# ASURA: A Best-Answer Estimation System for NTCIR-8 CQA Pilot Task

Daisuke Ishikawa[†]
[†]National Institute of Informatics
dais@nii.ac.jp

## ABSTRACT

This paper describes a best answer estimation system called ASURA. The features of ASURA were decided on the basis of the results of experiments that were conducted to determine how people estimated which of a given set of answers was the best. There are two ASURA models, ASURA-1, which has 5 features, and ASURA-2, that has thirteen features. We outline ASURA-1 and ASURA-2 in this paper, and we also report the results from the NTCIR-8 CQA pilot task.

**Keywords:** Yahoo! Chiebukuro, Community QA, Best Answer Estimation, SVM

## 1. INTRODUCTION

There has recently been an increase in the amount of research on community Q&A sites such as "Yahoo! Chiebukuro"[3] and "Oshiete! Goo"[4], etc.

These Q&A sites have frameworks for selecting the most appropriate answers. For example, in Yahoo! Chiebukuro, users can select the "Best Answer", which represents the most convincing and Satisfying answer in their opinion.

We wanted to find out whether or not it is possible for a computer to select the Best Answer (the most appropriate answer)[5]. So, we used ASURA, a best answer estimation system to do this. The features of ASURA were decided on the basis of the results of experiments conducted to determine how people estimate which of a given set of answers was the best. There are two ASURA models, ASURA-1, which has 5 features, and ASURA-2, which has thirteen features. We outline ASURA-1 and ASURA-2, and present the results from the NTCIR-8 CQA pilot task.

The structure of this paper is as follows. The outline and operating environment of ASURA are described in Section 2. The five-feature model, ASURA-1, is described in Section 3. The 13-feature model, ASURA-2, is described in Section 4, and the official result for ASURA in the NTCIR-8 CQA pilot task is given in Section 5. The paper concludes in Section 6.

## 2. OUTLINE OF ASURA

First, ASURA is a model used to analyze the factors used by a person when trying to select the best answer. The model was designed on the basis of the factors obtained from the results of experiments into how a person estimated the best answer. The ASURA-1 model is a simple best answer estimation system that takes into consideration only answers, and it has five features. The ASURA-2 model is a best answer estimation system that takes into consideration the compatibility of the question and the answer in addition to that from ASURA-1, and it has thirteen features. The details for each model are described in Sections 3 and 4.

The operating environment of ASURA is as follows.

- Learning environment:
  - Machine learning: SVM (TinySVM 0.09[7])
  - Solver Type: C-SVM (default)
  - Kernel: linear (default)
  - Training data: 300 questions extracted from each category at random at the same rate as the test data
    - Questions: 300 items
    - Answers: 600 items (300 best answers, 300 normal answers)
  - Test data: 1500 questions (official test collection)
    - Questions: 1500 items
    - Answers: 7443 items (1500 best answers, 5943 normal answers)
  - Classification: Binary classification (Positive data is a best answer and negative data is a normal answer.)
- Computing environment:
  - OS: CentOS 5.3 (x86_64, 64-bit)
  - CPU: Xeon 2.0 GHz Quad Core
  - Memory: 16 GB
  - Disk Array: 1 TB × 12 (RAID 6.0, 4 Gbps FC)

## 3. ASURA-1: FIVE FEATURES MODEL

The ASURA-1 model is a simple best answer estimation model designed based on the factors that a person uses to select the best answer.

In our previous work[5], we used two categories for best answer estimation experimentation: "Consultations on love, and problems with personal relationships" and "Personal computers and peripheral devices." The categories "Consultations on love, and problems with personal relationships" and "Personal computers and peripheral devices" were categorized into the question types for "Social surveys" and "Information searches", respectively [6].

*NTCIR-8 Workshop Meeting, 2010,* Tokyo, Japan.

From each of the categories, we extracted 50 random questions that received between two and four answers. The answers appended to the 50 extracted questions were shuffled, and the question/answer sets were then constructed. These 50 questions/answer sets were then distributed to the assessors. These two assessors are referred to as author 1 and author 2, and they are researchers in the field of Informatics. These assessors were asked to select one answer as the "Best Answer" from among those given (2 - 4 answers).

