.

○ 😣 not funny at all ○ 😔 can be better ○ 😊 funny ○ 😆 hilarious

Question: What is orange and sounds like a parrot?
Answer: A carrot.

○ 😣 not funny at all ○ 😔 can be better ○ 😊 funny ○ 😆 hilarious

Evaluation Results



	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

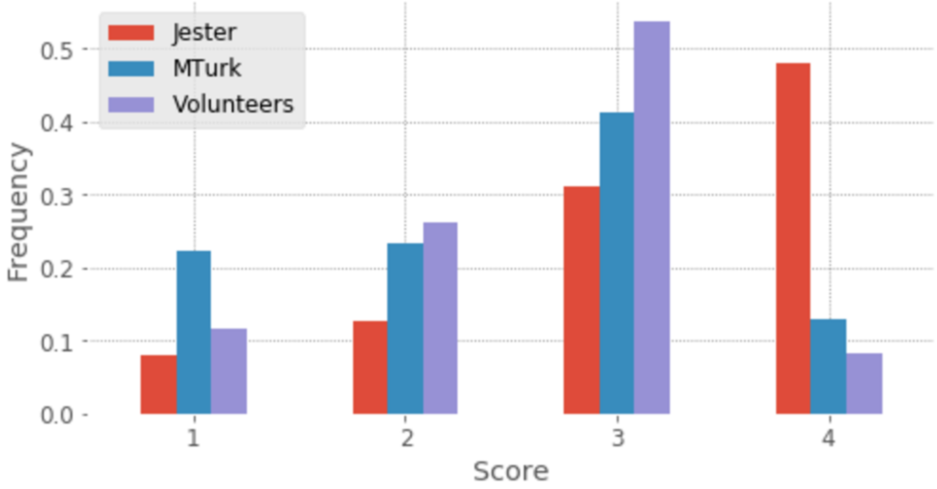
Evaluation Results -2

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.*



Q: *How many programmers does it take to change a lightbulb?*
A: *NONE! That's a hardware problem*

Evaluation: Summary

- Crowdsourcing is suitable for humor evaluation.
- Age/language proficiency/gender influence joke ratings.
- Jokes degrade over time → re-using evaluation is questionable.

Questions?

pbras@yandex.ru