What Does 3.3 Mean? Using Informal Evaluation Methods to Relate Formal Evaluation Results and Real World Performance

Jonas Sjöbergh and **Kenji Araki**[1]

Abstract. We have created an automatic humor generation system for Japanese that generates two man comedy routines or humorous responses to free text input. Evaluating humor is rather difficult since humor is subjective and many factors influence the perceived funniness. We evaluated our system in several ways. First, a traditional evaluation with evaluators ranking comedy performances from 1 (boring) to 5 (funny) gave a result of 3.3. To complement this evaluation and to see if 3.3 is good enough for real world usage we entered a comedy performance created using our system in a funny robot competition with a ¥500,000 prize. We did not win the ¥500,000, but we made it to the final and could use the audience reactions during the live performance for evaluation. We also did sentiment analysis of blog postings covering the competition. These evaluations, while informal, indicate that our system performs competitively compared to the other contestants, that were all handmade systems. That the system could compete against human made contributions shows that a score of 3.3 can be "good enough" for real world applications. We believe that evaluations like these can be useful despite being very informal, since they measure what we want to know: do people expecting something funny think the system is funny?

1 Introduction

Humans often use humor but computers are so far bad at this. Humor makes daily interaction smoother and more enjoyable, helps people deal with troubling events, and generally makes people feel better. We believe it would be good if machines too could use humor and understand when humans use humor, though this is a difficult task. Work on computational humor has been done on both humor recognition [12, 6, 8] and humor generation [2, 13, 4], see [3] for a good overview of the field.

Humor is subjective, and many other factors also influence whether something is perceived as funny or not. This makes evaluating and comparing computational humor systems difficult. Most evaluations consist of asking humans to rate how funny they think generated jokes are, which is a reasonable evaluation method but has some problems. It is difficult to compare systems evaluated at different times, under different circumstances, or by different evaluators, since so many factors other than the quality of the joke also influence perceived funniness. The same joke presented to the same evaluator at different times can receive different scores, depending on for instance the mood of the evaluator. We have also confirmed in previous experiments that the way a joke is presented has a fairly large impact. Having a joke read to you by a small robot was significantly funnier than just reading the same joke written on a paper [10]. The lack of standard baselines or gold standards to relate the evaluation results to also means that it is difficult to judge if an achieved result is good enough to be useful.

We have a humor generation system that has been evaluated in the traditional way of asking evaluators to rate how funny they thought the system output was. The latest version achieved a score of 3.3 [9] on a scale from 1 to 5. As an earlier version scored 2.8 in a different evaluation [7], it seems likely that the system has improved, but how good is 3.3?

Reported results from other systems are often low, but it is hard to compare with results from evaluations under different conditions. It is also somewhat hard to construct a relevant baseline or upper bound. Also, as far as we know, there are no other systems that generate full performances connecting different jokes etc., though there are several systems generating stand alone jokes. Thus, there is no similar research to relate to either. It is of course possible to hire professional comedians to create performances similar to what the system generates and then compare the system to human level performance. A quick subjective check of the system output shows that it is quite far from human level performance, though, so spending money on this is currently not that appealing. There are no freely available sources of human made humor that are similar enough to the system output to be useful as comparisons.

Simple baselines for a lower evaluation limit can be constructed by instantiating the patterns used by the system with random words instead of words chosen by the system. Such text is really boring for the evaluators to wade through though, and since we are not allowed to pay our evaluators it is hard enough to get reasonable numbers of participants already.

We did however want to complement this evaluation and preferably relate this score of 3.3 to real world performance. We found an opportunity for this in a competition for silly robots. The competition rules were that the system must have some mechanical parts, have no serious purpose, and must make people laugh. The funniest system, based on a two minute performance and if qualifying for the final also a 15 minute interview like performance chance, wins ¥500,000. The competition is described in more detail in Section 4.

We entered the competition with a performance generated using our system as a tool. Would a system scoring 3.3 be good enough to compete against human made contributions? The other contestants would thus be our "gold standard" to compare against, and one evaluation metric would be if our system could do reasonably well in comparison to the other teams or not.

We also tested the system capabilities by generating humorous re-

[1] Hokkaido University, Japan, email: {js, araki}@media.eng.hokudai.ac.jp

sponses to arbitrary requests from audience members at the live final. Would this score of 3.3 mean the system was robust and general enough to handle unrestricted input in real time and still entertain people?

