Robots Make Things Funnier

Jonas Sjöbergh and Kenji Araki

Graduate School of Information Science and Technology
Hokkaido University
{js, araki}@media.eng.hokudai.ac.jp

Abstract. We evaluate the influence robots can have on the perceived funniness of jokes. We let people grade how funny simple word play jokes are, and vary the presentation method. The jokes are presented as text only or said by a small robot. The same joke is rated significantly higher when presented by the robot. We also let one robot tell a joke and have one more robot either laugh, boo, or do nothing. Laughing and booing is significantly funnier than no reaction, though there was no significant difference between laughing and booing.

1 Introduction

While humans use humor often in daily interactions, computer systems are still far from able to use humor freely. When it comes to computer processing of humor in various ways, a good overview is [1]. Two main areas of computer implementations exist, humor recognition and humor generation. Systems generating quite simple forms of jokes have been constructed [2–4]. There are also systems that try to recognize whether a text is a joke or not [5, 6].

Our paper is mainly relevant for generation systems. What is considered amusing varies a lot from person to person. Many things have an influence on how funny something is perceived to be, such as the general mood at the time or who is delivering the joke. This makes evaluating and comparing jokes and joke generating systems complicated. We evaluate the effects of different ways of delivering simple jokes to evaluators. Our hypothesis was that jokes would be perceived as funnier when presented by a small robot than when presented in text form. We also hypothesised that the same joke would be perceived as funnier if another robot for instance laughs at the joke than with no reaction.

2 Evaluation Method

We automatically collected about 1,400 simple word play jokes in Japanese. This was done by searching the Internet for a few seed jokes and then automatically downloading for instance all items in a HTML list where one of the seed jokes occurred. The downloaded results were also manually cleaned up. We then selected 22 jokes from this list with the main criteria being that the joke should be very short (since the robot model we are using has a very limited amount of

memory for voice samples) and be understandable when spoken (some jokes in the list only work in written form).

Two robots were used, both of the same model: Robovie-i[1], a fairly small robot that can move its legs and lean the body sideways. It also has a small speaker, though the sound volume is quite low and the sound quality is poor. The main features of the Robovie-i are that it is cute, easily programmable, and fairly cheap. One of the robots used is gold colored and one is blue. Whenever a robot was about to speak, it first moved its body from side to side a little, and then slightly leaned looking up at and pointing the speaker (mounted on the stomach) towards the evaluator. Other than that, the robots did not move.

The robot voice was generated automatically using a text-to-speech tool for Japanese, AquesTalk[2]. The robots use different synthetic voices, so it is possible to distinguish which robot is talking only by listening. The text-to-speech conversion works fairly well, though the produced speech is sometimes difficult to understand. It is flat, lacking rhythm, intonation, joke timing etc. The voice is somewhat childlike, and sounds "machine like", like most cheap text-to-speech.

For the first experiment, ten jokes were divided into two sets, set 1 and set 2, of five jokes each. Jokes in set 1 were always presented first and then the jokes in set 2. To half of the evaluators (group A) the jokes in set 1 were presented using one of the robots and to the other half (group B) these jokes were presented only in text form. The same was done for the jokes in set 2 too, but if an evaluator had set 1 presented using the robot then the same evaluator would have set 2 presented using text and vice versa. Any one evaluator was only shown the same joke once, all jokes were shown to all evaluators, and the jokes were always presented in the same order. Evaluators were assigned to group A or B based on the order they arrived in, e.g. the first ten to arrive ended up in group A, the next ten in group B, etc. This means that all jokes were evaluated an equal number of times using the robot and using text.

For the second experiment twelve jokes were divided into three sets of four jokes each. These were then presented by having one robot tell the joke and the other robot either laugh a little and say "*umai*" ("good one"), say "*samui*" ("cold", as in "not funny"), or do nothing. As in the first experiment, the jokes were presented to all evaluators in the same order, and all evaluators were presented with each joke exactly one time. Set 1 was made up of jokes 0, 3, 4, and 8; set 2 of jokes 1, 2, 5, and 9; and set 3 of jokes 6, 7, 10, and 11. Evaluators were assigned to either group C, D, or E. All groups had different reactions for each set of jokes, so the second robot would laugh at four jokes each time, boo at four jokes, and make no reaction at four jokes, but different jokes for different groups. Which group had which reaction to which set of jokes is shown in Table 3. All jokes were presented with each reaction to the same number of evaluators.

