Automatically Choosing Appropriate Gestures for Jokes

Dai Hasegawa and **Jonas Sjöbergh** and **Rafal Rzepka** and **Kenji Araki**
Hokkaido University
{hasegawadai, js, kabura, araki}@media.eng.hokudai.ac.jp

Abstract

We examine the problem of automatically selecting gestures that are appropriate to use when telling a joke or a short story. Our current application of this is a joke telling humanoid robot that needs to be able to select natural gestures for arbitrary input. The topic is important because humans use body language and gestures, thus socially interactive robots should also be able to do so for more natural interaction. We asked evaluators to assign appropriate gestures from a set of gestures the robot can perform to 50 jokes from a corpus of jokes in Japanese. We then evaluated different methods for automatically selecting gestures on this data set. While human inter-agreement was rather low, indicating that this is a fairly difficult task and that some jokes have no obviously fitting gesture, the best method performs on par with humans and clearly outperforms the baseline.

Introduction

Body language and gestures are important and very frequently used parts of human communication. When creating embodied agents, social robots, or any kind of artificial agent that interacts with humans, it is a good idea to include gestures in the interaction to make it more natural. Some work has been done on generation of gestures together with speech, see for instance (Cassell, Vilhjálmsson, and Bickmore 2001) for a description of a state of the art system and an overview of the field. This system, like most systems, relies on a semantic knowledge base for a specific narrow domain to generate gestures that are human like and appropriate for the content. Other work has been done by Breazeal (2000), using among other things an emotional model of the robots "state of mind" to determine appropriate facial expressions and gestures. Analysis of human behavior, what gestures go with what types of contents etc., and a knowledge base based on a lot of human work can give very good results.

For languages other than English, it is generally difficult to find semantic knowledge bases of any reasonable size, though for very limited domains there are many. We wanted to examine how difficult the problem of assigning appropriate gestures in story telling, i.e. telling a short story, monologue, or joke, is and how far one can come without resorting to the use of deep semantics, since such resources are not generally available in more unrestricted domains. We do not want to limit the applications to a narrow domain, but aim instead for methods that can deal with unrestricted text.

We examine the problem of selecting gestures for a joke telling robot. The reason for using jokes is that it is easier to find willing evaluators to help out in the experiments if the experiments are entertaining, and jokes being innately funny help in this regard.

We programmed a robot to be able to perform a set of eight gestures meant to draw attention to or illustrate different body parts, thus taking advantage of the robot being an embodied agent. We then asked evaluators to select which motions were appropriate for each of 50 jokes from a database of word play jokes in Japanese. Based on this data set we evaluated three methods for automatically selecting gestures. While the human inter-agreement was quite low, the best method performed on par with humans and clearly outperformed the baseline.

The methods are general enough to be applicable in many other settings too. Any setting where unrestricted input is expected would have problems with approaches relying on pre-specified rules for which gestures go with what inputs etc. Examples where automatic means of assigning appropriate gestures would be useful include animating story telling or socially interacting robots, adding appropriate body language automatically to avatars in games when users are chatting freely with each other, adding gestures to characters in voice controlled games etc.

General Setting

We use a Speecys SPC-101C[1] robot, which is a small humanoid robot, see Figure 1. It is about 30 cm high and has 22 degrees of freedom of motion. It also has LEDs in the hands and chest, speakers for producing sounds, and a camera, though these were not used in our experiment.