Since the conditions of the contest were not under our control, a lot of the evaluation measurements would likely be hard to quantify or be rather subjective. What would be a reasonable method of measuring if the hundreds of spectators were entertained or not? We decided that since this was our first attempt at complementary evaluation methods, we would use very simple measure such as "a large part of the audience laughing is good, booing or uncomfortable silence after jokes is not good". We also complemented this with analysis of blog entries written by spectators discussing what they thought of the event, as a more easily measurable evaluation metric.

2 System

We have built a humor generation system for Japanese made up of modules that generate different types of jokes. It can thus easily be extended with new modules or adapted to new tasks by changing how the modules are used. Below we describe the modules and three variations of the system using these differently. More detailed descriptions can be found in [9]. Most modules are fairly simple though modules using sophisticated reasoning etc. to generate jokes could be added if methods that work reasonably well become available.

2.1 Database

This module represents the system reusing jokes it has heard other people tell. We automatically collected a database of word play jokes in Japanese using a few seed jokes, web queries, and pattern matching. A simple example of how this module works is searching the web for occurrences of a joke the system already knows. If two seed jokes (or jokes found earlier by the module) occur in the same context (e.g. an HTML list) the module downloads all other content that occurs in an identical left and right context (e.g all the other list items). The contexts are determined automatically by finding the longest common left and right contexts of the two known jokes in the page, similar to the method used in [14]. The database has almost 3,000 jokes, mostly word play jokes.

When a joke related to some word is needed, a joke containing this word that has not already been used is returned. An example would be one robot saying: "Speaking of 'life' *[assuming this word was mentioned earlier in the performance]*, 'shotgun wedding: a case of wife or death' comes to mind."

2.2 Proverbs

The proverb module has a database of sayings in Japanese. To generate a joke a proverb is twisted into a new variant by replacing content words with similar sounding dirty words. These are taken from a collection of a few hundred dirty words that have also been grouped into the categories sex related, feces related, and insults. If two or more dirty words from the same category sound similar enough to content words of the same word class, a joke is generated by replacing the original words with these.

The jokes are normally presented in patterns like "Recently my life has felt like *proverb*" "Oh? For me it has been more like *dirty proverb*". Instead of proverbs, other inputs can also be used, such as in the real time system detailed below.

An example in English similar to what the system produces in Japanese (direct translation of word play jokes do not work well) is one robot saying: "Recently my life has felt like 'jack of all trades, master of none', you know." The other robot then makes a vulgar joke by modifying this to "For me it has been more like 'jacking off all trades, masturbate none', I must say.".

2.3 Riddles

When a word play riddle is needed for a word, the module checks if it has any similar sounding dirty words. A riddle is formed along the lines of: "A *word* is a *word*, but what kind of *word* is *hint*?", "What?", "*Answer: similar sounding dirty word*". Hints are generated by searching a large corpus for phrases like "a *dirty word* is *hint*". Hints found are assumed to be reasonable descriptions if they co-occur more strongly with the answer than with the original word.

The riddle module can also be used with other patterns. When using the system as a tool, it gave jokes like "Give me an example sentence for '*content word*'", replying with an example sentence for a similar sounding dirty word.

The riddle jokes are in general quite weak, similar to: "'speaking' *[a word used earlier]* is 'speaking', but what is a naughty kind of speaking?", "What?", "Spanking the monkey!" ("speak" sounds like "spank" and "spanking the monkey" is a sexual euphemism).

2.4 Responses

Japanese stand-up comedy is usually performed in pairs, and the interaction between the straight man and the funny man is important. Generally the straight man must chastise the funny man when he says something stupid or misunderstands something. We have a module that generates responses like these, mainly using a database of generic responses. This module was only used in the stand-alone system. Example outputs are things like "Idiot, please stop." or "Yeah, yeah. Very funny..."

2.5 Basic Stand-Alone System

The basic version of the system is completely automatic and generates stand-up comedy performances in Japanese, to be performed by two performers. This system uses the different joke modules to generate jokes and ties them together into a coherent performance by simple transitions like "Speaking of *word from one joke*, *another joke related to this word*". It uses text-to-speech to generate sound files and uploads these into two small robots (see Figure 1) that perform the generated stand-up comedy. This is the version of the system that was evaluated in a previous experiment [9] in the traditional way of having evaluators rate the funniness of the output.

