How to evaluate humorous response generation, seriously?

Pavel Braslavski
Ural Federal University
pavel.braslavsky@urfu.ru

Vladislav Blinov
Ural Federal University
vladislav.blinov@urfu.ru

Valeria Bolotova
Ural Federal University
lurunchik@gmail.com

Katya Pertsova
University of North Carolina
pertsova@email.unc.edu

ABSTRACT

Nowadays natural language user interfaces, such as chatbots and conversational agents, are very common. A desirable trait of such applications is a sense of humor. It is, therefore, important to be able to measure quality of humorous responses. However, humor evaluation is hard since humor is highly subjective. To address this problem, we conducted an online evaluation of 30 dialog jokes from different sources by almost 300 participants – volunteers and Mechanical Turk workers. We collected joke ratings along with participants' age, gender, and language proficiency. Results show that demographics and joke topics can partly explain variation in humor judgments. We expect that these insights will aid humor evaluation and interpretation. The findings can also be of interest for humor generation methods in conversational systems.

KEYWORDS

computational humor; conversational systems; evaluation; crowdsourcing

ACM Reference Format:
Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, and Katya Pertsova. 2018. How to evaluate humorous response generation, seriously?. In *Proceedings of 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3176349.3176879

1 INTRODUCTION

There is a rapid proliferation of natural language interfaces – chatbots, intelligent assistants, conversational agents. These interfaces are used for a variety of tasks – from ordering pizza to serving as an intelligent companion or even a confidant of the user. The technology behind such applications as Apple's Siri or Microsoft's Cortana is changing the style of human interaction with mobile devices, cars, consumer electronics, smart homes, etc. This interaction becomes more emotional, personal, and even intimate. Sense of humor is an important quality of a conversational agent: humor makes a human-computer conversation more engaging, helps to establish a more trusting relationship between the user and the agent, creates a feeling of agent's empathy and personality. The need for humor is evident from the users' requests to personal assistants: "tell me a joke" is a very frequent request [10]. A humorous response is a good option for out-of-domain requests [4], it can soften the negative impact of inadequacies in the system's performance [5] and is a good option if the system is not able to generate an appropriate response.

The sense of humor of modern mobile personal assistants is seemingly based on a relatively small number of hand-crafted stimulus–response pairs, which leads to repetitions. Ideally, a personal assistant should be able to produce fresh and funny responses for a wide variety of input utterances. Thus, it becomes crucial to evaluate what users find funny. Automatic evaluation of dialog systems that relies on proximity metrics to reference utterances received a great deal of attention thanks to end-to-end training of conversational systems. As a recent study showed [12], outcomes of such evaluation do not correlate with human judgments well. Automatic methods for evaluation of humorous responses should be even worse due to diversity of potentially funny responses in a given context. Subsequently, a controlled lab evaluation remains the main viable option so far. Humor evaluation is challenging since the perception of humor is highly subjective and is conditioned on the situation and the socio-cultural background. Our preliminary small-scale evaluation experiments have shown that assessors' agreement is very low [6].

In this follow-up study, we attempt to investigate the problem of humor evaluation on a larger scale using crowdsourcing. The main idea of the study is to analyze personal characteristics of the assessors and estimate the contribution of these characteristics to variations in judgments. We collected a set of 30 two-turn jokes from different sources. Almost 300 people – volunteers and paid Mechanical Turk workers – took part in the online evaluation. Along with joke ratings, we collected assessors' age, gender, language proficiency, self-assessed sense of humor, and Big5 personality traits[1].

We conclude that crowdsourcing platforms are an efficient and appropriate option for humor evaluation. Annotators' demographics such as age and language proficiency can partly explain their disagreement. Although gender does not affect average humor scores, certain topics are appreciated by men and women differently. Comparison of obtained ratings with those based on evidence from joke sources (e.g. likes and retweets in case of funny tweets) suggests that we cannot straightforwardly reuse the same humor scores in different settings. We expect that these new insights will help to

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176879

[1] Due to limited space we do not report the Big5-related analysis of results in the paper.

design better humor evaluation methods and interpret their outcomes. The findings can also be used to improve humor generation strategies in conversational systems.