As a starting point to determine the difficulty of choosing appropriate gestures for arbitrary input, we examine how easy it is to use the fact that the robot is an embodied agent. Thus, the gestures reflect this, and all gestures are meant to display a body part. There are of course very many other

[1] http://www.speecys.com/mirai.html

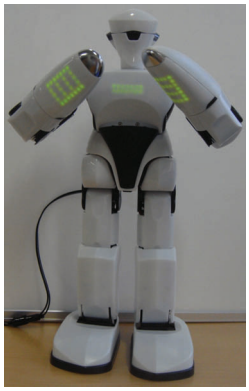

Figure 1: The robot used.

gestures that are of use in joke or story telling, such as gestures for adjectives like "big" or "small", words for feelings like "tired" or "cold", and much much more. As a first step to see if the methods are feasible we limit the gestures to only one type, body parts. It is difficult for some of the methods to compare if "big" is better than "buttocks" for jokes about someone's big buttocks, for among other reasons that the usage of adjectives and nouns are quite different. The gestures used are listed in Table 1 and snapshots from performances of the gestures are shown in Figure 2. In later stages we will of course add different types of gestures too. These are the gestures used in the experiment:

- For the "hand" gesture the robot puts out its right hand in front of itself and shakes the hand from side to side.
- For the "foot/leg" gesture the robot takes a few steps forward.
- For the "head" gesture the robot points with both hands towards the head and gently shakes its head.
- For the face gesture the robot turns its head slightly to the right and points to the middle of the face with the right hand.
- For the "stomach" gesture the robot was meant to touch its belly with both hands but in the end it turned out to not be possible to do that given the joints in the robot body. The final pose looks more like both arms pointing straight forward at stomach level.
- For the "chest" gesture the robot touches its chest slightly below the face with both hands.
- For the "buttocks" gesture the robot bends slightly forward shooting its behind out. It also pulls the left hand behind itself and waves it towards the buttocks and out a few times.
- For the "both hands" gesture the robot raises both hands above its head.

The gestures were not generated with much thought put into them in this exploratory experiment, but we do plan to study videos of comedy performances and model gestures for joke telling in Japanese on these at a later stage.

Japanese	English
te	hand
ashi	foot/leg
atama	head
kao	face
hara	stomach
mune	chest
shiri	buttocks
ryoute	both hands

Table 1: The labels for the gestures used in our experiments.

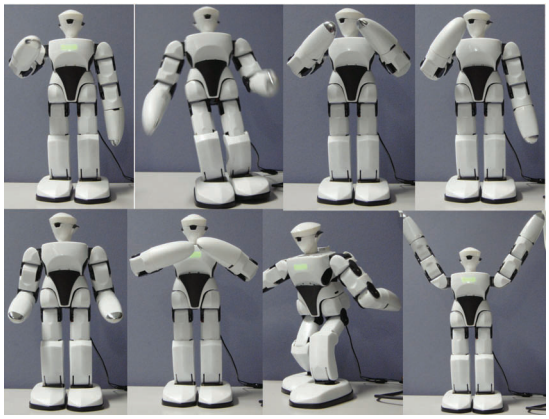

Figure 2: The gestures. Top row from left to right shows "hand", "foot/leg", "head", and "face". Bottom row shows "stomach", "chest", "buttocks", and "both hands".

We use jokes from the same joke database as the one collected in (Sjöbergh and Araki 2008). This database was collected by automatically downloading jokes in Japanese from the Internet using pattern matching and a few seed jokes. If for instance two seed jokes occurred in the same HTML list, all other list items were also downloaded. It contains almost 3,000 jokes. The jokes are word play jokes in Japanese, of a type called *oyaji gyagu* or *dajare* (roughly corresponding to "puns" in English).

We selected 50 jokes from the corpus that seemed to have some relation to a body part, for example jokes containing words such as "ear ring", "contact lens", "chopsticks" or "eating". Many jokes are very abstract and even a human would find it difficult to come up with any gesture relevant to the contents.

We showed the robot performing the different gestures and had evaluators assign appropriate gestures to each joke. We then evaluated different methods for automatically selecting gestures by comparing with the manually selected gestures.

Methods

The methods described here are given a list of the available gestures that the robot can perform and a joke. Given these, the methods output which gesture would be most appropri-

ate when telling this joke and optionally when in the joke this gesture should be applied.

The list of the available gestures can be given in many different ways, it is for instance possible to give several written labels for the methods to use that all refer to the same actual movement of the robot. The labels of the gestures the robot is currently programmed to perform are listed in Table 1.