The generated performances normally contain one proverb joke, one riddle joke, and two database jokes. The responses from the straight man in the performance can also lead to more jokes based on these being included in the performance.

2.6 System as a Tool

Since the completely automatic system sometimes makes mistakes, such as generating jokes that are incomprehensible or sentences with broken grammar etc., we decided to use the system as a humor generation tool when competing against human made performances in the silly robot competition. Humor is a very difficult problem, so having

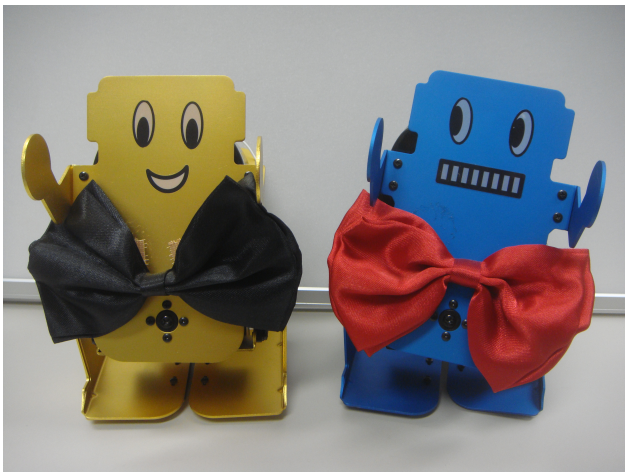

Figure 1. The robots performing the generated comedy.

a system that can be used as a joke generation tool is still a good first step, we believe.

All the jokes in the performance were generated by the system but we manually selected which jokes to include in the performance, removing jokes that were too hard to understand (using rare or difficult words) and correcting minor mistakes. We also manually reordered the jokes and wrote the transitions between them. Parts of this work could be automated, for instance checking how common words or expressions are. Finally, since the text-to-speech output is hard to understand even under good conditions, we expected it to not be enough for a live performance in front of hundreds of audience members. Instead we had two students read the generated output into a microphone and then made it sound "robotic" by using a phaser filter. In the end, we were also given the use of a big monitor, which we used to show subtitles of what the robots were saying.

Database jokes were not used in the competition entry version of the system. We wanted to see if the automatically generated jokes were funny enough to compete against man made contributions and thus did not want to include any man made jokes. This led to pretty much all the jokes in the performance being dirty jokes.

2.7 Real Time System

We also constructed a real time version of the system to be used at the live event if we made it to the final. It takes an input statement and calls the different modules to see if any of them can produce a relevant joke. Since the database is the fastest option, first the system extracts content words from the input and sees if there is any joke in the database containing one of these content words that has not already been used. If so, it outputs one randomly selected such joke. If this fails, the system tries the proverb module, treating the whole input statement as if it was a proverb and then proceeding as described earlier. If this also fails, the riddle module is called with each content word from the input until a riddle joke can be generated. If this too fails, the system outputs statements along the lines of "La la la, I cannot hear you, lalala".

2.8 Humor Enabled Chat-Bot

The real time system has also been used as a component in a simple humor enabled chat-bot for Japanese [11]. The chat-bot contains the modules described above, and some modules that are not humor related. Non-humor modules include such things as mentioning trivia information related to the user input, reacting on user input that seems unusual, simple backchannel statements, talking about the weather, and more. The chat-bot also has a humor related module for detecting if the user tells a joke and then reacting appropriately (laughing). This module uses the database of jokes collected for the database module above and if the user writes something that is similar enough to any of the database jokes, the chat-bot laughs.

In a small evaluation with 20 evaluators, our humor enabled chat-bot outperformed two other chat-bots for Japanese, though both of these were fairly simplistic chat-bots. Our system received the highest scores for all evaluation criteria, including "human likeness", "understandability", "wish to continue chatting", "interesting/funny", and "the system you would prefer to chat with".

Other research on humor enabled chat-bots has also been done, mainly for English [5, 1, 4].

3 Traditional Evaluation

As mentioned, we previously evaluated the stand-alone system using the traditional method of having evaluators watch the generated performances and rate them from 1 (boring) to 5 (really funny). 33 evaluators took part in the evaluation. Each evaluator watched two performances and was then asked to rate each performance as well as the separate parts of each performance. See Table 1 for the results for the whole system and [9] for more detailed evaluation results.

Table 1. Mean (and standard deviation) evaluation scores (1=boring, 5=funny) and the number of 4s or 5s assigned by 33 evaluators rating two performances each.