2 RELATED WORK

Our experimental design is similar to psychological humor appreciation studies. For example, a widely adopted 3 WD (*Witzdimensionen*) test assesses an individual's perception of jokes and cartoons along three basic humor types: incongruity-resolution, nonsense, and sexual humor [18]. While psychological tests aim at revealing individual humor-related traits based on a fixed set of stimuli, our goal is to reliably evaluate humorous content from an average user's point of view.

Jester dataset [7, 16] contains over 5 million ratings of 100 jokes from 150K users on a continuous [−10, +10] scale. Jokes were collected 'from friends and newsgroups' with filtering out 'highly offensive jokes'. However, the main focus of Jester was on the joke recommendation algorithm, rather than on specific issues of humor evaluation. Both users and jokes were treated as a 'black box'.

In computational humor studies, humorous vs. serious content is usually considered as such based on its source: for example, humor collections vs. news [15, 23]. Another approach relies on expert opinion. For example, humor scores of tweets related to a TV show in the *#HashtagWars* collection reflect the choice of the show's editorial staff [17].

There are two main approaches to evaluating computer–generated humor. In one of them, users evaluate their overall experience of using the system and the system's sense of humor in particular [11, 15, 22]. The second approach is to evaluate individual automatically-generated jokes on a rating scale [9, 20, 21]. Many of these studies rely on crowdsourcing [19, 21, 22], but unfortunately they do not report crucial details about the evaluation process. Humor evaluation experiments mentioned above also do not take into account the demographics of assessors and their personal traits.

3 EXPERIMENTAL DESIGN

The online questionnaire consisted of an introduction followed by three sections: 1) demographics; 2) Big5 test plus two sense of humor self-assessment items; and 3) 10 screens with three dialog jokes on each to evaluate, 30 jokes in total.

We collected demographic information about age range, gender, and English proficiency. To obtain Big5 factor scores we used the Ten Item Personality Measure (TIPI) [8]. Additionally, participants were asked to assess two items related to their humor appreciation and productivity on a 7-point scale:

(1) *I laugh often; it is easy to make me laugh.*
(2) *I usually pun, tell jokes or funny anecdotes in a social situation.*

We collected 30 two-turn dialog jokes from several sources: Jester dataset, Siri, *Jokes & Riddles* category of Yahoo!Answers[2], Reddit's *funny* subreddit[3], tweets from funny Twitter accounts[4], as well as automatically retrieved answers to Yahoo!Answers questions from our previous experiments [6]. The breakdown of the collection by joke source is shown in Table 1.[5]. In the case of Siri we went through image results for a query `[siri jokes]` and picked up funny question/answer pairs that have no references to Siri herself. When composing the collection we tried to maintain topical variety. In addition, we tried to pick jokes of varying levels of funniness based on ratings presented in the original sources – points on Reddit, thumbs up & down on Yahoo! Answers, likes and retweets on Twitter, and user ratings in Jester data. Note that the Jester dataset contains 'canned jokes' only; Siri jokes can also be assigned to the same category due to the popularity of the application and its limited joke inventory. Other jokes are closer to spontaneous conversational jokes: the odds are high that the assessors have not seen them before. See [1] for distinction between canned and conversational jokes.

Table 1: Joke sources, average scores, and rank correlations with rankings in source data

Source	Count	Avg. score	Spearman's ρ
Jester	7	2.32	0.57
Siri	3	1.76	–
Yahoo!Answers	5	1.73	0.05
Automatically generated	5	1.80	0.97
Reddit	5	2.37	0.80
Twitter	5	1.82	-0.30
Total	30	2.01	

The jokes were presented to participants on 10 screens, three jokes at a time. The order of the screens and jokes were the same for all participants. The jokes were judged on a four-point scale, with corresponding emoticons in the evaluation interface: not funny at all (1, ☹), can be better (2, 😐), funny (3, 🙂), and hilarious (4, 😂).