We found several factors when selecting the best answer from the results of the best answer estimation experiment by human assessors in our previous work. ASURA-1 is a model that consists of three factors: detailed, evidence, and polite, from among the factors found. The descriptions of these factors is as follows.

- <Detailed> Explanations were given in detail: Detailed explanations that were difficult to understand were not considered to be Best Answers. In some cases, simple responses were selected as Best Answers.

- <Evidence> Answers with sources (URLs, etc.) included as evidence for information: This includes answers based on personal experience. (We used the following order of priorities: personal experience > others' experiences > no experience, conjecture)

- <Polite> Answers written in polite Japanese: In many cases, the answers selected as Best Answers were written in polite Japanese.

These three factors can be calculated through simplification as follows.

- <Details>: Count the number of characters in the answer, and compare with the average number of characters in the answer group.

- <Evidence>: Check whether the character string 'http' or 体験 or 経験 is included in the answer [1]

- <Polite>: Count the number of appearances of です (the sentence copula 'desu') and ます ('masu', the polite form of a verb in the answer, and compare with the number of the average appearances of 'desu'(です) and 'masu'(ます) of the answer group.

The ASURA-1 model is composed of the following five features based on the calculation listed in Table 1.

The learning results from ASURA-1 that consist of these five features are described below. The training data and test data of ASURA-1 were generated based on these five features. The training data used 300 questions extracted from each category at random [by/at?] the same rate as the test data. An example of this training data is shown in Figure1. The test data used 1500 questions distributed from the CQA pilot task. The machine learning software, TinySVM[7] was used to learn these training data, and the performance was measured by using the test data. The performance of ASURA-1 is shown on the left in the table2 as a result of the learning.

```
-1 1:293 2:2 3:0 4:319.5 5:2 # Q1-A1 NA
1 1:346 2:2 3:0 4:319.5 5:2 # Q1-A2 BA
1 1:244 2:2 3:0 4:242.5 5:1.5 # Q2-A1 BA
-1 1:241 2:1 3:0 4:242.5 5:1.5 # Q2-A2 NA
-1 1:253 2:0 3:0 4:240.5 5:1 # Q3-A1 NA
1 1:228 2:2 3:0 4:240.5 5:1 # Q3-A2 BA
-1 1:235 2:1 3:0 4:348 5:1 # Q4-A1 NA
1 1:461 2:1 3:0 4:348 5:1 # Q4-A2 BA
1 1:226 2:0 3:1 4:241 5:0.5 # Q5-A1 BA
-1 1:256 2:1 3:1 4:241 5:0.5 # Q5-A2 NA
```

**Figure 1: Training data example of ASURA-1**

## 4. ASURA-2: 13 FEATURES MODEL

ASURA-2 is a model that refers to the question and the answer while ASURA-1 is one that refers only to the answer. The ASURA-2 model calculates < details > more strictly on the basis of the number of keywords. Additionally, ASURA-2 has added features based on the compatibility of the question and the answer.

Moreover, this model has added category information to the features.

The 13 features of ASURA-2 are shown below in Table 3.

The learning results from ASURA-2 consisting of these 13 features are Described below. The training data and the test data of ASURA-2 were generated based on these 13 features. The training data used 300 questions extracted from each category at random at the same rate as the test data. Figure 2 shows an example of these training data. The test data used 1500 questions distributed from the CQA pilot task. The TinySVM for machine learning was learned by using these training data, and the performance was measured by using these test data. The performance of ASURA-2 is shown on the right in table 2 as a result of the learning.

## 5. OFFICIAL RESULTS

This section describes the official results[2] from ASURA.

Table 4 shows the results of ASURA-1, ASURA-2, BASELINE-1(sort by random), BASELINE-2(sort by length of answer) and BASELINE-3(sort by time-stamp of answer).