	Script 1	Script 2	Both
Score	3.4 (0.9)	3.2 (1.0)	3.3 (1.0)
4 or 5	16 (48%)	14 (42%)	30 (45%)

The total impression of the system was 3.3 on average with very low variation between performances but as can be expected very high variation between different evaluators.

It is difficult to judge if 3.3 is good enough for any specific use or if it is very funny at all. Another measure is that 45% of the evaluators gave the performances a 4 or a 5 which reasonably indicates that they liked them. Since computational humor systems generally fare quite badly (humor is difficult and the system outputs are thus generally not that funny), we believed that this was a quite good result, but how good? To find out more we tried two other tests for the system.

4 Silly Robot Competition

4.1 Background

Bacarobo 2008[2] was the second installment of the Bacarobo silly robot competition. The rules are that each system has to be mechanic in some way, must have no serious purpose, and must make people laugh. The funniest system wins ¥500,000. Bacarobo is held in

[2] http://bacarobo.com/

Japan, and contestants can enter performances in either Japanese or English, though so far only systems for Japanese have made it to the finals. In 2007, 11 systems made it to the live performance final, but in 2008 only 5 teams were to be selected for the final. In the end, they received too many good contributions and put 6 teams in the final.

The competition is organized by *Maywa Denki*[3], which is an art performance company. The finals are held in cooperation with *Yoshimoto Kogyo*, a large Japanese entertainment company that employs most of Japan's famous stand-up comedians. The final is held in a *Yoshimoto* theater. In 2008 it was in the *Yoshimoto ∞Hall* in Shibuya, Tokyo. It seats slightly over 200 guests, who had to pay an entrance fee of 3,000 yen and the event was more or less sold out. The audience was a fairly even mix of (overwhelmingly Japanese) men and women. People were of all ages, even a few children which was not so good for our very foul mouthed robot performance.

The first stage of the competition was sending a video including a two minute competition performance and an additional three minutes of time allotted for explaining what your robot is doing. At the live final, each team was given a short presentation and then entered the stage. After performing the previously submitted two minute performance live, a rather long (about fifteen minutes) interview like explanation session followed, the contents of which varied a lot between teams. The jury consisted of four minor celebrities: the company boss of *Maywa Denki*, a manga writer/paper columnist, a movie director, and a professor from Tokyo University. The live event was led by a famous comedian.

4.2 Other Teams

We here give a brief description of the five other teams that made it to the final. They are presented in the performance order in the final.

Generally, a lot of performance time was used by the accompanying people performing and interacting with the robots, while the robots themselves were fairly passive.

4.2.1 Magaru-Ken & Tono-Metto

Magaru-ken & Tono-metto consisted of a samurai haircut robot and a sword robot that were used as lie detectors and nervousness detectors. For instance, the sword is held in the hand and when the holder starts sweating it bends to indicate stress. The performance consisted of a man wearing the haircut robot on his head and holding the sword in his hand being asked questions by a woman in a white lab coat. One question was "Have you ever bought a used school girl swim suit in an Internet auction?", to which he answered "No" but the robots indicated nervousness and possible dishonesty. The other questions were in similar vein. During the interview part of the competition they used the robots on the jury members, asking for instance if they had rented their own movies to raise the rental statistics.

4.2.2 YKRN

YKRN was a face recognition robot that had the face of a Japanese female celebrity stuck on a movable arm that swung around looking for human faces. When it found a face the eyes lit up and it stopped moving around. For faces of people it recognized it had various programmed reactions, such as trying to kiss any jury member that came close and to completely ignore the comedian presenting the final. This was the winning team.

[3] http://www.maywadenki.com/

4.2.3 Our System: Yonashi

The third performance was our system, with two robots performing a comedy routing created using the system as a tool as described above.

4.2.4 Kangaeru Robo

Kangaeru robo was a replica of a thinking man sculpture that had a computer screen instead of a face and one more computer screen showing what the robot itself was seeing. It also did face recognition by recognizing a big green board with a cut out for your face that volunteers from the audience were asked to put to their faces. When it recognized a face it switched the stick figure face it normally had for the audience member's face, and changed the face in the image of what the robot was seeing to be the stick figure face instead of the audience member's original face. The name of the system came from the fact that the robot was said to know what people it switched face with were thinking. The robot was asked what the audience member was thinking, and answered with various embarrassing things, typed into the system by a team member back stage.