Volunteers were recruited via online social networks. Paid workers were recruited through the Mechanical Turk (MT) crowdsourcing platform. MT workers were required to have a US locale and were offered 0.3$ per HIT.

4 RESULTS AND DISCUSSION

4.1 General Trends

Overall, 167 volunteers and 112 MT workers took part in the experiment.[6] Table 2 reports the breakdown of total 279 participants by age, gender, and English language proficiency (see #-columns).

As can be seen from the Table, the population has a skew towards younger people (under 30); however, the next two age groups (31–40 and 41-50) are fairly well presented in the sample. The population is gender-balanced, although we undertook no special efforts to ensure it. Almost all MT workers reported being English native speakers, whereas volunteers' self-reported proficiency in English is more diffuse.

Self-assessed humor appreciation/productivity levels do not correlate with joke scores.[7] It means that self-reported sense of humor cannot be used for unbiasing the joke scores.

[2]https://answers.yahoo.com/dir/index?sid=396546041
[3]https://www.reddit.com/r/funny/
[4]These accounts were obtained through various lists, such as http://www.hongkiat.com/blog/funny-twitter-accounts/
[5]The complete list is available at http://bit.ly/HumorEval
[6]We rejected MT workers who spent less than five seconds per page or used only two out of four levels on the evaluation scale.
[7]However, the correlation is high in a group of 15 participants who assessed *both* their humor appreciation and productivity as very high.

Table 2: Average joke scores and number of participants by group (MT – MT workers, V – volunteers)

	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

Table 2 summarizes average joke scores for each group of participants. According to an independent-sample t-test, the difference between average scores of volunteers and MTurk workers for 30 jokes is not statistically significant.

Mean assessments by men and women are slightly but insignificantly distinct. This result supports numerous psychological studies that find no gender-specific differences in sense of humor. There are still gender-specific topical preferences: studies suggest that men are more likely to enjoy aggressive and sexual humor, whereas women like 'nonsense' humor more (see references and discussion in [13]). Our data support these hypotheses, see section 4.2.

Average joke scores decrease whith age. This is in line with some psychological studies suggesting that decline in cognitive abilities in the elderly may be associated with lower comprehension of humor (but greater humor appreciation at the same time). Some studies suggest that *humor type preferences* change with age (see details in [13]).

We can also observe a monotone decrease of average scores with increase in language proficiency, which is surprising at first glance. According to [3], most frequent cases of failed humor appreciation by non-native speakers are connected with the failure to understand the meaning of words and their connotations. Grasping word ambiguities is crucial in our experiment since many jokes in our collection involve some kind of wordplay. We hypothesize that participants had no time pressure to comprehend the jokes in contrast to real-life conversations (volunteers spent more time assessing the jokes, see below). Non-native speakers may also be less aware of 'canned' jokes.

Figure 1 shows average time spent on assessment along with average joke scores by MT workers vs. volunteers on each out of 10 screens of joke triples. The score bars demonstrate again that volunteers and MT workers assess jokes in a very similar way. The time chart shows that MT workers complete tasks faster; there is less variation in time intervals. This observation can be explained again by MT workers' language proficiency and their more professional attitude. The discrepancy of time spent on the first page between volunteers and MT workers can be explained by the fact that this is the shortest page out of 10 (three jokes make up 230 characters), the jokes on this page are quite simple (i.e. no tricky wordplay), and volunteers are very enthusiastic at

