AAAI Technical Report FS-12-02
Artificial Intelligence of Humor

Human Judgment on Humor Expressions in a Community-Based Question-Answering Service

Masashi Inoue
Yamagata University
National Institute of Informatics

Abstract

For understanding humorous dialogue, a collection of humorous expressions is needed. In addition to humorous expressions, their annotations are important to be used as language resources. In this paper, we analyzed how human assessors annotate humorous expressions extracted from an online community-based question-answering (CQA) corpus, which contains many interesting examples of humorous communication. We analyzed the annotation results of a collection of humorous expressions as done by 28 annotators in terms of the degree of humor and categorization of humor. We found the assessments to be quite subjective, and only marginal inter-annotator agreements were observed. This result suggests that the variability in humor annotations is not noise resulting from erroneous assessment but is rooted in personality differences of the annotators. It would be necessary to incorporate the individual differences in humor perception for properly utilizing the resources. We discuss the possibility to improve the collection process by applying filtering techniques.

Introduction

Humor plays an important role in human communication, and its linguistic and psychological frameworks have been studied (Attardo 1994). Researchers had been interested in the assessment of sense of humor as the matter of personality. For example, a measure of humor has been developed that is shown to be well correlated with observed behavior and peer ratings (Martin and Lefcourt 1984). Recently, a new kind of communication has emerged in recent times: online communication, e.g., community-based question-answering (CQA) services. This paper focuses on humorous responses to questions in a CQA service and the analysis of how people perceive humorous expressions. The results of our study show that there is little consistency in the annotation of humorous expressions by human annotators.

Humor in Community-based Question-answering Services

Before the emergence of online communication, question-answering services were provided as one-to-one interactions between knowledgeable experts and laymen. Thanks to the asynchronous nature of online written communication, multiple users can now answer the same question simultaneously. In other words, a question or topic posted to the CQA service often elicits multiple answers provided by different users. While most answers are serious in nature, some are humorous. We present examples of serious and humorous answers as extracted from the service:

Question: My PC freezes when I play an online war game. Is this because of the machine spec? Can I fix it if I upgrade my RAM? I'm using Pen M 1.5G, memory 256M.

Serious Answer: Your 256M memory is too small let alone for online 3D war games! That size of memory can barely run XP and you want to play games...

Humorous Answer: That's a pacifistic PC...

Upon observing other examples, we note that there are no word-level or phrase-level similarities among humorous answers to different questions.

Computational Humor Analysis

Humor analysis is necessary to enable humor generation by artificial intelligence. Some computational approaches have been developed that use online text to understand humor expressions. For example, $16,000$ one-liners and self-contained short jokes have been collected using bootstrapping on ten seed expressions (Mihalcea and Strapparava 2005). Similarly, humorous text has been collected from the news website Onion, where all articles are assumed to be humorous (Mihalcea and Pulman 2007). When online communication is considered, it would be more interesting to see the examples from interactive communications rather than broadcasting. For example, expressions were collected from twitter stream using its hash tag #humor (Reyes, Rosso, and Buscaldi 2012). Another example of a source of humorous text is online bulletin boards such as Slashdot (Reyes et al. 2010). Selection of comments with the "funny" tag resulted in the extraction of $159,153$ items containing humorous text. Users of the website are expected to write something humorous occasionally and they would receive feedback on the writing via the tag system.

Here, an important distinction should be made in the form of humor generation. In the first examples, humor expres-

sions were *prepared* and presented. In the last example, humor expressions were *spontaneous* that were stimulated by the post of other users. The prepared humor has been studied by using scripted comedy or drama before the emergence of online communication. Therefore, we believe that the spontaneous humor communication is an important aspect to exploit digital text fully. Also, we believe that it is more useful in generating humor expressions by machines that interact with human.

