

# Collecting Humorous Expressions from a Community-based Question-answering-service Corpus

Masashi Inoue, Toshiki Akagi

Yamagata University/National Institute of Informatics  
3-16, 4 Jyonan, Yonezawa-shi, Yamagata, Japan  
mi@yz.yamagata-u.ac.jp

## Abstract

We proposed a method of collecting humorous expressions from an online community-based question-answering (CQA) corpus where some users post a variety of questions and other users post relevant answers. Although the service is created for the purpose of knowledge exchange, there are users who enjoy posting humorous responses. Therefore, the corpus contains many interesting humour communication examples that might be useful in understanding the nature of online communications and variations in humour. Considering the size of 3, 116, 009 topics, it is necessary to introduce automation in the collection process. However, due to the context dependency of humour expressions, it is hard to collect them automatically by using keywords or key phrases. Our method uses natural language processing based on dissimilarity criteria between answer texts. By using this method, we can collect humour expressions more efficiently than by manual exploration: 30 times more examples per hour.

**Keywords:** humor, CQA, collection

## 1. Introduction

Humour plays an important role in human communication, and its linguistic and psychological frameworks have been researched (Attardo, 1994). However, because of its volatility, humorous expression has been extensively subjected to few linguistic analyses - especially corpus-based ones. Most humour is observed in casual conversations that are not recorded for analysis. Therefore, the analysis of humorous expression has been limited to the scripted expressions found, for example, in TV sitcoms, poems, novels (Attardo, 2001), and advertisements (Duncan, 1979)(Sternthal and Craig, 1973). This situation has changed nowadays as the Internet is increasingly being used as a means of casual communication. Online textual communication is becoming richer and richer, and not only is its quantity increasing but also its diversity. Many people help others who are in need of information, either by using their real names or anonymously. An interesting aspect of this type of online communication, community-based question answering (CQA), is that people do not always use the service to obtain just the information. Some users use these systems to enjoy the communication itself. We can collect examples of such interactions which include humour expressions. This paper focuses on humorous responses to questions in a CQA service and proposes a simple method of automatically collecting such examples. The experimental results show that the method can be used to collect examples far more efficiently than by manual searching.

## 2. Computational Humour Collection

Using online text, there are computational approaches in collecting and understanding expressions. For the efficient collection of expressions, resource selection is important. For example, colloquial expressions have been collected from an online bulletin board(Inoue et al., 2011). Regarding humour expressions, 16,000 one-liners, and self-contained short jokes have been collected by using boot-

strapping on ten seed expressions (Mihalcea and Strapparava, 2005). The resources were web pages whose URLs contained any of six keywords that are related to humour. Similarly, humorous text have been collected from the news web site, Onion, where all articles are assumed to be humorous(Mihalcea and Pulman, 2007)

In contrast to these previous works, we are interested in the situation where humorous text is buried in the midst of much non-humorous text. Humour also depends on the preceding text to a great extent. Such situations can be found in online bulletin boards such as Slashdot (Reyes et al., 2010). Each news topic there has numerous comments and the comments are marked with tags including "funny". By selecting comments with the "funny" tag, 159,153 humour text were extracted. In this site, users are expected to state something humorous, as there is a tag for the purpose of feedback. In the CQA corpus we use in this study, users are expected to provide informative answers to the question and humorous answers are considered noises. Therefore, we have to extract humour expression without explicit cues provided by the systems.

## 3. Community-based Question Answering Services

Before online communication emerged, question-answering services were provided as one-to-one interactions between knowledgeable experts and the common people. With the asynchronous nature of online written communication, many people can now answer the same question at the same time. The first commercial service was started in 2000 by OKWave in Japan. Similar services followed, such as Jinriki Kensaku Hatena in 2001 and Yahoo! Chiebukuro in 2004. Then, Yahoo! Answers started in 2005 and became a global CQA service with 200 million users worldwide. As a language resource, CQA corpus can be directly used to assist building a QA system that mimics human answers (Momtazi and Klakow,

2010). We use the CQA corpus as the source of particular expressions.