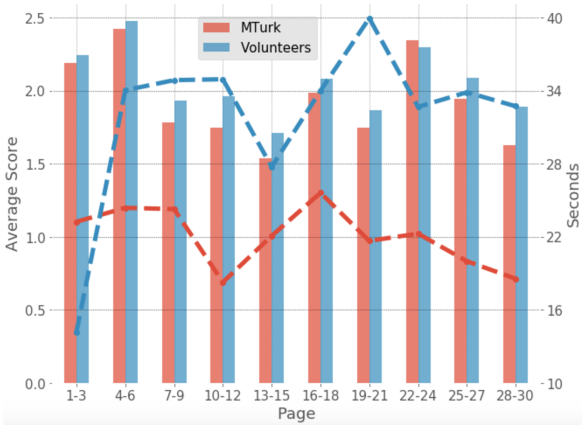

Figure 1: Time spent (line, right y-axis) and corresponding average scores (bars, left y-axis) for joke triples

the very beginning of the experiment and less so at the end. There is seemingly no 'fatigue effect' towards the later pages that some volunteers reported in private conversations. It is interesting to note that time spent per page and average triple scores are oppositely correlated in volunteers (−0.21) and MT workers (0.47).

4.2 Individual Jokes

In this section we make observations about the specific jokes. For example, below is the joke with the highest variation in ratings:

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

It is obvious that jokes about the tragedy of 2001 may seem inappropriate to many people. The greatest difference between average scores (0.97) can be observed between the groups of native speakers and participants with an average knowledge of English.

Despite the fact that men and women assess jokes in a similar way, their judgments are quite different on some jokes. Here's the joke which shows the largest difference (0.50) between men's and women's scores:

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

It can be assumed that women are more responsive to some topics, *sweets* in particular. This is confirmed by experiments on author profiling that list *chocolate* among most distinctive 'female words' [2]. The claims that men prefer cynical or violent humor is partially supported by the fact that male participants assessed the *9/11* joke 0.32 points higher than female participants.

The third column in Table 1 contains average scores by joke source. A rather low score of three Siri jokes is not necessarily indicative of her sense of humor in general. Two of these jokes used rather highbrow or complex words (the third one is the *chocolate* joke mentioned above):

Q: *What is the meaning of life?*
A: *I Kant answer that. Ha, ha.*

Q: *Why did the chicken cross the road?*
A: *I am not perspicacious about the peregrinations of poultry.*

Their low scores support the findings reported in the literature that funnier jokes tend to use more common, easily recognizable words [14, 19]. The third column in Table 1 shows rank correlations between joke 'source' rankings and rankings based on average scores obtained in our experiment. Original Yahoo! Answers jokes were ranked based on a difference of thumbs up & down counts. Automatically matched jokes were annotated by three assessors on a four-point scale in our previous experiment [6]. Reddit posts were ranked based on their points; tweets were ordered according to normalized scores (defined as a sum of retweets and likes divided by the number of an account's followers). The outcomes are mixed: in case of automatically matched jokes and Reddit, the rankings are very close; in case of Yahoo!Answers and Twitter, the ordering is quite different (negative rank correlation of tweets can be partially explained by the aggressive normalization – maybe popular accounts post inherently funnier tweets). Almost identical ranking of automatically matched jokes in both experiments can be explained by a very similar experimental design. We can assume that if one is interested in relative ranking of jokes rather than in obtaining absolute scores for individual jokes, a small group of assessors may be sufficient (though they will have a low agreement).

We took a closer look at the Jester jokes and compared score distributions for seven jokes in Jester data vs. our experiment. We mapped Jester's continuous values to our four-point scale and calculated Kullback-Leibler divergence (with Jester being the reference distribution). The KL-scores range from 0.04 (almost perfect coherence) to 0.48; average score over seven jokes is 0.26. The following joke has the highest divergence score:

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

Figure 2 compares the Jester scores for this joke to those obtained in our experiment. We can explain the discrepancies by the negative effect of a 'canned' joke – it sounds a bit dull in 2017 (recall, the first version of Jester went online in 1998). This observation as well as low-rank correlations with 'original' scores discussed earlier cast doubt on the success of transferring joke ratings between different domains and re-using old jokes.

