IPSJ Transactions on Databases Vol. 4 No. 2 1–9 (July 2011)

Recommended Paper

Community QA Question Classification: Is the Asker Looking for Subjective Answers or Not?

Naoyoshi Aikawa,[†1,†2] Tetsuya Sakai[†1] and Hayato Yamana[†2]

Community question answering (CQA) sites such as Yahoo! Chiebukuro are known to be very useful resources for automatic question answering (QA) systems. However, CQA users often post questions expecting not some general truths but rather opinions of different people. We believe that a QA system should act according to these different question types. We therefore define two question types based on whether the questioner expects *subjective* or *objective* answers, and report on an automatic question classification experiment. We achieve over 80% weighted accuracy using uni-gram and bi-gram features learned by Naïve Bayes with smoothing. We also discuss the inter-annotator agreement and its impact on automatic classification accuracy, as well as what kind of questions tend to be misclassified.

1. Introduction

Providing an appropriate answer in response to the user's question is one of the practical challenges in natural language processing and information retrieval. Some automated question answering (QA) systems that use the Web as the knowledge base are already useful to some extent [1),2)]. However, a highly effective QA system needs to (a) understand the user's intent given the question; (b) identify documents that contain the correct answer(s) to the question; and (c) extract or generate an appropriate answer string. None of these subproblems is trivial.

Utilizing community QA (CQA) data is one promising approach for solving the difficulties in automated QA. CQA sites such as Yahoo! Chiebukuro[⋆1] (Japanese Yahoo! Answers) and Oshiete! Goo[⋆2] provide a mechanism for people to post questions, post answers to these questions, give feedback to the posted items, and share all of the data with the world. Because the data from a CQA site are already well-structured (each record consists of a question and a set of answers), utilizing the CQA data for automated QA, for example, by searching the CQA data for an existing question that is similar to a newly posted question [3),4)], may be more effective and efficient than looking for answers across the entire Web.

However, even a quick look at existing CQA data would strongly suggest that completely automated QA may not always satisfy the user. It appears that some askers are *not* looking for computer-generated answers, however precise they may be. More specifically, there are questions that expect different personal opinions as answers rather than one "correct" answer. Moreover, there are questions that initiate a conversation between users. In such cases, returning a computer-generated answer may in fact hurt user satisfaction.

From the above viewpoint, this study considers the problem of classifying newly posted questions into two classes: those that expect *subjective* answers, and those that expect *objective* answers. Our ultimate goal is to build a QA system that operates in two modes: (1) Given a question that expects subjective answers ("subjective question"), contact a selected set of users and prompt them to post an answer; and (2) Given a question that expects objective answers ("objective question"), trigger an automatic QA system (e.g., search the entire Web).

In our problem setting, we assume that answer information and user information are unavailable at the time of question classification. This is in contrast to some existing studies [5),6)]. We believe that this setting is practical because we would like to build a system that can classify a question as soon as it is posted to the system, even if the asker is anonymous or a newcomer. Additionally, we note that it is usually not difficult for human judges to determine whether a question is subjective or not by just looking at the question text itself.

In this study, we use the NTCIR-8 Community QA Pilot Task question set [7),8)], which is part of the Yahoo! Chiebukuro data. When we evaluate classification result, we used weighted accuracy which can reflect the confidence of annota-

†1 Microsoft Research Asia
(This work was done while the first author was an intern at Microsoft Research Asia.)
†2 Waseda University
⋆1 Yahoo! Chiebukuro (http://chiebukuro.yahoo.co.jp/)

⋆2 Oshiete! Goo (http://oshiete.goo.ne.jp/)

tions. We tried SVM and Naïve Bayes learning machines with the combination of cue words, n-gram, dependency, and maximal repeats features. As a result, we found Naïve Bayes with n-gram or maximal repeats features achieves approximately 80% weighted accuracy. We also analyzed what kind of questions were misclassified by our best classifiers.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 defines our question classification criteria (subjective vs. objective), and Section 4 describes classification methods we explored. Section 5 describes our experimental settings, and Section 6 reports our experimental results and analyzes misclassified data. Finally, Section 7 concludes this paper.