Baseline

The baseline method simply selects the gesture that is most common (giving the highest score) in the data, which is either the "hand" motion or the "stomach" motion, depending on the evaluation criteria used.

Word Co-Occurrence in a Large Corpus

The co-occurrence methods use statistics on the co-occurrences of content words in the jokes and the gesture labels to determine which gesture to use, basically gestures with labels that have high co-occurrence with the contents of a joke are assumed to be appropriate. We tried two different approaches to measuring co-occurrences, though the only real difference between them is the data they use. One uses a corpus of slightly over one million web documents that we have downloaded ourselves (originally for a different research project) and one uses the "Japanese Web as a Corpus" (Erjavec, Kilgarriff, and Erjavec 2007) corpus of about 12 million tokenized and part-of-speech tagged sentences of Japanese, also originally taken from the Internet.

For our own web corpus we downloaded 20 GB (gigabytes) of web pages in Japanese and indexed them with the Hyperestraier search engine. This corpus was originally intended for use in a different project, and for reasons related to that has a fairly strong bias towards pages with "dirty words". Given an input joke, content words (nouns, verbs, and adjectives) are extracted using the MeCab[2] syntactic analyzer for Japanese. The co-occurrences of each content word with each gesture label are found in the corpus and used to determine which gesture to use. Stop words are ignored when extracting the content words.

Since the log-likelihood ratio has been suggested as a good measure for rare events (Dunning 1993), for a pair of a content word and a gesture label we calculate the log-likelihood of the pair given the number of occurrences of the words separately and of the words together. Words have to co-occur in the same sentence, not just the same page, to count as co-occurring.

When log-likelihood ratios have been calculated for all gesture labels and all content words from the joke, the label that had the highest log-likelihood ratio with any content word is selected. If there are many content words in the joke, a gesture can be assigned to each content word, beginning with the one that has the highest log-likelihood with a gesture label. In this paper, only the top suggestion was output, even if many possibly suggestions were found.

If no gesture label co-occurs with any content word, the method outputs the "nothing" label, indicating that it could not find any appropriate gesture. The "nothing" label is also used as one of the labels in the normal log-likelihood calculations. The idea is that if no gesture is more strongly connected to the joke than the "nothing" label, then probably none of the gestures are really appropriate and outputting "nothing" (no appropriate gesture found) is reasonable. There are likely better ways to determine if the gesture labels are connected strongly enough than comparing to the word for "nothing", though this seems to work fine as a threshold and has the added bonus of making the implementation very simple.

Since Japanese has no space between words, naive searching gives noise in the form of short words seeming to co-occur with other words when they are in fact not present but the same letter sequence is present as a substring of a longer word. As this method is already very time consuming even with a fairly small corpus of 20 GB, especially for longer jokes with many content words or for jokes with very common content words, we did not run the (quite slow) syntactic analyzer on the corpus data to remove this noise. In the same vein, it is also possible to use only the top N pages with occurrences of both the content word and the label and calculate the log-likelihood based on how often they co-occur in the same sentence in these pages. Since for $N = 1,000$ this still seems to give reasonable log-likelihood ratios we use this threshold in the experiments. It is much faster than using the whole corpus, though $N = 1,000$ is still quite slow.

It is of course much faster to just check how many pages the words co-occur in, since this is what the search engine has indexed. This does however not give satisfactory results.

Two other things to counter this are to index sentences instead of pages, and to syntactically analyze the whole corpus before searching. To try this we used the Japanese Web as a Corpus, JPWAC (Erjavec, Kilgarriff, and Erjavec 2007). It is also built from web pages, and has been part-of-speech tagged and split into sentences. Since the JPWAC was morphologically analyzed with a different tool than MeCab that we use, what constitutes a word is different. To remove this problem, we analyzed all sentences with MeCab and used the word boundaries assigned by MeCab in our experiments.