Evaluators were found by going to a table in a student cafeteria and setting up the robots and a sign saying that in exchange for participating in a short robot experiment they would get some chocolate. Only native speakers of Japanese

[1] http://www.vstone.co.jp/top/products/robot/Robovie-i.html
[2] http://www.a-quest.com/aquestal/

could participate. The evaluations were done one person at a time, so if many arrived at the same time some would have to wait their turn. Evaluators were asked to grade all jokes on a scale from 1 (boring) to 5 (funny).

As the cafeteria background was fairly noisy, compounded by the poor speaker, it was sometimes hard to hear what the robot was saying. In such cases the joke was repeated until the evaluator heard what was said. A more quiet background would have been better but finding a large number of evaluators prepared to go to a different experiment place would have been very difficult. We have since then managed to find better speakers, though not available in time for the experiment.

3 Results

In general, the evaluators were happy to participate, though most people passing by ignored the evaluation. In total, 60 evaluators, 17 women and 43 men, participated in the experiment. The scores of the jokes vary wildly from person to person. The lowest mean score for all jokes for one person was 1.3 and the highest 3.9 for the first experiment and 1.3 and 3.8 for the second experiment.

3.1 Robot vs. Text

The results of the first experiment are presented in Tables 1 and 2. Table 1 shows the mean scores of the different sets of jokes using the robot and text, and Table 2 shows the scores of each joke. These are of course also influenced by how interested each evaluator was in this type of jokes in general.

Table 1. Mean evaluation scores for the three sets of jokes using different presentation methods. Which group evaluated which set using which method is given in parenthesis.

Set	Robot	Text
1	2.5 (A)	2.2 (B)
2	3.0 (B)	2.4 (A)
All	2.8	2.3

The total average scores in Table 1 is perhaps the most interesting result to focus on. It gives a good comparison between the two methods. Any specific evaluator is giving a score to the same number of jokes for both methods, and every joke is present an equal number of times for both methods. As hypothesized, the robot presentation method gets a higher mean score than text, 2.8 compared to 2.3. Though the standard deviation in the scores is quite high, 1.2 for both methods, the difference is significant ($\alpha = 0.01$ level, Student's t-test).

Looking at the individual jokes in Table 2, nine of the ten jokes were perceived as funnier when presented by the robot than by text, though in many cases the

Table 2. Mean evaluation scores using different presentation methods.

Joke	Total	Robot	Text
0	2.2	**2.5**	2.0
1	1.9	**2.0**	1.9
2	2.8	**3.0**	2.5
3	3.0	**3.2**	2.9
4	1.8	**2.0**	1.5
5	2.5	**3.1**	1.9
6	2.3	**2.7**	2.0
7	2.8	**3.2**	2.4
8	2.7	**2.8**	2.6
9	3.1	3.1	**3.2**
Average	2.5	2.8	2.3
# Highest Score		9	1

difference was small. Accounting for the multiple number of comparisons, only two jokes (jokes 5 and 7) were significantly funnier at the $\alpha = 0.05$ level using the robot.

The Pearson correlation between the jokes in text form and the same jokes presented by the robot is 0.73, indicating a fairly high correlation. This indicates that the robot improves the impression of the jokes, and not that the robot is simply funny in itself (which would make all jokes told by the robot be about as funny). Some jokes are improved more than others, which could depend on many things. Some factors are the quality of the robot voice for the words in question, if the joke is new or old, if the joke is a joke on the person telling the joke etc. Joke 1 (scored similarly in text and using the robot) is for instance a very well known Japanese joke that most people no longer find interesting while joke 5 (scored a lot higher using the robot than using text) is a joke that has a repeated sound that sounds funny, but does not come through quite as well in text it seems.

3.2 Laughter, Booing, or No Reaction

The results of the second experiment evaluating the influence of a second robot either laughing, booing, or giving no reaction at all to the telling of a joke, are presented in Tables 3 and 4. Table 3 shows the mean scores of the different sets of jokes having different reactions, and Table 4 shows each individual joke.

As before, the total average score for each presentation method in Table 3 is perhaps the most interesting result to focus on, since any specific evaluator is giving a score to the same number of jokes for every reaction type, and every joke is present an equal number of times with each reaction. As hypothesized, the mean scores are higher with some form of reaction than with no reaction, averaging 2.8 for laughter and 2.6 for booing, compared to 2.2 for no reaction.