In any evaluation, ASURA-2 exceeds ASURA-1. In GA-nDCG and GA-Q, the performance of ASURA-2 is good. However, the performance of BASELINE-2 is good in BA-Hit@1 and GA-nG@1. In GA-Hit@1, the performance of ASURA-2 and BASELINE-2 is the same.

Next, the graph of the evaluation results for each category is shown in Figure 3. In all the graphs except GA-Hit@1, the performance of ASURA-1, ASURA-2, and BASELINE-2 is higher than that of BASELINE-1 and BASELINE-3. In GA-nDCG and GA-Q, the performance of ASURA-2 is better than that of ASURA-1 for most categories. Moreover, the performance of ASURA-2 is higher than that of BASELINE-2 in category 5 (internet) and 6 (sports) in GA-nDCG and GA-Q. However, the performance of BASELINE-2 is higher than that of ASURA-2 in category 10 (news) and 11 (travel) in GA-nDCG and GA-Q. As for ASURA-2, the categories that demonstrate good and bad performances were observed while the performance of BASELINE-2 is comparatively steady in every category.

## 6. CONCLUSION

We described ASURA, a system for estimating the best answer in the CQA pilot task.

ASURA-1 is a five-feature model based on the factors of the

[1] 経験, pronounced as 'keiken', is the Japanese word for 'experience in general', and 体験, 'taiken', means something like a single experience of something

**Table 1: 5 features of ASURA-1**

| | |
|---|---|
| Feature 1 | Number of characters in answer. |
| Feature 2 | Number of appearances of 'desu', です and 'masu', ます. |
| Feature 3 | Existence of character string of 'http' or 'keiken', 経験, or 'taiken', 体験.<br>(exist = 1, not exist = 0) |
| Feature 4 | Average number of characters in answer group. |
| Feature 5 | Average number of appearances of 'desu', です and 'masu', ます in answer group. |

**Table 2: Learning results of training data from 300 questions**

| ASURA-1: five features model | |
|---|---|
| Accuracy | 67.76837% (5044/7443) |
| Precision | 34.70568% (1020/2939) |
| Recall | 68.00000% (1020/1500) |
| System/Answer | p/p p/n n/p n/n: 1020 1919 480 4024 |
| Elapsed time | 7min 43.11sec (99%CPU) |

| ASURA-2: 13 features model | |
|---|---|
| Accuracy | 67.76837% (5044/7443) |
| Precision | 34.78849% (1028/2955) |
| Recall | 68.53333% (1028/1500) |
| System/Answer | p/p p/n n/p n/n: 1028 1927 472 4016 |
| Elapsed time | 17min 55.44sec (99%CPU) |

'best answers' selected by a human. ASURA-2 is a 13-feature model that has features based on the compatibility of the question and the answer to the ASURA-1 model.

From the official results we found that ASURA-2 exceeds ASURA-1 in every case. In GA-nDCG and GA-Q, the performance of ASURA-2 is higher than that of BASELINE-2 while the performance of BASELINE-2 is higher than that of ASURA-2 in BA-Hit@1 and GA-nG@1.

Our future work is to further analyze the categories that did not perform well and to verify the effectiveness of each feature of the proposed model.

## 7. ACKNOWLEDGMENTS

'Yahoo! Chiebukuro Data' provided to National Institute of Informatics by Yahoo Japan Corporation was used in the implementation of this research.

## 8. REFERENCES

[1] Ishikawa, D., Sakai, T. and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task(Part I): The Test Collection and the Task, *NTCIR-8 Proceedings*, 2010. (to appear)

[2] Sakai, T., Ishikawa, D., and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task(Part II): System Evaluation, *NTCIR-8 Proceedings*, 2010. (to appear)

[3] Yahoo!JAPAN: Yahoo! Chiebukuro, http://chiebukuro.yahoo.co.jp/

[4] OKWave: Osiete! goo, http://oshiete.goo.ne.jp/

[5] Ishikawa, D., Kuriyama, K., Seki, Y., and Kando, N.: Investigation of Possibility of Best-Answer Estimation in Q&A Site (in Japanese), *IPSJ SIG Technical Reports*, 2010-FI-97, 2010.