2. Related Work

There are several existing studies that tackled the problem of question classification for CQA.

Kim, et al. [9] examined the criteria that question askers use for selecting the best answer by manually classifying 465 question-answer pairs and its comments in Yahoo! Answers. They classified questions into four types: Information (finding specific facts or understanding phenomena), Suggestion (seeking advice, recommendations, or viable solutions), Opinion (surveying other people's thoughts or tastes, or initiating discussions about social issues), and Others. After the classification, the distribution of questions over question types was as follows: Information-type 35%; Suggestion-type 25% and Opinion-type 39%. They found that the criteria of selecting the best answer differ across question types. Additionally, they pointed out that the socio-emotional factor plays an important role in selecting the best answer especially for Opinion-type questions.

Kuriyama and Kando [10] analyzed questions and answers in Yahoo! Chiebukuro. They manually classified 500 questions into three major types: Information-Search-type, Social-Research-type, Non-Question-type. Information-Search-type includes fact, authenticity, definition/description, method/means, cause/reason and effect/result. Social-Research-type includes advice, opinion, preference, recommendation and experience. Non-Question-type includes assertion and incomprehension. As a result of their analysis, they found question type distributions are quite different across question categories as defined in Yahoo! Chiebukuro. They also argue that presenting each question with not only its category but also its question type may enhance the usability of Yahoo! Chiebukuro.

Adamic, et al. [11] tried to capture user behavior and category characteristics. They used question clustering instead of classification. The number of clusters was set to three, and the resultant clusters were similar to the aforementioned categories as defined by Kim, et al. [9] Based on their findings, they predicted what kind of answerers are likely to be awarded the best answer. For example, for factual questions, answerers who focus narrowly on a specific topic tend to receive high ratings.

Rather than classifying questions directly, Liu, et al. [12] considered whether their answers are reusable or not. With their question data, which included 56–83% open/opinion questions, 78% of the best answers were reusable. However, more than 52% of them have non-unique best answers. This suggests that prompting users to provide answers is a safer strategy than automatic answering even for those open/opinion questions that are reusable.

All the works mentioned above mainly focused on mining some characteristics of questions, answers or user behavior in CQA. In contrast, our work was motivated by a concrete CQA application, namely, a system that triggers automatic QA for objective questions and prompts human answerers for subjective questions. We devised our question classification criteria specifically for this practical application and provided them to the annotators.

Harper, et al. [6] classified questions whether they are informational or conversational. Their main finding was that classification accuracy can be improved significantly when user information is combined with question text and category information. In their work, user information includes the number of people the user interacted with and the ratio of question posts and answer posts.

Li, et al. [5] tried to predict question subjectivity orientation, which is a problem similar to our question classification task. However, unlike our problem setting, they assumed the availability of answer data. To utilize unlabeled data, they showed the effectiveness of building two classifiers, one based on the question text and the other based on the answer text, and co-training them.

The above two existing studies are the most closely related to ours. One im-

portant difference of our work is that we carefully discuss what is subjective and objective, whereas these two works discuss little about what is conversational or what is subjective. Additionally, we do not rely on answer or user information, hence our approach is more widely applicable than their works.

3. Question Classification Task

In this section, we describe our criteria of question classification. The criteria were used as instructions when two annotators labeled the questions in our experiment. Our goal is to predict whether the asker needs a human response or not, and this is reflected in the criteria.

3.1 Annotation Criteria

The annotation criteria we used are as follows:

Type OBJ: Asker expects one or more **objective** answers.

- An objective answer is one that is based on some common knowledge or universal truth; it does NOT directly reflect the answerer's personal opinion. This type can be answered by a person with high expertise.
- Examples: facts, definitions, methods, how-to's.
- For these questions, the system is expected to look for answers automatically using IR or QA techniques.

Type SUB: Asker expects one or more **subjective** answers.

- A subjective answer is one that directly reflects the answerer's personal opinions or judgments. Questions which initiate a discussion also belong to this category.
- Examples: "What do you think about . . . ?", opinions, discussions.
- For these questions, the system is expected to actively prompt people to provide answers.