Collecting Humor Candidates

In this paper, we examined humorous communication in a CQA service, Yahoo! Chiebukuro, which began in 2004. In 2005, this service was renamed as Yahoo! Answers, a global CQA service with 200 million users worldwide. We used the Yahoo! Chiebukuro CQA corpus (first edition)[1] that was collected from April 2004 to October 2005, and its total size was 4.1 GB. All text used in our experiments was in Japanese. We extracted humorous answers from the CQA corpus by following the procedure reported in (Inoue and Akagi 2012). After identifying a typical answer as the most serious example, a character-based trigram was used as the dissimilarities measure between serious and humor answers. As a result, candidate answers were ranked by calculating how humorous they are. Questions on the Yahoo! Chiebukuro corpus are posted according to various categories. In this study, we used two categories that contain more humorous expressions than others: Love and human relationships (Love) and Chiebukuro (Miscellaneous). The numbers of candidate answers according to these categories are summarized in Table 1.

Evaluation Conditions

Annotators and Questionnaire

We asked 34 participants, all of whom were undergraduate students majoring in computer science, to fill a questionnaire. It contained a list of questions extracted from the CQA corpus. Each question was accompanied by up to a maximum of five candidate humorous answers that are considered to be humorous by the automatic data collection tool. The questionnaire was administered in hard copy, and the participants, i.e., annotators completed it in privacy. First, annotators were asked to categorize questions, but this information was not used in this study. Then, they were asked to evaluate the automatically collected candidate answers according to their perceived degree of humor. They were also asked to categorize all the answers into one out of four humor types. Out of the 34 annotators, three did not return the questionnaire and three did not complete them properly Therefore, we acquired 28 annotations at the end of the survey, as listed in Table 1. Next, we provide more detailed explanations of the annotation procedure.

Degree of Humor

The annotators rated candidate answers on a scale of zero to three. Each rating is explained as follows.

[1]http://www.nii.ac.jp/cscenter/idr/en/yahoo/yahoo.html

Table 2: Type of Humor

Answer Type	Description
A	Just punning
B	Tripping up questioners
C	Sidestepping questions
D	Others (Please specify)

[0:] This answer is not intended to be humorous.
[1:] This answer is intended to be humorous but is not humorous.
[2:] This answer is intended to be humorous and is somewhat humorous.
[3:] This answer is intended to be humorous and is quite humorous.

These descriptions were provided to assist annotators and to clarify how they perceived the expressions.

Types of Humor

We set the following four types of humor on the basis of the observation of collected humorous expressions as listed in Table 2. Since these types of humor are not concrete concepts, we referred to them by letters A, B, C, and D, instead of naming them. Examples were provided for types A, B and C, but for type D, annotators were asked to describe the type of humor according to their understanding.

Evaluation Results

Score Distribution

Owing to space restrictions, we herein present results only for the "Love 2004" category out of the four categories, because this category was annotated by a relatively larger number of participants. As shown in Figure 1, assessments of humor varied greatly among annotators. Consistency in assessment was observed in the case of higher mean ratings (quite humorous) or lower means ratings (not humorous at all) for some topics.

Score Consistency

Next, we determined whether any annotator behaved differently from other annotators, by calculating the correlations between all combination pairs of annotators. Figure 2 shows the value of Kendall's tau, where a brighter color indicates a higher correlation. From this figure, we can see that the ratings by annotator 10 and 12 were the most different from those by other annotators; they can be removed for the purpose of consistency.

Categorization Consistency

As explained earlier, we set four types of humor (see Table 2) and asked annotators to categorize candidate answers into one of the four types. Ideally, humorous answers should have been categorized into the same type by all 13 annotators. However, the categorization result was contrary to this ideal result. As shown in Figure 3, most answers were categorized

Table 1: Summary of the candidate humor expressions

Category	Love 2004	Miscellaneous 2004	Love 2005	Miscellaneous 2005
Number of Candidate Answers	12	13	6	5
Number of Annotators	13	2	6	7