4.2.5 Push-Kun & Yome-Push

Push-kun & Yome-push were two trash cans with legs that walked around and spoke to each other and their maker. They also did a dance number that finished with the female robot giving birth to a small baby robot. One of the robots was in the final in the previous Bacarobo contest too.

4.2.6 Tsuneo & His Friends

Tsuneo & his friends consisted of four high school girls and three robots. One robot was a vacuum cleaner stuffed in a shell that looked like a pig that "ate" (by vacuuming it up and throwing it out the other end) glitter fed to it by one of the girls. The second robot was another vacuum cleaner stuck inside a green elephant shaped shell that blew into three party blowout-noisemakers when touched. The last robot was a small sleeping man that snored very loudly and whose stomach grew when snoring. In the end, it exploded.

4.3 Results

Our contribution, made by using our system as a tool and containing only computer generated jokes (no jokes from the database) made it to the final but did not win. When the robots performed the generated jokes the jury and audience members laughed at the "correct" places and most people seemed to enjoy the performance though this of course varied from person to person. One jury member laughed incredibly much, for instance. He later told us that he had had very high expectations just from reading the program where our team was indicated as "obscene robots".

The main result of this part of the evaluation is the fact that we made it to the final, though an even better result would have been to win the whole contest. The individual placements in the final were not given, only which team was the overall winner, so a more precise measure than between place 2 and 5 was not available. We believe just reaching the final against man made contributions is a fairly strong result. The other finalists were all very good/funny.

5 Live Audience Interaction

5.1 Details of the Performance

After the pre-generated two minute performance at the live event, we also showed the real time capabilities of the system. The system output was text only, and the text was shown on several big monitors around the theater. First the head of the jury requested a joke related to the sponsors of the event. This was set up beforehand. This was so as to set the mood, show how this part of the performance was to transpire, and to guarantee at least one example that the system could handle successfully (the audience did after all have to pay to see the event).

It turned out that we misspelled the sponsor name and input *Oronami shi* (instead of *Oronamin C*). The system part-of-speech and word boundary analyzer (Japanese does not mark word boundaries) found the word *nami* (wave) when misanalysing this. It output a database joke: *Nami to ieba, Konishi Manami no yoko ni, Shimanami Kaidou* ("Speaking of waves, right beside *Konishi Manami* I saw *Shimanami Kaidou*" – the last part of "beside" plus the name of a road sounds like the name of an actress). This got a good reaction.

Next, a small girl in the first row was asked for her favorite word. She said *Manafi*, which is the name of an animated character in a children's show. No module could generate a response so the failure mode kicked in and *kikoenai, kikoenai, kikoenai* ("Lalala, I cannot hear you.") was output. This also got a lot of laughs and reactions along the lines of "Very human like reaction!". This indicates that the system fails somewhat gracefully, and that the recovery options in the system are useful when the system cannot produce sophisticated output.

Prompted for her favorite food the girl said *ringo* ("apples"). The system output: *Ringo ga jimen ni ochita tte iu ton* ("People say the apple fell to the ground." – the last part sounds like "Newton" in Japanese); another database joke, giving an OK reaction.

Next a jury member asked for a joke regarding *"Touhoushinki"*. We did not know it at the time but it is the name of a Korean boy band. Not knowing the proper Chinese characters to use we input it using phonetics (*hiragana*). The system misanalysed this unknown word as several short words and returned a database joke related to *hou* (in the boy band name it roughly means "direction", but in the joke found it means roughly "ho" as an interjection in for example "Ho ho ho"). *Hou to ieba: "Houkama ha aitsu kamo!" "Houka, houka"* (Speaking of "ho": "That guy might be the pyromaniac!" "Hoho, is that so"). This got a good reaction.

Finally we were asked for a Christmas joke and got: *Kurisumasu ha kuri de sumasu* ("I make do with a chestnut for Christmas"); a proper Christmas joke with good reactions.

Apart from this interaction with the system, we were also asked a lot of questions by the jury members. Questions included such things asking for an explanation of the underlying technology, asking what we had come to Japan to do (implying that our system did not seem to be serious work), and asking why our clothes looked so strange.

5.2 Results

The results of this part of the evaluation is first that the system was robust enough to output jokes reasonably relevant to the input for all inputs except one. As can be expected, the database joke module was heavily used, mainly because it is called first and is applicable to a wide range of inputs. Large parts of the audience laughed at each joke, indicating that the system not only can output something, but that the output is also funny. Of course, parts of this could be because of things like the system misunderstanding the inputs being funny.