***** When the asker's question seems to expect both objective and subjective answers, classify the question as type SUB.

3.2 Examples of Annotation

"Who wrote Sherlock Holmes?" and "What does NEET stand for?" are typical type OBJ questions, known as factoid questions. However, in CQA data, such simple questions are relatively rare, and there are more complex type OBJ questions.

For example, we view "Please teach me how to make Aurora Sauce." as a type OBJ question even though this is not a single-truth question. This is because, even though there may be many ways to make Aurora Sauce, we assume that the user's intention is to make good Aurora Sauce and that a few good methods for making Aurora Sauce (retrieved automatically) will satisfy the user. Whereas, we view "Please tell me your home recipe for Aurora Sauce." as a type SUB question, as we assume that the user wants to survey different ways people make Aurora Sauce and requires personal responses.

Some questions are indeed very difficult to annotate even with detailed annotation criteria. Consider the following spectrum of questions: "how to solve $x^2 = x - 1$", "how to use Excel", "how to learn Japanese", "how to get a job", "how to get a girlfriend" and "how to spend my entire life". Probably many people will agree that the first question is type OBJ and that the last question is type SUB. The ones in between are more controversial, and we believe that this kind of uncertainty is inevitable. We therefore devised an evaluation method that takes into account this uncertainty, which we shall discuss in Section 5.

3.3 Discussions on the Criteria

The main difficulty of our problem setting is the definition of subjective and objective. In this section, we discuss other possible criteria and their problems.

One possible strategy for classifying questions is *to examine the variety of possible answers*. Factoid questions often have a unique answer or a small number of correct answers, while opinion questions inherently have many possible answers. However, this strategy can be counterintuitive in some cases. Consider this example: "[situation description]. . . Is it his fault or mine?" The number of possible answers may be one or two, but this question requires a subjective judgment in order to answer it. On the other hand, while "how to make Aurora Sauce" may have many correct answers, each answer does not necessarily require a subjective judgment and we therefore would like to view the question as objective. As was mentioned earlier, automatically retrieving one recipe for Aurora Sauce may satisfy the user.

Some dictionaries[⋆1] define subjective as *"Things existing in the mind"* or

⋆1 Dictionary.com (http://dictionary.reference.com/)

"Things related to person's emotion". But these definitions are also sometimes problematic. One example is "Is Yahoo! Auction popular?" While popularity can be considered as conceptual or emotional, it can also be quantified by some statistics.

Another definition of subjective is *"characteristics of individual; personal"*. This seems precise, but it is merely a paraphrase so that we need to define again what is personal and what is general.

Indeed, "personal" is probably an important keyword for defining subjective, so we define a subjective answer as "one that directly reflects the answerer's personal opinions or judgments", which we believe is a reasonably clear definition. We admit this may still be unclear in some situation, so we added a more practical guideline in the criteria: if the question can be satisfied by a single answer provided by a person with high expertise, then the question is objective.

4. Classification Method

To classify questions to type OBJ or type SUB, we tried some common methods for text classification.

4.1 Learning Machines

We tried to use two classifiers as follows:

(1) Support Vector Machine (SVM) [13]
(2) Naïve Bayes [14] (Bayesian Filter [15])

We used SVM because it is known to be highly effective for various classification tasks. As for the implementation, we used LIBSVM⋆1 with Radial Basis Function (RBF) kernel. Parameters were manually tuned with the training data. We also tried Naïve Bayes classifier with add 1 smoothing. While Naïve Bayes have several variants, we selected Bayesian Filter which is an effective method for spam filtering. Spam filtering is also a binary text classification task, so Baysian Filter is promising for question classification. For SVM classifiers, we used feature selection techniques [16]. We implemented Chi Square Test, Information Gain and Mutual Information based method, and selected Chi Square Test as it performed best. For Naïve Bayes, we did not use feature selection because we found that it only hurts the accuracy when used with smoothing.