## 4. Target Corpus

The target corpus was built with data taken from the Yahoo! Chiebukuro CQA service mentioned in the previous section. The corpus was distributed by NII-IDR for research purposes<sup>1</sup>. The data was collected from April 2004 to October 2005 and its size was 4.1 gigabyte (Yahoo! Chiebukuro data, first edition). All text used in our experiments was in Japanese. The corpus statistics are summarised in Table 4.. For each topic there was a question (Q) posted by a user and several answers to the question by other users. One of the answers was considered to be chosen as the 'best answer' (BA) by the question poster, and the user who provided the answer was rewarded by being given points. Other answers are called 'normal answers' (NA). There was an attempt to predict if there will be a BA for a particular question (Liu et al., 2008). Although it seems that BAs are associated with the seriousness of an answer, we found that the BAs are not necessarily characterised as the opposite of humorous answers. The question is posted with category information. The number of categories changes in different months; therefore, we considered common categories throughout the corpus.

## 5. Humorous Answers

For a question, there are often multiple answers in the CQA. Examples of serious and humorous answers extracted from the service are as follows<sup>2</sup>:

**Question:** What is 'reason'?

**Serious Answer:** In dictionary terms, it means  
1) a psychological function for judgement that is based on reasoning or an ability to think logically and conceptually; 2) ···

**Humorous Answer:** That's what you don't have!

**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...

If you want to play games, minimum is 512M. I recommend 1G or more... Other than games, your PC may freeze by doing simple encoding... Add real memories! Otherwise, you'll see your PC frozen every time you run memory hungry tasks.

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

**Question:** I have to write a report and do not know where to start. This is my first time and I'm confused.

**Serious Answer:** First, you compose a rough story. Then, make the table of contents. In the beginning, ···

**Humorous Answer:** Stop playing with this Chiebukuro service to begin with.

By observing these examples, we see that there are not word-level or phrase-level similarities among humorous answers for different questions. This nature of humour makes the collection of expressions difficult.

## 6. Method used to Collect Humorous Text

### 6.1. Manual and Local Feature-based Collection

There are two approaches to finding humorous text automatically. The first is the micro approach, in which particular words or phrases that are often found in humorous text are assigned high scores and text containing many high-scoring words or phrases is retrieved as humorous text. The second is the macro approach, in which a narrative characteristic of text is believed to generate the humour the text. Our initial observation suggested that the macro approach was more reasonable because humour is not derived from a word or a phrase but the context. For purposes of comparison, we collected 2,030 humorous answers from the corpus manually. By using these separated humorous and non-humorous text, we calculated  $tf \cdot idf$  scores of the vocabulary. We observed that keywords having the highest scores do not necessarily correspond to being humorous and so we did not continue with this type of keyword-based micro approach.

### 6.2. Dissimilarity-based Collection

The procedure we followed to extract humorous answers from the QA corpus is summarised in Figure 1. Candidate answers were ranked by calculating degree of humour. As features for macro approach, we employed length of answers and dissimilarity between answer texts. Our observation suggests that serious answers tend to be long. In contrast, humorous answers are often short but not all short answers are humorous. Therefore, we based our scoring method on the dissimilarity between the longest answer ( $\mathbf{la}$ ) assumed to be a typical serious answer and the other answers. This approach is rationalised by the intuition that keywords or phrases are not good clues to finding humorous text in the CQA corpus. For answers to be humorous, they should not be expected by the questioner. Therefore, there are some factors of surprise and the vocabulary or expression may be different from the topic. Since we observed that character-based features are more robust than word-based ones by accommodating expression deviations, we use character-based n-grams. We also found that unigram and bigram were too short to bear a semantic chunk; thus, we used trigram. The similarity is calculated as follows:

$$\text{sim} = 1 - \frac{1}{N_{\mathbf{la}}} \sum_{k=1}^{N_{\mathbf{la}}} \left\{ \frac{\text{freq}_{\mathbf{la}}(k) - \text{freq}_{\mathbf{sa}}(k)}{\text{freq}_{\mathbf{la}}(k) + \text{freq}_{\mathbf{sa}}(k)} \right\}^2 \quad (1)$$