Not only did people laugh a lot at the generated system output, they also seemed very impressed (an "ohhh" of feeling impressed went through the theater) when during the explanations that followed we mentioned that the system searches the Internet for jokes automatically. That such things can be done seemed to be more impressive than the first two minute performance.

We also had people come up to us after the show was over to tell us that they liked our system. Many also wanted to take photos of us and the robots. Other contestants also took pictures of our system (mostly our laptop, not the robots, for some reason) back stage.

6 Sentiment Analysis of Blog Reports

6.1 Analysis Method

After the contest was over, we downloaded blog postings covering the Bacarobo 2008 event and did sentiment analysis of these. We did the sentiment analysis by hand, since there was not that much text to deal with and since automatic methods are somewhat lacking in accuracy. We ignored postings mentioning only that someone went to the event, since we wanted data on the impression of the different performances. We also ignored press release like postings just mentioning the contest and the name of the winning team. Only postings that actually described the contents of at least one performance from the contest were examined.

For each performance that reached the final we checked four different things: First, how many postings mention the performance in question at all, and second how much text was used for describing the performance (measured in number of characters). The basic idea is that if people care about a performance they will be more likely to mention it and also to write more about it. If a performance is not mentioned, it did not make enough of an impression for the writer to care. Generally, for these two points a higher score is better, though of course having only negative statements written about your system would give high scores without indicating a good performance.

The third and fourth things examined were the number of positive statements (like "the crowd laughed a lot", "I was impressed") and the number of negative statements written about the performance. Most postings spent most of the text on fairly factual descriptions of the contents of the performances, with some subjective impressions added at the end or sometimes interspersed throughout, so there were not that many positive and very few negative statements.

What should be considered a positive statement is also somewhat up for debate, since quite a few comments called the performances "stupid". Since the goal of the competition is to be stupid, it may not be negative in this context, though in many cases it was not very positive either. We did not count statements referring to performances as "silly/stupid" (*baka*) as positive or negative, since most of them probably just played on the name of the competition (*Bakarobo*, "silly robots"). Negative sounding comments like "the system looked scary" are not necessarily negative either. We counted statements as negative only when the writer was clearly criticizing the performance, such as by saying "I did not understand what the point was" or "the movements were very awkward". We did not take into account the strength of the sentiment, i.e. if the comment was very strongly negative or just mildly negative, though this could have been done since the analysis was done by hand. For negative comments there was not much difference in the strength, while there was some variation among the positive comments.

6.2 Results

The results of the sentiment analysis are shown in Table 2. The winning team of Bacarobo 2008 was YKRN.

Table 2. Results of sentiment analysis of blogs. For each team, the number of blogs mentioning the performance, the amount of text (number of Japanese characters) discussing it, and the number of positive and negative statements about the team were checked.

Team	Mentioned	Text Size	Pos.	Neg.
Magaru-ken & Tono-metto	9	8,681	15	1
YKRN	8	9,365	15	2
Our System	9	9,723	25	5
Kangaeru robo	6	8,189	11	0
Push-kun & Yome-push	6	7,209	14	5
Tsuneo & his friends	7	6,698	6	2

Our system is one of the three performances mentioned by the most number of bloggers, though one of these mentions seemed to be more concerned with the fact that we were the only foreign team than with the performance itself. It is also the system which has the largest amount of text dedicated to it, and it has the most positive statements. The reason for the high number of positive statements is just one blogger, who seemed to be very impressed with our system, though. Not counting this blogger our system still has the most positive statements, though the difference to the next best performance, the winning team YKRN, is then only 1.

Our system also has the most negative comments, tied with one other performance. The negative comments about our system and performance were almost exclusively that the jokes were too dirty, especially since there were children watching. There was also one negative comment about the t-shirt worn when starting the performance and answering questions from the jury, which while part of the performance is not really a property of the system itself. There were also quite a few comments about some sound problems we had during the beginning of our performance, though we did not count these as negative statements about our system since the fault was in the sound system of the theater and not in our system.

The positive statements were mainly that people were impressed with how advanced the system was (or they thought it was), in relation to the real time interaction with the audience and jury. Other positive statements were that the jokes were very dirty (two people liked that), that the robots were very cute (one comment), that it was nice to see a foreign entry (said by a blog in English), and that one person appreciated that we were prudent or commonsensical enough to also prepare the big monitors for subtitles.