4.2 Features

As for features, we tried some combination of features shown below:

(1) cue words (hand extracted words or expressions)
(2) word n-gram ($n = 1, 2, 3$)
(3) word dependency (a modifier and a modified relation)
(4) maximal repeats [17]

These features are known to be effective for other existing text classification task [18),19]. For (1), the first author extracted about 150 expressions, such as noun, verb phrases, and ending particles, which seemed useful for classification. For (2), when using n-gram we did not convert each word to its original form as this did not improve performance. For (3), we used CaboCha⋆2 as a dependency parser [20]. As for (4), we used maximal repeats, which minimally represent all occurrences of different repeats. Given a string, repeats are substrings that occur more than once, and maximal repeats are repeats that have at least one occurrence within the string such that no other repeat subsumes this occurrence. Given the string "unigram|bigram|hexagram", for example, "a", "gram" and "igram" are maximal repeats. Whereas, other repeats such as "gra" and "am" are not maximal repeats as their occurrences are completely subsumed by "gram". Maximal repeats can be calculated in $O(N)$-time using $O(N)$-space, where N is the length of given string, by using Suffix Tree [17] or Enhanced Suffix Array [21]. We applied maximal repeats to all the questions in the training set, so all frequent expressions in the questions can be considered as features.

In this study, we did not use any features that are specific to CQA data. However, features such as politeness and informativeness [22] may also be useful for our question classification task. We leave this to future work.

5. Experimental Settings

5.1 Dataset

In our experiments, we used the questions from the Yahoo! Chiebukuro data, which contains 3,116,009 questions posted between April 2004 and Oc-

⋆1 LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

⋆2 CaboCha (http://chasen.org/~taku/software/cabocha/)

Table 1 Annotation distribution and agreement for annotators A and B.

	B:DO	B:PO	B:PS	B:DS	Total
A:DO	**471**	**140**	49	3	663
A:PO	**37**	**41**	26	11	115
A:PS	11	29	**42**	**29**	111
A:DS	10	46	**114**	**441**	611
Total	529	256	231	484	1500

tober 2005. For evaluation, we used the NTCIR-8 Community QA Pilot Task question set [7),8)], which is a representative sample of the Chiebukuro data set. We manually classified these 1,500 questions for our subjective/objective classification task. When evaluating classification performance, the questions were sorted by timestamp and the first 1,000 questions were used for training while the remaining 500 questions were used for testing.

5.2 Annotation Settings

Using a simple click-based interface on a Web browser, two annotators (the first two authors of this paper) independently classified each question by relying on the criteria described in Section 3.1.

As we discussed earlier, some questions are hard to judge even when given detailed criteria. We therefore used four classes to handle judgment *confidence*: Definitely Objective (DO), Probably Objective (PO), Probably Subjective (PS) and Definitely Subjective (DS). We will refer to DO and DS "confident" labels and PO and PS as "unsure" labels. It is known that using multiple judges and non-binary judgments are useful for evaluating systems based on human judgments that can vary considerably across judges [8)].

5.3 Annotation Result

Table 1 shows the distribution of annotations over the four classes for the two annotators A and B, as well as the agreement between the two. It can be observed that:

(1) If we disregard the confidence (Definitely vs. Probably), the two annotators agreed on the subjective/objective annotation for as many as 1,315 questions (those shown in bold) out of the 1,500 (88%).

(2) For only 13 questions, the two annotators strongly disagreed with each other (i.e., the questions were labeled DS by one annotator and DO by the other).

(3) The ratio of type OBJ and type SUB is about half-and-half according to both two annotators.

(4) Annotator B used more "unsure" labels than A.

The substantial inter-annotator agreement (88%), which is supported by Cohen's kappa coefficient [23)] of 0.75, suggests that question classification task is reasonably well-defined and worth tackling. On the other hand, it is not a trivial task, as the distribution across type OBJ and type SUB is reasonably flat, as mentioned above in Finding (3). Moreover, the introduction of judgment confidence seems to have been useful, since the inter-annotator agreement between confident labels is far higher than that between unsure labels.

While our inter-annotator agreement is high, we closely examined the aforementioned 13 cases of strong disagreements and found that these disagreements are mainly caused by different interpretations of the intent behind the question. For example, Annotator A judged "What is the abbreviation for Kokkai-Gijido-Mae station?," as a DS question, thinking that "Kokkai-Gijido-Mae obviously does not have an abbreviation, so the asker must be joking or initiating a discussion." Whereas, Annotator B judged the same question as a DO question, interpreting it as a straight factoid question.