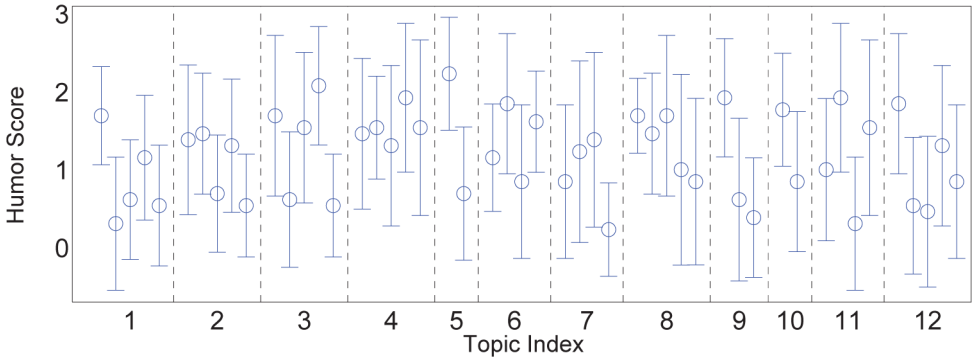

Figure 1: Distribution of humor ratings for twelve questions. Each question has multiple humorous answer candidates. The circles represent mean scores for each candidate, and the vertical bars represent standard deviations.

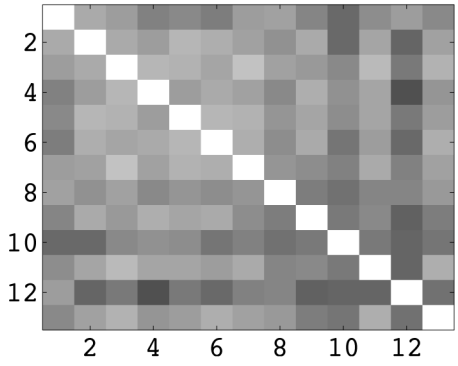

Figure 2: Correlation between ratings of annotators (Kendall's tau). Brighter squares indicate that the annotators represented on the x-axis gave a rating similar to that given by the annotator represented on the y-axis.

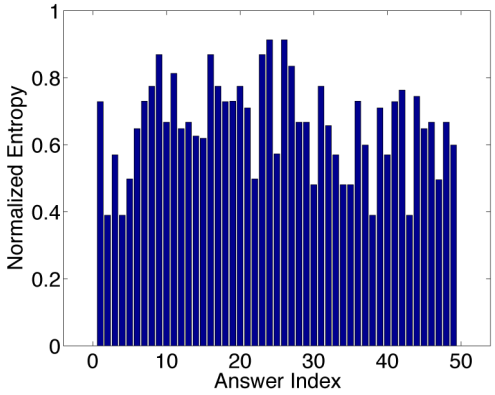

Figure 3: Entropy of assessment of humor types by annotators

differently by the annotators; the entropy was close to one. In the case of answers for which the best agreements were obtained, 10 out of 13 annotators assigned them to the same type. In the worst case, the answers were categorized into one of the four types as follows: 3 chose A, 1 chose B, 5 chose C, and 4 chose D.

Discussion

Our observation suggests that the variability in humor rating and labeling was not noise but was rather attributed to the differences in annotators' personalities. Similar perceptional differences can also be found in other domains. For example, couples often disagree about the status of their relationships (Busby and Holman 2009). The variability does not originate from noise resulting from erroneous labeling; nevertheless, the variability can be corrected in several possible ways (Wiebe, Bruce, and O'Hara 1999; Clemen and Winkler 1990; Klebanov, Beigman, and Diermeier 2008). Further, research efforts have also been directed toward accommodating label noise (Dawid and Skene 1979; Brodley and Friedl 1999; Lawrence and Schölkopf 2001). Our analysis suggests that there is a possibility to eliminate inconsistency in middle range (somewhat humorous or not so humorous) rating by focusing on the extremes. Also, we can separate a few non-standard annotators by calculating similarities between annotators even if they are consistent labellers. A big problem if that by applying further filtering, the number of annotations per item become smaller. Therefore, we need to investigate the applicability of existing and new methodologies to the problem of annotation reliability.