7 Discussion

Most results that can be taken from this evaluation are of course quite informal. We do however think they give an interesting complementary view and tell us things not possible to learn from more traditional and strict evaluation methods.

All in all, both the audience and the jury members seemed to enjoy both the pre-generated naughty joke performance, and perhaps especially the real time generation of jokes. Quite likely the latter was considered impressive because it was completely clear that the computer was doing all the work, while the pre-generated performance could just as well have been human made (like the other competition entries), though we did explain that all jokes were computer generated.

Other feedback apart from the fact that we reached the final, and the laughter and general happy impression of the jury and audience members there, include: the other participants (most of them university researchers) being impressed by our system and photographing it back stage (despite it looking like an ordinary laptop); audience members commenting to us after the show that they thought it was funny and that the robots were cute; and the blogs covering the event, analyzed in the previous section.

Reaching the final of a competition and getting good reactions from an audience are of course quite informal ways of evaluating the system. We believe it shows that the system is "funny enough" to be useful though, since the goal of the system is to entertain people. This is much harder to tell from the results of a more traditional though strict and scientific evaluation. The sentiment analysis of blogs covering the event showed that our system was actually one of the systems that received the most mentions (and the most text dedicated to it) and the most positive comments in blog postings about the event. The sentiment analysis results are fairly even among most teams, so this is not a very strong result, but at least it shows that our system is not markedly worse than any of the human generated performances. Some systems received markedly less coverage than our system.

Our system did however get the most negative comments too (tied with one other system). All of the relevant comments were about the system generating jokes that are too dirty, or at least inappropriate when performing in front of children. Making a more child-friendly version is thus one of our current projects.

Apart from the results already discussed, it was also very interesting for us just to try to apply the system in these settings that were very different from the normal research settings. There were many problems to overcome that would not have been found if we had not entered this competition. Examples include hardware problems when adapting to the requirements of having several hundreds of audience members that had to hear the robot performance at the same time, real time requirements in the audience interaction, trying to adapt to the very varied composition of the audience (children, foreigners with less knowledge of Japanese etc.), and much more. A result that we were quite pleased with ourselves was the simple fact that we managed to pull off the live performance at all.

In conclusion, while there is a lack of baselines or other comparisons to relate the previous system evaluation result of 3.3 to, these informal evaluations showed that people are entertained enough to still be satisfied after having paid a quite high entrance fee (seeing a professional comedian would have been cheaper). It also shows that the system is robust enough to handle unrestricted input. It is not handled perfectly by any means, but well enough to keep people entertained, the main goal of the system. The system thus seems like a quite useful application of computational humor.

As there are not that many humor competitions for robots to participate in (though there were Bacarobo contests in both 2007 and 2010 top), performing this type of evaluation may not be easy. With no humor contest available, similar evaluations can perhaps be done by using the system during "open stage night" at local comedy clubs or similar activities. Organizing a competition yourself likely requires a lot of work, but could perhaps be done with less effort as a part of some other event such as a school festival.

We would also like to mention a suggestion we received from a very helpful reviewer. Given that the competition we participated in was videotaped, it would be possible to do evaluations by showing videos of all performances to evaluators and having them rate them for instance on a scale from 1 to 5. This would give more controlled results while still allowing for comparison to other systems.

8 Conclusions

We created an automatic humor generation system for Japanese. A traditional evaluation where human evaluators scored the generated content on a scale from 1 to 5 gave a score of 3.3. We wanted to relate this score to actual real world performance and to test if this score of 3.3 would mean that the system was good enough to entertain people.

As a complementary evaluation we entered a comedy performance created by using our system in a "silly/funny robot" competition with a ¥500,000 prize, where our system would compete against human made contributions. Our two robots performing stand-up comedy generated by our system and some manual filtering was good enough to reach the final of the competition. The final consisted of a live performance at a comedy theater in front of a paying audience.

The audience response was good. A session of requests for jokes on unrestricted topics from audience and jury members also went well, with a lot of laughter. Sentiment analysis of blogs covering the event showed that our system compared very well to the other teams, though we did get the most negative statements too. These were about the system generating jokes that are too dirty. Though we did not win the 500,000 yen prize, the experience still showed that the score of 3.3 seems to indicate that our system is robust and funny enough for real world applications.