5.4 Evaluation Metrics

As evaluation metrics, we used weighted accuracy, as well as standard accuracy measure, to take annotators' confidence into account.

We first discuss evaluation based on a single annotator, where each question is labeled with DO, PO, PS or DS. Let Q be the set of questions used for evaluation, and let $N = |Q|$. Let n_{con} and $n_{uns}(= N - n_{con})$ denote the number of questions with confident (DO and DS) and unsure (PO and PS) labels, respectively. To reflect the degree of agreement between the predicted class with the annotator's label, we define a score for each question $q(\in Q)$ as follows:

- $score(q) = 2$, if the prediction is correct and the label is DO or DS (confident).
- $score(q) = 1$, if the prediction is correct but the label is PO or PS (unsure).
- $score(q) = 0$, otherwise, or the prediction is wrong.

Table 2 (a) shows the score for every case. Based on this, we define weighted accuracy as follows:

Table 2 Scores used when calculating weighted accuracy.

(a)	DO	PO	PS	DS
(b)	DO/DO DO/PO	PO/PO DO/PS	PS/PS PO/DS	DS/DS PS/DS
predicted OBJ	2	1	0	0
predicted SUB	0	0	1	2

$$weighted\ accuracy = \frac{\sum_{q \in Q} score(q)}{2 \cdot n_{con} + n_{uns}} . \quad (1)$$

That is, weighted accuracy is the sum of the scores normalized by its maximum possible value. Note that it is reduced to standard (unweighted) accuracy metrics if all non-zero scores are set to 1.

We extended the above evaluation method for the case of two annotators as follows. We first removed questions for which the two annotators disagreed and had equal confidence, i.e., (DO, DS) and (PO, PS). Then, we assigned the scores as shown in Table 2 (b). Note that, for example, a question with labels (DO, PO) is considered DO, and a question with labels (DO, PS) is considered collectively as PO (Compare with Table 2 row (a)). Based on this score assignment, we also computed the weighted accuracy based on the two annotators. In our experiment, we use this combined annotation data except for Section 6.2.

6. Results

6.1 Classification Results with Different Features

Table 3 shows the classification results using features described in Section 4. In each column, FS means Feature Selection and NB means Naïve Bayes, and standard accuracy values are shown on the left and weighted accuracy values are shown on the right. From this table, we can observe that:

(1) Weighted accuracy is consistently higher than standard accuracy.
(2) Naïve Bayes consistently achieves higher accuracy than SVM.
(3) The combination of Naïve Bayes and maximal repeats achieves the highest accuracy (79.5%) and weighted accuracy (81.0%).
(4) Merely using unigram and bigram features achieve over 80% weighted accuracy by Naïve Bayes with smoothing.
(5) For maximal repeats, smoothing is unnecessary or even hurts the classification accuracy.
(6) The cue words are unnecessary for maximal repeats.
(7) Dependency does not work as well as bigram feature.
(8) Using trigrams does not lead to higher accuracy.

Table 3 Accuracy and weighted accuracy of each learning result (cue word [cw], n-gram [ng], dependency [dep], maximal repeats [mr]).

acc - w.acc (%)	SVM	SVM (FS)	NB	NB (smooth)
cw	67.9 - 68.5	67.9 - 68.5	72.9 - 74.3	72.3 - 74.1
1g	69.6 - 70.6	72.9 - 74.7	70.6 - 71.6	73.8 - 75.2
1g+dep	71.0 - 72.8	72.3 - 74.1	72.7 - 73.8	75.7 - 77.1
1g+2g	71.2 - 73.0	**73.2 - 75.3**	73.2 - 74.5	77.8 - **80.2**
1g+2g+cw	**72.5 - 74.4**	72.5 - 74.4	73.8 - 75.3	**78.9 - 80.8**
1g+2g+3g	70.6 - 72.6	71.2 - 72.9	73.6 - 75.0	76.7 - 78.6
mr	71.5 - 72.7	71.7 - 72.7	**79.5 - 81.0**	78.2 - 79.8
mr+cw	71.2 - 72.8	72.3 - 73.1	79.3 - 81.0	78.0 - 79.6