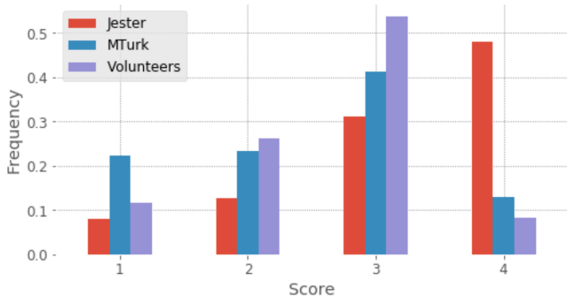

Figure 2: Distribution of joke #4 scores by Jester users/MTurk workers/volunteers (Jester continuous ratings mapped to four-point scale)

5 CONCLUSIONS AND FUTURE WORK

The results show that paid crowdsourcing workers and volunteers assess jokes in a very similar way. Thus, we can conclude that crowdsourcing is a viable and efficient option for humor evaluation. The findings also suggest that age, gender, and language proficiency can partly explain variation in funniness scores. We expect that these new insights will help to design humor evaluation and interpret its outcomes. They can also be directly applied to humor generation in conversational systems.

In the future, we are going to analyze the impact of personality traits on humor evaluation. It is expected that Big5 can better reflect the annotator's humor appreciation level. We will also focus on how joke topics influence subjects' scores. We plan to include education level as an additional feature in our future experiments.

REFERENCES

[1] Salvatore Attardo. 1994. *Linguistic Theories of Humor.* Walter de Gruyter.
[2] Angelo Basile et al. 2017. N-GrAM: New Groningen Author-profiling Model. In *CLEF'2017 Evaluation Labs and Workshop.*
[3] Nancy Bell and Salvatore Attardo. 2010. Failed humor: Issues in non-native speakers' appreciation and understanding of humor. *Intercultural Pragmatics* 7, 3 (2010), 423–447.
[4] Jerome R Bellegarda. 2014. Spoken Language Understanding for Natural Interaction: The Siri Experience. In *Natural Interaction with Robots, Knowbots and Smartphones.* 3–14.
[5] Kim Binsted. 1995. Using humour to make natural language interfaces more friendly. In *AI, ALife and Entertainment Workshop.*
[6] Vladislav Blinov, Kirill Mishchenko, Valeriya Bolotova, and Pavel Braslavski. 2017. A Pinch of Humor for Short-Text Conversation: an Information Retrieval Approach. In *CLEF.* 3–15.
[7] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval* 4, 2 (2001), 133–151.
[8] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *J. Res. Pers.* 37, 6 (2003), 504–528.
[9] Bryan Anthony Hong and Ethel Ong. 2009. Automatically extracting word relationships as templates for pun generation. In *CALC.* 24–31.
[10] Jiepu Jiang et al. 2015. Automatic online evaluation of intelligent assistants. In *WWW.* 506–516.
[11] Peter Khooshabeh et al. 2011. Does it matter if a computer jokes. In *CHI.* 77–86.
[12] Chia-Wei Liu et al. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP.* 2122–2132.
[13] Rod A. Martin. 2007. *The Psychology of Humor: An Integrative Approach.* Elsevier.
[14] Rada Mihalcea and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In *CICLing.* 337–347.
[15] Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence* 22, 2 (2006), 126–142.
[16] Tavi Nathanson, Ephrat Bitton, and Ken Goldberg. 2007. Eigentaste 5.0: constant-time adaptability in a recommender system using item clustering. In *RecSys.* 149–152.
[17] Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. In *SemEval.* 49–57.
[18] Willibald Ruch. 1992. Assessment of appreciation of humor: Studies with the 3WD humor test. *Advances in personality assessment* 9 (1992), 27–75.
[19] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *PKDD.* 1065–1074.
[20] Oliviero Stock and Carlo Strapparava. 2005. HAHAcronym: a computational humor system. In *ACL (demo).* 113–116.
[21] Alessandro Valitutti et al. 2013. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints.. In *ACL (2).* 243–248.
[22] Miaomiao Wen et al. 2015. OMG UR Funny! Computer-Aided Humor with an Application to Chat. In *ICCC.* 86–93.
[23] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor Recognition and Humor Anchor Extraction. In *EMNLP.* 2367–2376.