[6] Kazuko Kuriyama, Noriko Kando: Analysis of Questions and Answers in Q&A Site, *IPSJ SIG Technical Reports* 2009-FI-95, 2009.

[7] TinySVM: http://chasen.org/ taku/software/TinySVM/

**Table 3: 13 features of ASURA-2**

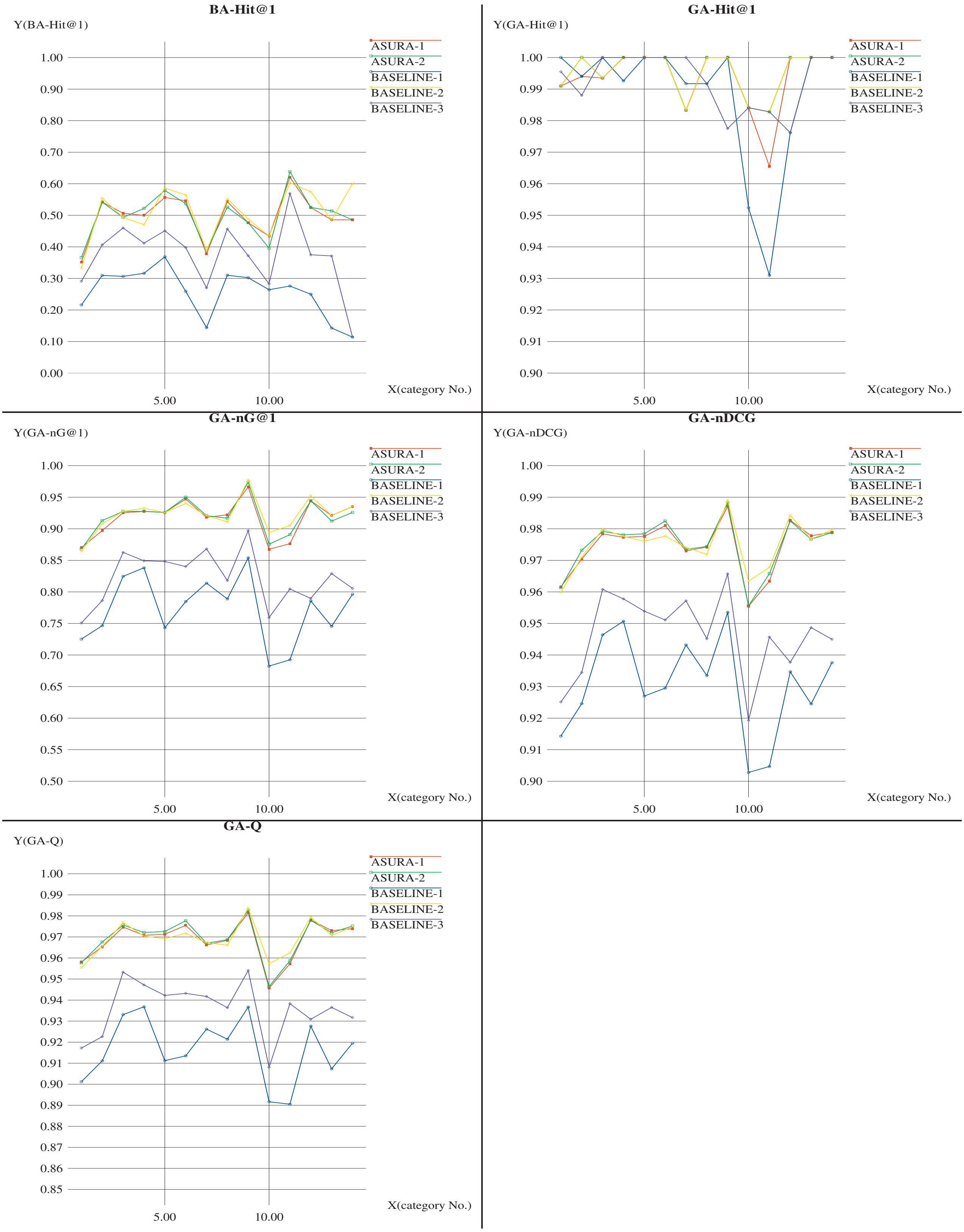
| | |
|---|---|
| Feature 1 ∼ 5 | is the same as for ASURA-1. |
| Feature 6 | Kind number of keywords based on hiragana in answer. |
| Feature 7 | Kind number of keywords based on non-hiragana in answer. |
| Feature 8 | Kind number of keywords based on hiragana in question. |
| Feature 9 | Kind number of keywords based on non-hiragana in question. |
| Feature 10 | Number of characters in question. |
| Feature 11 | Number of agreements of keywords based on hiragana in question and answer. |
| Feature 12 | Number of agreements of keywords based on non-hiragana in question and answer. |
| Feature 13 | Category number (1-14) in question is acquired, 100 is added to category number, and number is set to 1 as feature. (For example, feature:value is 101:1 in case of category=yahoo, category number=1) |

```
-1 1:177 2:2 3:0 4:203.5 5:2 6:9 7:10 8:11 9:11 10:246 11:1 12:3 101:1 # Q1-A1 NA
1 1:230 2:2 3:0 4:203.5 5:2 6:9 7:12 8:11 9:11 10:246 11:1 12:0 101:1 # Q1-A2 BA
1 1:127 2:2 3:0 4:126 5:1.5 6:6 7:10 8:4 9:5 10:99 11:0 12:1 101:1 # Q2-A1 BA