Table 3. Mean evaluation scores for the three sets of jokes using different presentation methods. Which group evaluated which set using which method is given in parenthesis.

Set	No reaction	Laughter	Booing
1	1.9 (E)	3.1 (D)	2.9 (C)
2	2.0 (C)	2.2 (E)	2.5 (D)
3	2.7 (D)	3.1 (C)	2.4 (E)
All	2.2	2.8	2.6

Again, the standard deviation in the scores is quite high, 1.0 for laughter and no reaction, and 1.1 for booing. The differences between laughter and no reaction and between booing and no reaction are significant on the $\alpha = 0.01$ level (Student's t-test, α adjusted for multiple comparisons), while the difference between laughter and booing is not significant.

Table 4. Mean evaluation scores using different presentation methods.

Joke	Total	No Reaction	Laughter	Booing
0	2.6	1.6	3.0	**3.0**
1	2.5	2.3	2.5	**2.8**
2	2.0	1.9	**2.1**	**2.1**
3	3.0	2.5	**3.3**	3.1
4	2.7	1.7	**3.2**	3.1
5	2.3	1.9	2.1	**2.8**
6	2.5	2.2	**3.2**	2.2
7	2.7	2.5	**3.0**	2.5
8	2.4	1.9	**2.9**	2.5
9	2.0	1.8	1.8	**2.4**
10	3.0	**3.4**	3.1	2.5
11	2.6	2.6	**2.9**	2.2
Average	2.5	2.2	2.8	2.6
# Highest Score		1	7	5

Looking at the individual jokes in Table 4, only for one joke out of twelve was the no reaction presentation better than the other methods.

Thus, despite some problems with hearing and understanding what the robots said, the robots did make things funnier. The difference in mean score of 0.5 between text and robot is rather large, considering that the average score was only 2.5. The same is true for the difference between no reaction and laughter (or booing), 0.6 (0.4) for an average score of 2.5. Evaluations of the impressions of robots performing *manzai* (Japanese stand-up comedy) have shown similar re-

sults to ours. The overall impression of the robots was rated higher than viewing amateur comedians perform the same routine on TV [7].

4 Conclusions

We evaluated the impact of different presentation methods for evaluating how funny jokes are. We found that the same joke was perceived as significantly funnier when told by a robot than when presented only using text. The average scores were 2.8 (robot) and 2.3 (text), which is a quite large difference in this scale from 1 to 5. This means that it can be difficult to compare the evaluations of different joke generating systems (or other sources of humor) evaluated at different times, since even the presentation method used has a very large impact on the results. There are likely many other factors that influence the evaluation results too, making it difficult to compare different systems.

We also evaluated the impact of having another robot laugh, boo, or do nothing when a joke was told. This too made a significant difference to the perceived funniness of a joke, with an averages of 2.8 (laugh), 2.6 (boo), and 2.2 (no reaction). The robot always laughed and booed in the exact same way. A more varied set of reactions would probably be even funnier.

In future experiments we want to examine the effect of speech only (is the voice or the robot funny), and include a small training evaluation to remove any effect of a joke being funny because it is the first appearance of the robot.

Acknowledgments

This work has been funded by The Japanese Society for the Promotion of Science (JSPS). We also thank the anonymous reviewers for insightful comments.

References

1. Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., O'Mara, D.: Computational humor. IEEE Intelligent Systems **21**(2) (2006) 59–69
2. Binsted, K.: Machine Humour: An Implemented Model of Puns. PhD thesis, University of Edinburgh, Edinburgh, United Kingdom (1996)
3. Binsted, K., Takizawa, O.: BOKE: A Japanese punning riddle generator. Journal of the Japanese Society for Artificial Intelligence **13**(6) (1998) 920–927
4. Yokogawa, T.: Generation of Japanese puns based on similarity of articulation. In: Proceedings of IFSA/NAFIPS 2001, Vancouver, Canada (2001)
5. Taylor, J., Mazlack, L.: Toward computational recognition of humorous intent. In: Proceedings of Cognitive Science Conference 2005 (CogSci 2005), Stresa, Italy (2005) 2166–2171
6. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Proceedings of HLT/EMNLP, Vancouver, Canada (2005)
7. Hayashi, K., Kanda, T., Miyashita, T., Ishiguro, H., Hagita, N.: Robot manzai - robots' conversation as a passive social medium. In: IEEE International Conference on Humanoid Robots (Humanoids 2005), Tsukuba, Japan (2005) 456–462