Other than the corpus data being different, the method does not differ. We still calculate the log-likelihood ratios for the words in the jokes and the gesture labels, based on the co-occurrences in the sentences in the corpus. This corpus is however smaller than our own (already quite small) corpus, with word frequencies of for instance the gesture labels being about 50% higher in our corpus.

Hypernyms, Hyponyms, and Synonyms using WordNet

Instead of using co-occurrences in a large corpus, we also tried using an ontology. Gestures with labels closely related to the contents of a joke, measured by closeness in the ontology link structure, are assumed to be appropriate. We used the Japanese WordNet (Isahara et al. 2008) to find which words in the input were most similar to which gesture labels. The Japanese WordNet is currently a direct translation of the English WordNet (Miller 1995), though it lacks some

[2]MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/

of the information in the original. This means that it is not made with the structure of the Japanese language in mind, and that many words common in Japanese but with no corresponding common expression in English are not present.

For a given input the system again used MeCab to extract nouns and verbs. Since the verbs and nouns in the Japanese WordNet are not connected and all gesture labels are nouns we first extracted nous related to the verbs in the input using the Internet. The Internet was searched for patterns of the form "* *wo* <verb>", "* *ni* <verb>", and "* *de* <verb>". The "*wo*", "*ni*", and "*de*" particles in Japanese are markers of direct object, indirect object, and instrument or location, respectively. From the top 100 snippets returned from each of these queries, the top five nouns matching these patterns were extracted.

These extracted nouns and the nouns found in the input were then used. Similarity in WordNet was measured as how many links the closest path from a gesture label to a noun contained. Hypernym and hyponym links count as a distance of one, while all synonyms in a synset count as having a distance of zero. If the number of links between a gesture label and a noun is n then the score for the gesture label is set to $\frac{1}{n+1}$, i.e. a word counts as having a score of 1 compared to itself, and otherwise the score gets lower with the distance. The total score for a gesture label is the sum of the scores for the distance to each of the nouns found.

The "nothing" label is considered as having a score of 0.25, so if no label has a sum higher than this the method suggests using no gesture at all. This means that if one noun is within three steps, or two nouns occur within a distance of four etc. the gesture will be used. If several gestures are above the threshold they can all be output, but in our experiments only the highest scoring gesture was used.

Evaluation

We first showed the robot performing the eight gestures to volunteer evaluators. They then got a list of 20 jokes and were asked to assign which gestures were appropriate for each joke. Since several gestures can be appropriate for the same joke, any number of gestures could be assigned. If no gesture seemed appropriate, it was also possible to select "nothing" meaning none of the gestures available would be appropriate. The evaluators could have any gestures they liked shown again at any time. No evaluator assigned more than two gestures to any joke and the vast majority of jokes received only one gesture for each joke, on average 1.05 gestures per joke and evaluator.

A total of 25 evaluators took part in the evaluation, all of them were university students with a balanced mix of men and women. With 25 evaluators evaluating 20 jokes each, each of the 50 jokes was assigned gestures from 10 different evaluators. The agreement between the evaluators was quite low. Fleiss's Kappa measure of inter-rater agreement (Fleiss 1971) is only 0.19, indicating low agreement.

When evaluating the automatic methods we calculated two different measures. The first is the "average percent", which means that for each suggested gesture, the percentage of the evaluators that also assigned this gesture is checked.

Method	Average (%)	Over Half
Highest Possible Score	48	29 (100%)
Best Human	34	25 (86%)
Human Average	28	17 (59%)
Worst Human	19	10 (34%)
Baseline	14	5 (17%)
Annotated Sentences (JPWAC)	19	7 (29%)
Unannotated Web Pages	25	13 (45%)
WordNet	23	12 (41%)

Table 2: The agreement on the suggested gestures.

The total score is then the average of these percentages for each suggestion. This calculation allows a method to assign any number of gestures to each joke, though in the evaluations the methods only output their top suggestion. The maximum achievable score is 80%, since no joke had a higher percentage of evaluators assigning the same label than eight evaluators out of ten. Only three jokes had 80% agreement on the same gesture, so since all methods suggest gestures for each input, the actual maximum achievable is lower. Always suggesting the gesture with the highest agreement for each joke would give a score of 48%.