-1 1:125 2:1 3:0 4:126 5:1.5 6:4 7:5 8:4 9:5 10:99 11:0 12:1 101:1 # Q2-A2 NA
-1 1:138 2:0 3:0 4:124.5 5:1 6:8 7:9 8:14 9:18 10:319 11:1 12:2 101:1 # Q3-A1 NA
1 1:111 2:2 3:0 4:124.5 5:1 6:6 7:7 8:14 9:18 10:319 11:0 12:1 101:1 # Q3-A2 BA
-1 1:118 2:1 3:0 4:231 5:1 6:8 7:9 8:17 9:18 10:428 11:2 12:1 101:1 # Q4-A1 NA
1 1:344 2:1 3:0 4:231 5:1 6:14 7:13 8:17 9:18 10:428 11:2 12:3 101:1 # Q4-A2 BA
1 1:109 2:0 3:1 4:124 5:0.5 6:7 7:9 8:9 9:11 10:187 11:2 12:3 101:1 # Q5-A1 BA
-1 1:139 2:1 3:1 4:124 5:0.5 6:2 7:4 8:9 9:11 10:187 11:1 12:0 101:1 # Q5-A2 NA
```

**Figure 2: Training data example of ASURA-2**

**Table 4: Official Results for ASURA and BASELINE**

| team name | BA-Hit@1 | GA-Hit@1 | GA-nG@1 | GA-nDCG | GA-Q |
|---|---|---|---|---|---|
| ASURA-1 (5 features) | 0.4813 | 0.9940 | 0.9140 | 0.9734 | 0.9680 |
| ASURA-2 (13 features) | 0.4840 | 0.9953 | 0.9166 | 0.9742 | 0.9689 |
| BASELINE-1 (random) | 0.2713 | 0.9920 | 0.7751 | 0.9311 | 0.9169 |
| BASELINE-2 (answer length) | 0.4847 | 0.9953 | 0.9170 | 0.9735 | 0.9680 |
| BASELINE-3 (timestamp) | 0.3820 | 0.9940 | 0.8213 | 0.9460 | 0.9359 |



(Category No. = Category Label):
1= yahoo, 2= entertainment, 3= health, 4= lifeguide, 5= internet, 6= sports, 7= love,
8= education, 9= school, 10= news, 11= travel, 12= business, 13= career, 14= manners

**Figure 3: Per-category Evaluation for ASURA and BASELINE**