Table 4 Classification result based on one annotator ($*$ indicates that the difference between annotators A and B in weighted accuracy is significantly different with two-sided signed test at $\alpha = 0.05$).

acc - w.acc (%)	SVM	SVM (FS)	NB	NB (smooth)
A:1g+2g+cw	69.6 - 70.1	69.6 - 70.1	70.6 - 71.1	74.6 - 75.3
B:1g+2g+cw	74.2 - 76.1$*$	76.2 - 78.5$*$	72.6 - 74.9	75.4 - 79.8
A:mr	67.6 - 67.9	69.4 - 70.2	76.8 - 77.2	75.2 - 76.0
B:mr	71.4 - 74.3$*$	72.8 - 75.3	78.0 - 81.3$*$	77.6 - 80.7$*$

Observation (1) suggests that questions that are difficult for annotators to classify (i.e., unsure questions) are also difficult for the systems to classify, as weighted accuracy uses a lighter penalty for misclassified unsure questions than standard accuracy does. For (5), one possible explanation is that maximal repeats are robust to data sparsity, because each maximal repeat occurs in the training data at least twice by definition. For (6), there is a clear explanation. Manually selected cue words tend to be those that occur frequently in the question text. However, maximal repeats already cover such frequent words.

6.2 Classification Results with Different Annotation Data

Whether two annotators are necessary or not to measure classification performance is an important matter of concern. **Table 4** shows the classification

result using single annotator (A and B) labeled data. We can observe that the classification result of annotator B is always higher than that of annotator A, and sometimes the difference is statistically significant. One possible explanation of such difference is annotator A considers the implicit intent behind the question text more often than annotator B. As learning machines cannot consider such implicit intents, evaluation based on annotator A's judgments may be more challenging.

By comparing Table 4 with the "1g+2g+cw" and "mr" rows of Table 3, it can be observed that while the absolute values differ according to the annotation data, the general trends remain similar. In particular, regardless of the annotation data used, "NB (smooth) with 1g+2g+cw" and "NB with mr" are the top performers. In absolute terms, the performances based on the combined annotations are slightly higher than those based on individual annotations. Thus, it is possible that it is easier for the machine to predict an average user's judgment than to predict a particular individual's judgment. This seems quite intuitive.

Overall, while we confirmed that combined annotations and individual annotations generally yield similar results, we believe that hiring multiple assessors is useful in order to represent different points of view in the gold standard data. Similar observations have been made in evaluating CQA answer ranking [24]. While we cannot conclude how many assessors are required for reliable experiments from our present study, it is possible that ensuring the quality of annotators is more important than the number of annotators [24].

6.3 Failure Analysis

We examined questions that were misclassified by our classifiers. Based on this analysis, we argue that the following techniques are necessary:

(1) more training data or a semi-supervised method.
(2) a machine learning algorithm which can consider feature dependency.
(3) external knowledge, especially about news.
(4) a method which can treat deep semantics.

Some failures seems to be due to lack of training data. By examining the misclassified test questions, we observed some overfitted features (i.e., words) from the training data as well as promising features in the test data that were not covered by the training data. **Figure 1** shows the effect of increasing training data on the weighted accuracy for our top two classifiers. This graph suggests that more training data will increase coverage of good features and reduce overfitting.

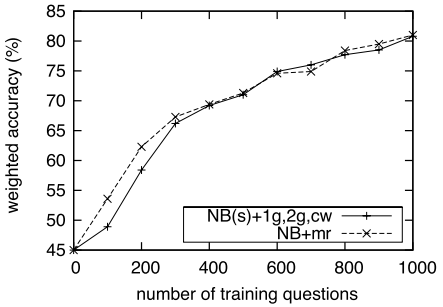

Fig. 1 Learning curves of two best classifiers.