The second measure is the "over half" measure. This counts the number of suggested gestures that agree with gestures assigned by at least half the evaluators (i.e. at least five). If at least half the evaluators agree on the same gesture, it is reasonable to think that this is in some sense a "correct" or "common sense" gesture to use. There were only 29 jokes for which at least half of the evaluators agreed on a gesture.

The baseline consists of simply always assigning the gesture that would give the highest score, which is the "hand" gesture for "average percent" and "stomach" for "over half".

We also calculated scores for each of the evaluators. For "average percent", the calculation is the same as for the automatic methods, i.e. for each suggested gesture what is the percentage of (other) evaluators that agree on this gesture. For "over half", removing one evaluator (the one to be evaluated) from the set of 10 for each joke leaves 9 suggestions, so no really fair comparison of agreeing with over half of the other evaluators can be done. We calculated a generous version of this, which means agreeing with at least four other evaluators (with the evaluator himself being the fifth making half of the evaluators agree), giving a measure that is slightly easier for the evaluator than for the automatic methods to score highly on, thus giving a good upper bound on how well the system could be expected to work in an ideal case on this data set. Since each evaluator only evaluates 20 jokes (and the automatic systems 50) the scores are also adjusted by a factor 2.5 to be more easily comparable.

The results of our experiments are shown in Table 2. It can be seen that while the automatic methods clearly outperform the baseline, and even the worst performing method is almost on par with the human evaluator that had the lowest agreement with the other evaluators, they still have some way left to reach the average human level. The hy-

Method	Average (%)	Over Half
Annotated Sentence (JPWAC)	20	8 (28%)
Unannotated Web Pages	27	14 (48%)

Table 3: The results after adjusting the gesture labels.

ponyms/hypernyms in WordNet perform quite well, as does searching in our own local web corpus. Using the Japanese Web as a Corpus (JPWAC) data that has sentence and word boundaries annotated (and also part-of-speech, though we did not use that information), but on the other hand is smaller, performs quite a lot worse. The word co-occurrence data are of course very sparse, so perhaps this explains why adding annotation and removing noise does not compensate for the lack in size.

One problem for all methods that was quite obvious to us was the simplistic choice of labels for the gestures that the robot performs. While we started with a list of body parts and then created some gestures to draw attention to these body parts, the thinking of the human evaluators on what this gesture could be used to mean was not restricted to body parts. The gesture for "both hands" was for instance used to mean "surprise" or "being upset" by the evaluators, and the gesture for "head" which involved pointing at the head with the hands and gently shaking the head was often used to indicate disagreement, sadness, or pain.

To examine how much of an impact the gesture labels might have, we also did one more small experiment. We ran the co-occurrence methods one more time, with two new labels added. The gesture for "head" was also given the label "*iie* (no)", and the gesture for "both hands" was given the label "*bikkuri* (surprise)", the two gestures that seemed to be used the most with a different intention than just drawing focus to the actual body part. These words do not occur in the Japanese WordNet, so the WordNet method cannot be improved with these labels. While this is a very slight modification compared to doing an extensive check of what humans think the actual gestures represent, it gives an indication of how much difference using more appropriate labels for the gestures could make.

The results of this modification are shown in Table 3. Even with this very simple modification the results are improved in both cases, though not by a huge amount. The best method now performs comparable to the human average, and quite a bit better than the lowest scoring human evaluator.

Discussion

All in all, the results are quite promising. Even these rather simplistic methods and naive ways of labeling the gestures, the performance is almost at the same level as the average human agreement.

One experience from the current evaluation is that human evaluators are not very good or consistent in assigning gestures to jokes in a task like this, shown by the low inter-rater agreement. For future experiments we would prefer to have volunteers tell the jokes and videotape them to see what gestures they actually do use and build the gold standard on this. It will likely be more difficult to find people willing to be videotaped than to just answer a questionnaire, though. For jokes, another alternative is to have professional comedians create the gold standard, assigning gestures that professionals believe are appropriate for the various jokes.