<sup>1</sup><http://www.nii.ac.jp/cscenter/idr/en/yahoo/yahoo.html>

<sup>2</sup>They are accessible by the topic IDs q1346901713, q125073382 and q121864282 on the service web site.

Table 1: Summary of the Yahoo Chiebukuro CQA corpus

Questions	All Answers	Normal Answers	Categories
3,116,009	16,593,794	10,361,777	139

1. For each question, an answer set (**A**) is stored.
2. Lengths of all answers in **A** are calculated and the longest answer is set as a serious answer (**la**).
3. Character-based trigrams are calculated for **A**.
4. Similarity between short answers **sa** and **la** are calculated using Equation 1.
5. The answer with the lowest similarity score is presented to a human evaluator as a candidate of a humorous text example.

Figure 1: Humour scoring procedure.

where  $N_{la}$  is the number of trigram patterns in **la**, the term  $\text{freq}_{la}(k)$  is the frequency of the trigram pattern  $k$  in **la** and  $\text{freq}_{sa}(k)$  is that in **sa**. This similarity is based on the dissimilarity measure using the number of shared n-grams (Kešelj et al., 2003). This measure focuses on the difference of text style such as authorship rather than topical differences for which other measures such as cosine similarity may be used. Our choice is based on the assumption that humour in answers is related to a poster’s writing style rather than to the topic of the post.

Many of the serious answers contain similar information. Such redundancy may impose certain cognitive load to the users and there is a need for summarisation (Liu, 2008). This fact motivates us to save the most dissimilar answer to the typical serious answer for manual checking.

### 6.3. Corpus Dependent Parameters

For the purpose of enhancing our algorithm, we conducted a preliminary experiment and decided to use the following values.

- Questions whose answer set size  $|A| \geq 2$  are considered.
- There exists a long answer whose length  $l \geq 150$  characters.
- The degree of similarity  $0.001 \leq s \leq 0.05$ .

The first condition is necessary to set dissimilarities among answers. The lower bound in the third condition is set to eliminate the effects of spam or erroneous answer posts that contain only meaningless characters and result in zero dissimilarity values.

## 7. Experimental Results

### 7.1. Collection Effectiveness

Before processing the entire collection, we examined the effectiveness of our procedure by using a sub-corpus of ‘Love and human relationship’ category that contains a relatively larger number of humour expressions. The procedure 1 to 4 of Figure 1 was conducted and in the last step, instead of presenting the most dissimilar answer as the candidate, we showed the two most dissimilar ones for manual validation. Among 600 topics, there were 18 humour answers. Fifteen of them were the most dissimilar answers and three were in the second dissimilar ones. There was not any topic that contained more than one humorous answer. From this result, we assume that it is efficient to present only the most dissimilar candidate to the human validator, even though we may miss some humour expressions.

### 7.2. Collection Efficiency

The time spent for collecting humorous expressions from the corpus is summarised in Table 7.2.. From 19 months of data, we used two months because the entire corpus is too big to be analysed. December 2004 and July 2005 were considered to avoid seasonal effects. The manual collection procedure took 134 hours to investigate all the answers for randomly chosen 2,030 questions and 47 humour answers were found. The automatic procedure identified candidates for the humorous answers and manual checking followed. The computation (CPU) and manual investigation took 138.9 hours to find 267 humour answers from 23,168 topics. Manual checking of humorous answer candidate alone took 28.9 hours. That is, in terms of efficiency, by using our automatic method, we can collect 30 times more humorous expressions for the same human work hours. The automatic method could be enhanced by improved implementation or by using more powerful computers.