In CQA, users often ask about current events. Such questions tend to initiate discussions. For example, "Why was Mr. Fukushima (football player) fired?" is difficult to judge whether the question is type OBJ or SUB. But if one has heard of this news and knows that the reason for firing him has not been disclosed to the public, he can judge, based on this knowledge, that this question is for initiating a discussion. Another good example is "Who is Hiroshi?" Hiroshi is a very common first name in Japan and the question appears to make no sense, but at the time when the question was posted there was a rising star comedian named Hiroshi (without a surname) is it probable that the asker had this particular Hiroshi in mind. These examples suggest that question text is often not sufficient for classifying questions, and that external knowledge sources, especially those that cover current affairs like sports and entertainment, would provide useful clues.

In our experiment, adding trigrams hurt performance. This may be due to noisy features, or redundant features already covered by unigrams and bigrams. While we already use noise reduction by means of feature selection and smoothing, there probably is room for reducing redundant features. The Naïve Bayes classifier assumes that features are independent of one another so cannot handle the above feature dependency problem. Some SVM kernels may solve this problem, but our experimental results are negative. Therefore it is possible that we need a more sophisticated learning algorithm.

Consider the example "Where is heaven?" which we view as a type SUB question. Any bag-of-words learning machines may judge it as type OBJ for the following reason. In this example, "is" is a neutral word so the problem boils down to classifying "where" and "heaven." Now, suppose that, in the training data, "where" occurred in 100 questions (10 times in type SUB and 90 times in type OBJ), "heaven" occurred in 4 questions (3 times in type SUB and once in type OBJ), and that they never co-occurred. Then, the conditional probabilities that the question is type SUB are P(SUB|where) = 0.1 and P(SUB|heaven) = 0.75. By combining these probabilities, classifiers are likely to judge the question as type OBJ. While some learning machines can interpret the occurrence counts (100 and 4) for measuring confidence, this does not help in this case as the low conditional probability for "where" will be given high confidence. Hence, to manage this problem, utilizing some kind of background knowledge or common sense that "heaven is a conceptual place and we do not even know if it exists" may be necessary.

7. Conclusions

In this paper, we defined a subjective/objective question classification task from the viewpoint of building a system that either prompts people to provide an answer or retrieves answers automatically. Two annotators annotated the NTCIR-8 CQA questions using our criteria, and the inter-annotator agreement was 88% (Cohen's kappa: 0.75). We showed that using Naïve Bayes (Baysian Filtering) with n-gram or maximal repeats features can achieve approximately 80% classification accuracy. We also showed that questions for which the annotators had high confidence are also easier for the machine to classify than the low-confidence ones. Our weighted accuracy measure, which takes into account this judgment confidence, seems to be useful for evaluating this task. We also argued that using multiple annotators is useful for reliable evaluation.

Our future work includes: developing a more effective subjective/objective classification algorithm; using a different set of annotation criteria for question classification and examining its effect on classification accuracy; investigating the effect of increasing the number of annotators; and expanding our language scope.

Acknowledgments We thank our colleagues at MSRA for good discussions. We also express our thanks to Yahoo! Japan and NII for providing us with the Yahoo! Chiebukuro data.

References

1) Soricut, R. and Brill, E.: Automatic Question Answering: Beyond the Factoid, *Proc. HLT-NAACL* (2004).
2) Roussinov, D., Fan, W. and Flores, J.R.: Beyond keywords: Automated question answering on the web, *Communications of the ACM*, Vol.51, No.9, pp.60–65 (2008).
3) Jeon, J., Croft, W.B. and Lee, J.H.: Finding similar questions in large question and answer archives, *Proc. 14th ACM International Conference on Information and knowledge management*, pp.84–90 (2005).
4) Jeon, J., Croft, W.B., Lee, J.H. and Park, S.: A framework to predict the quality of answers with non-textual features, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.228–235 (2006).
5) Li, B., Liu, Y. and Agichtein, E.: CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.937–946 (2008).
6) Harper, F.M., Moy, D. and Konstan, J.A.: Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites, *Proc. 27th International Conference on Human Factors in Computing Systems*, pp.759–768 (2009).
7) 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*, pp.421–432 (2010).
8) Sakai, T., Ishikawa, D. and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task (Part II): System Evaluation, *NTCIR-8 Proceedings*, pp.433–457 (2010).
9) Kim, S., Oh, J.S. and Oh, S.: Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective, *Proc. American Society for Information Science and Technology*, Vol.44, No.1, pp.1–15 (2007).
10) Kuriyama, K. and Kando, N.: Analysis of Questions and Answers in Q & A Site (in Japanese), *IPSJ SIG Technical Report*, Vol.2009-DBS-1, No.19, pp.1–8 (2009).
11) Adamic, L.A., Zhang, J., Bakshy, E. and Ackerman, M.S.: Knowledge sharing and yahoo answers: Everyone knows something, *Proc. 17th International Conference on World Wide Web*, pp.665–674 (2008).
12) Liu, Y., Li, S., Cao, Y., Lin, C.-Y., Han, D. and Yu, Y.: Understanding and summarizing answers in community-based question answering services, *Proc. 22nd International Conference on Computational Linguistics*, pp.497–504 (2008).
13) Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features, *Proc. 10th European Conference on Machine Learning*, Berlin/Heidelberg, pp.137–142 (1998).