With gestures modeled on what gestures humans actually use (within the constraints of what the robot body is actually capable of doing) and gesture labels assigned by examining what the gestures are used to mean, we expect the results to be improved further. This would also likely help in making evaluation data using a similar evaluation method "cleaner". If the gestures are more natural the agreement between the evaluators should be higher. Of course, by adding more gestures and gestures with meanings that are difficult to compare, in many ways the task for the system will also become harder.

The methods for selecting gestures can of course also be improved. One simple option is combining the different information sources, using both the synonymy and hyponymy information in WordNet and the word co-occurrence information in the large corpora. Adding more sophisticated analysis of the corpus texts is another option, though there is a lack of available tools. Also, even the simpler annotation used (part-of-speech, word boundaries, sentence boundaries) already has problems with both the web data and especially the jokes. The jokes contain non-standard choice of alphabets (Japanese uses several different alphabets mixed together), words written to imitate the sound of someone speaking in dialect instead of the standard orthography, made up words, etc., all of which makes the analysis of the text difficult and the analysis results less than ideal.

Another modification of the methods we want to add is to first determine which content words are likely to be accompanied by gestures. Given this, gestures appropriate for just these content words can then be assigned. In related research this is often based on studying how humans use gestures. For instance, humans tend to use gestures for new information, or surprising information, see for instance (Cassell et al. 1994).

A lot of research has been done on human usage of gestures, see for instance (McNeill 1992), and we would like to make use of such results. Combining manual work on semantic models or simple rules for selecting gestures for high accuracy on common cases with automatic selection for rare cases to improve the coverage seems like a good idea.

Conclusions

We examined the difficulty of the task of selecting appropriate gestures to go with arbitrary input, in our case word play jokes told by a joke telling robot. Human evaluators were asked to choose which of a set of programmed robot body gestures were the most appropriate for 50 jokes in Japanese. We then evaluated different methods for assigning gestures automatically based only on the text. While the human inter-agreement was quite low, the automatic methods performed on par with the agreement between the human evaluators.

This shows that even using quite naive labels for the gestures and using shallow analysis (no hand written semantic knowledge etc.) quite good results can be achieved. The best performing method uses only unannotated text and tools for finding word class information, word boundaries, and sentence boundaries in the jokes. Having access to semantic information or deeper analysis methods can likely improve the results, but since our goal is to be able to handle arbitrary text such resources are hard to come by, but our shallow methods work on any input text.

For the future we want to model gestures on gestures used by humans in similar contexts, and to label the gestures based on what humans perceive them to mean. We would also like to use a different evaluation method, since the inter-rater agreement was rather low. We also want to combine the different gesture selection methods and add more sophisticated analyses.

Acknowledgements

The authors would like to thank the Nissan Science Foundation for their financial support.

References

Breazeal, C. 2000. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Cassell, J.; Steedman, M.; Badler, N.; Pelachaud, C.; Stone, M.; Douville, B.; Prevost, S.; and Achorn, B. 1994. Modeling the interaction between speech and gesture. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 153–158.

Cassell, J.; Vilhjálmsson, H. H.; and Bickmore, T. 2001. BEAT: the behavior expression animation toolkit. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 477–486.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.

Erjavec, T.; Kilgarriff, A.; and Erjavec, I. S. 2007. A large public-access Japanese corpus and its query tool. In *CoJaS 2007*.

Fleiss, J. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Isahara, H.; Bond, F.; Uchimoto, K.; Utiyama, M.; and Kanzaki, K. 2008. Development of Japanese WordNet. In *LREC-2008*.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Miller, G. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41.

Sjöbergh, J., and Araki, K. 2008. A complete and modestly funny system for generating and performing Japanese stand-up comedy. In *Coling 2008: Companion volume: Posters and Demonstrations*, 109–112.