One of the potential approaches that have been explored to overcome the problem of variability in annotation is crowdsourcing. This approach entails outsourcing annotation work to anonymous annotators on the Internet at relatively lower costs. The categorization variability is expected to be stabilized by averaging. For example, crowdsourcing has been used to obtain the estimate of grammatical acceptabilities (Gibson, Piantadosi, and Fedorenko 2011) and also to estimate the quality of experience (QoE) for multimedia (Chen et al. 2009). Although crowdsourcing has some disadvantages, such as a nonmotivated annotator and data corruption by spammers, computational approaches are available for detecting and overcoming these disadvantages (Snow et al. 2008; Raykar et al. 2010).

Conclusions

In this study, we analyzed human judgement on humorous expressions extracted from a CQA corpus. For this purpose, we conducted a questionnaire-based evaluation using automatically collected candidates of humorous answers to questions in the corpus. Our results show that there is very little agreement between the assessments of humorous expressions by the annotators, in terms of both degree of humor and categorization of humor. This finding also suggests that we cannot use the average ratings of annotation as the gold standard for developing humor-related applications. We have to eliminate the cause of inconsistency or categorize different perception by different annotators. An interesting topic to explore is whether it is possible to model personality differences that affect humor perception. It is thus far unclear whether previous research findings on personality differences in the perception of humor (e.g., (Lamb 1968)) can be applied to online communication, where we can only guess the personality of users on the basis of the text they output rather than direct observation.

References

Attardo, S. 1994. *Linguistic Theories of Humor*. Mouton de Gruyter.

Brodley, C., and Friedl, M. 1999. Identifying mislabeled training data. In *Journal of Artificial Intelligent Research*, volume 11, 131–167.

Busby, D. M., and Holman, T. B. 2009. Perceived match or mismatch on the gottman conflict styles: Associations with relationship outcome variables. *Familyh Process* 48(4):531–545.

Chen, K.-T.; Wu, C.-C.; Chang, Y.-C.; ; and Lei, C.-L. 2009. A crowdsourceable qoe evaluation framework for multimedia content. In *MM*.

Clemen, R. T., and Winkler, R. L. 1990. Unanimity and compromise among probability forecasters. *Management Science* 36:767–779.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1):pp. 20–28.

Gibson, E.; Piantadosi, S.; and Fedorenko, K. 2011. Using mechanical turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass* 5(8):509–524.

Inoue, M., and Akagi, T. 2012. Collecting humorous expressions from a community-based question-answering-service corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Klebanov, B. B.; Beigman, E.; and Diermeier, D. 2008. Analyzing disagreements. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, 2–7.

Lamb, C. W. 1968. Personality correlates of humor enjoyment following motivational arousal. *Journal ofPersonality and Social Psychology* 9:237–241.

Lawrence, N. D., and Schölkopf, B. 2001. Estimating a kernel fisher discriminant in the presence of label noise. In *Int'l Conf. Machine Learning*, 306–313.

Martin, R. A., and Lefcourt, H. M. 1984. Situational humor response questionnaire: Quantitative measure of sense of humor. *Journal of Personality and Social Psychology* 47(1):145–155.

Mihalcea, R., and Pulman, S. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Mihalcea, R., and Strapparava, C. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Raykar, V. C.; adn Linda H. Zhao, S. Y.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *JMLR* 11:1297–1322.

Reyes, A.; Potthast, M.; Rosso, P.; and Stein, B. 2010. Evaluating humor features on web comments. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)*.

Reyes, A.; Rosso, P.; and Buscaldi, D. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering* 74:1–12.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263.

Wiebe, J. M.; Bruce, R. F.; and O'Hara, T. P. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 246–253.