14) Lewis, D.: Naive (Bayes) at forty: The independence assumption in information retrieval, *Proc. 10th European Conference on Machine Learning*, Berlin/Heidelberg, pp.4–15 (1998).
15) Graham, P.: A Plan For Spam, http://www.paulgraham.com/spam.html (2002).
16) Yang, Y. and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, *Proc. 14th International Conference on Machine Learning*, pp.412–420 (1997).
17) Gusfield, D.: *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press (1997).
18) Kudo, T. and Matsumoto, Y.: A Boosting Algorithm for Classification of Semi-Structured Text (in Japanese), *IPSJ Journal* (2004).
19) Okanohara, D. and Tsujii, J.: Text Categorization with All Substring Features, *Proc. SIAM International Conference on Data Mining* (*SDM*), pp.838–846 (2009).
20) Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S. and Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing, *Natural Language Engineering*, Vol.13, No.2, pp.95–135 (2007).
21) Abouelhoda, M.: Replacing suffix trees with enhanced suffix arrays, *Journal of Discrete Algorithms*, Vol.2, No.1, pp.53–86 (2004).
22) Ishikawa, D., Kuriyama, K., Sakai, T., Seki, Y. and Kando, N.: Analysis of Best-Answer Estimation in Q&A Site and Its Application to Machine Learning (in Japanese), *Japan Society for Information and Knowledge 18th Annual Conference Proceedings*, Vol.20, No.2, pp.73–85 (2010).
23) Landis, J.R. and Koch, G.G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol.33, No.1, pp.159–174 (1977).
24) Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K. and Lin, C.: Using graded-relevance metrics for evaluating community QA answer selection, *Proc. 4th ACM International Conference on Web Search and Data Mining*, pp.187–196 (2011).

(Received December 8, 2010)
(Accepted April 10, 2011)

(Editor in Charge: *Miyuki Nakano*)

Naoyoshi Aikawa worked as an internship student at Microsoft Research Asia in 2010. He received a Master of Engineering degree from Waseda University in 2011.

Tetsuya Sakai received a master's degree from Waseda University and joined Toshiba in 1993. He received a Ph.D. from Waseda University in 2000. He was a visiting researcher at the University of Cambridge from 2000 to 2001. He was the director of the Natural Language Processing Laboratory at NewsWatch, Inc. from 2007 to 2009. He is currently a lead researcher at Microsoft Research Asia. He has received several awards for his research in the area of information access, including two IPSJ Best Paper Awards, the Yamashita Award and the FIT Funai Best Paper Award.

Hayato Yamana received his Doctor of Engineering degree at Waseda University in 1993. He began his career at the Electrotechnical Laboratory (ETL) of the former Ministry of International Trade and Industry (MITI). He was subsequently appointed Associate Professor of Computer Science at Waseda University in 2000, and has been a professor in that department since 2005. From 2003 to 2004, he was IEEE Computer Society Japan Chapter Chair. From 2010, he is the chairman of Database system SIG of IPSJ. He has written, co-written and translated a number of books including Google Hacks (translation supervisor).