### 7.3. Validation and Categorisation

We evaluated the automatically collected candidate humour responses in terms of degree of humour by asking 34 volunteers to rate them by allocating scores from zero to three. They were also asked to categorise the questions into one of nine types and the responses into one of four humour types. The results showed that our simple filtering by selecting the response with the highest dissimilarity score was a safe decision that kept humorous responses in the pool. For example, for a topic, the answer with highest dissimilarity score also has the highest humour score, 1.42 while other answers have considerably lower scores: 0.14, 0.43 and 0.14. Regarding question type, we found inconsistency in question types selection among volunteers.

Table 2: Time spent for the collection humorous expressions.

	Items Found	Items Searched	Time Spent	Time per Item
Manual	47	2030	134 hours	0.35/hour
Automatic with CPU	267	23168	138.9 hours	1.92/hour
Automatic without CPU	267	23168	28.9 hours	9.24/hour

## 8. Conclusions

We proposed a method of collecting humorous expressions from a CQA corpus. The method uses natural language processing based on dissimilarity criteria between texts. By using our method, we can collect humour expressions more efficiently than by manual exploration. The disadvantage of our method is that we may miss some humour expressions by taking only the answer with highest humour score. Although it is difficult to actually calculate recall rate, we may estimate the trade-off between efficiency and accuracy in collection humour expressions. The next step will be the analysis of the collected expressions to see whether all varieties of humour are observed or whether some types are missing. An interesting question to study is whether we can collect different kinds of humour expressions using different similarity measures by the same method or whether we have to use a completely different procedure.

We address the problem of collecting humours in CQA corpus. Expressions that evoke certain responses in the readers are identified as stimuli. The other approach may utilise the result of humour which is laughing. In several spoken dialogue corpora, there is laughing included. Laughing can be detected automatically by signal processing (Khiet and Truong, 2007). We can assume that there is a humorous expression just before the laughing starts. However, it should be noted that laughing is not always caused by humour expressions (Partington, 2006). Therefore, the accuracy of such an approach should be investigated especially when the corpus is not speech but a textual one where obvious laughing does not appear frequently. In the case of text, we can use slang and emotiocons such as “LOL” or “:-)” in English (Reyes et al., 2010), or “www” in Japanese text as well as an explicit amused expression.

## 9. Acknowledgements

We thank for Moeko Okada and anonymous reviewers for suggesting references. This research was partially supported by the Grant-in-Aid for Scientific Research 21500266.

## 10. References

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

S. Attardo. 2001. *Humorous texts: a semantic and pragmatic analysis*. Walter de Gruyter.

C. P. Duncan. 1979. Humor in advertising: a behavioral perspective. *Journal of the Academy of Marketing Science*, 7:285–306.

M. Inoue, T. Matsuda, and S. Yokoyama. 2011. Web resource selection for dialogue system generating natural

responses. In *HCI International 2011 – Posters’ Extended Abstracts*, volume 173, pages 571–575.

V. Kešelj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram based author profiles for authorship attribution. In *Pacific Association for Computational Linguistics*, pages 255–264, Halifax, Canada.

P. Khiet and D. A. Truong. 2007. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158.

Y. Liu, J. Bian, and E. Agichtein. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’08, pages 483–490.

Y. L.-Y. Liu. 2008. Understanding and summarizing answers in community-based question answering services. In *The 22nd International Conference on Computational Linguistics*, pages 497–504.

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

R. Mihalcea and C. Strapparava. 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)*, Vancouver, October.

S. Momtazi and D. Klakow. 2010. Yahoo! answers for sentence retrieval in question answering. In *The LREC Workshop on Web Logs and Question Answering*, Valletta, Malta.

A. Partington. 2006. *The Linguistics of Laughter*. Routledge.

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

B. Sternthal and C. S. Craig. 1973. Humor in advertising. *Journal of Advertising*, 37:12–18.