We believe using complementary evaluations methods like how far you can go in a competition, sentiment analysis of things written about your system, or informal measurements like "does the audience seem to be laughing a lot at the correct times of a live performance" can give interesting results not easily found using traditional evaluation methods. The evaluation results are of course very informal in many cases, but since it seems difficult to measure things that we actually want to know, such as "do people who paid to see something funny actually find the system funny enough", informal measurements are better than nothing, and they can hopefully tell us interesting things. Competitions like this can perhaps also be a good way to compare the performance of different systems that produce output that would be difficult to compare by more formal methods.

Just applying the system to real world conditions also showed us many things, e.g. the system is fast enough for real time applications and robust enough to handle input from a very varied audience. We also noticed and fixed real world application related problems like scaling the audio output to be heard by hundreds of people at once. Going out into the real world taught us many new things about our system and it is an experience we recommend to others too.

Acknowledgements

We would like to thank Shinsuke Higuchi and Hajime Wakahara for doing the voices for the robots.

Most of the system was developed in a project funded by the Japanese Society for the Promotion of Science (JSPS).

REFERENCES

[1] Agnese Augello, Gaetano Saccone, Salvatore Gaglio, and Giovanni Pilato, 'Humorist bot: Bringing computational humour in a chat-bot system', in *CISIS'08: Proceedings of the 2008 International Conference on Complex, Intelligent and Software Intensive Systems*, pp. 703–708, (2008).

[2] Kim Binsted, *Machine Humour: An Implemented Model of Puns*, Ph.D. dissertation, University of Edinburgh, Edinburgh, United Kingdom, 1996.

[3] Kim Binsted, Benjamin Bergen, Seana Coulson, Anton Nijholt, Oliviero Stock, Carlo Strapparava, Graeme Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, and Dave O'Mara, 'Computational humor', *IEEE Intelligent Systems*, **21**(2), 59–69, (2006).

[4] Pawel Dybala, Michal Ptaszynski, Shinsuke Higuchi, Rafal Rzepka, and Kenji Araki, 'Humor prevails! – implementing a joke generator into a conversational system', in *Proceedings of AI-2008*, eds., Wayne Wobcke and Mengjie Zhang, volume 5360 of *Lecture Notes in Artificial Intelligence*, pp. 214–225. Springer, (2008).

[5] Dan Loehr, 'An integration of a pun generator with a natural language robot', in *Proceedings of the International Workshop on Computational Humor*, (1996).

[6] Rada Mihalcea and Carlo Strapparava, 'Making computers laugh: Investigations in automatic humor recognition', in *Proceedings of HLT/EMNLP*, Vancouver, Canada, (2005).

[7] Jonas Sjöbergh and Kenji Araki, 'Automatically creating word-play jokes in japanese', in *Proceedings of NL-178*, pp. 91–95, Nagoya, Japan, (2007).

[8] Jonas Sjöbergh and Kenji Araki, 'Recognizing humor without recognizing meaning', in *Proceedings of WILF 2007*, eds., Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, volume 4578 of *Lecture Notes in Computer Science*, pp. 469–476, Camogli, Italy, (2007). Springer.

[9] Jonas Sjöbergh and Kenji Araki, 'A complete and modestly funny system for generating and performing Japanese stand-up comedy', in *Coling 2008: Companion volume: Posters and Demonstrations*, pp. 109–112, Manchester, UK, (2008).

[10] Jonas Sjöbergh and Kenji Araki, 'Robots make things funnier', in *Proceedings of LIBM'08*, pp. 14–19, Asahikawa, Japan, (2008).

[11] Jonas Sjöbergh and Kenji Araki, 'A very modular humor enabled chat-bot for Japanese', in *Pacling 2009*, pp. 135–140, Sapporo, Japan, (2009).

[12] Julia Taylor and Lawrence Mazlack, 'Toward computational recognition of humorous intent', in *Proceedings of Cognitive Science Conference 2005 (CogSci 2005)*, pp. 2166–2171, Stresa, Italy, (2005).

[13] Hans Wim Tinholt and Anton Nijholt, 'Computational humour: Utilizing cross-reference ambiguity for conversational jokes', in *CLIP 2007*, Camogli, Italy, (2007).

[14] Richard Wang and William Cohen, 'Language-independent set expansion of named entities using the web', in *Proceedings of ICDM'07*, pp. 342–350, Omaha, U.S., (2007).