ILP-based Opinion Sentence Extraction from User Reviews for Question DB Construction

Masakatsu Hamashita* and **Takashi Inui**
University of Tsukuba
1-1-1 Tenoudai, Tsukuba, Ibaraki 305-8573, JAPAN
{m.hama@mibel.,inui@}cs.tsukuba.ac.jp

Koji Murakami and **Keiji Shinzato**
Rakuten Institute of Technology - Boston, Rakuten USA
2 South Station Suite 400, Boston, MA 02110
{koji.murakami,keiji.shinzato}@rakuten.com

Abstract

Typical systems for analyzing users' opinions from online product reviews have been researched and developed successfully. However, it is still hard to obtain sufficient user opinions when many reviews consist of short messages. This problem can be solved with an active opinion acquisition (AOA) framework that has an interactive interface and can elicit additional opinions from users. In this paper, we propose a method for automatically constructing a question database (QDB) essential for an AOA. In particular, to eliminate noisy sentences, we discuss a model for extracting opinion sentences that is formulated as a maximum coverage problem. Our proposed model has two advantages: (1) excluding redundant questions from a QDB while keeping variations of questions and (2) preferring simple sentence structures suitable for the question generation process. Our experimental results show that the proposed method achieved a precision of 0.88. We also give details on the optimal combination of model parameters.

1 Introduction

Typical systems for analyzing users' opinions from online product reviews have been researched and developed successfully (Liu, 2012; Jo and Oh, 2011; Kouloumpis et al., 2011; Pozzi et al., 2016). However, it is still hard to obtain sufficient user opinions when many reviews consist of short messages. In this situation, it would be practical to elicit additional opinions by actively asking users questions instead of just waiting for user posts. We define this procedure as an active opinion acquisition (**AOA**).

Suppose an example which is a review post consisting of just one sentence below:

u1 *This wine has a really refreshing aroma!*

It is possible to capture the user opinion "*refreshing aroma*" from **u1**. Here, in the case of an AOA-oriented system (AOAS), the system asks a question like **s1** after **u1**.

u1 *This wine had a really refreshing aroma!*

s1 <u>*How was the aftertaste?*</u>

u2 *The aftertaste was bitter.*

Then, it is also possible to obtain the additional opinion "*bitter aftertaste*" from **u2**. This example shows that an AOAS can efficiently collect user opinions by asking users questions.

Here, a question database (QDB), that is, a set of large quantities of question examples, is an essential resource for realizing dialogues between a user and an AOAS (Murao et al., 2003) because it would enable an AOAS to ask users precise questions in various situations. Nio and Murakami (2018) proposed a question-conversion method for constructing QDBs automatically. This method runs through a machine translation-like architecture and then converts an affirmative sentence to an interrogative form such as:

The aroma was a bouquet.
$\rightarrow$ *How was the aroma?*

*Currently, Gunosy Inc.

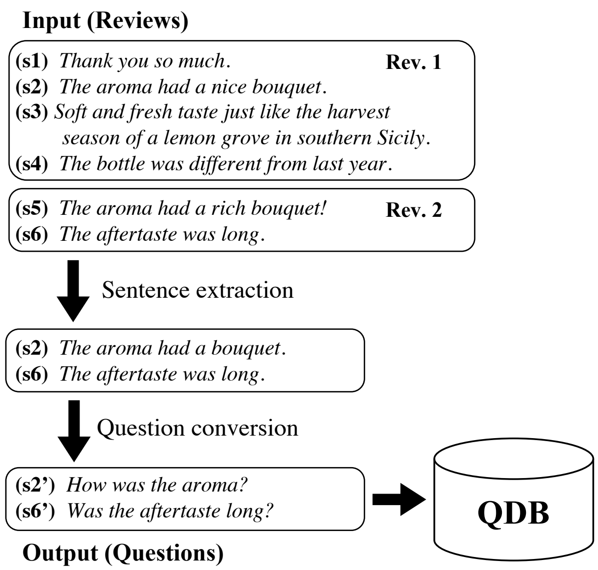

Figure 1: Relationship between sentence extraction and question conversion. Given multiple user reviews, the sentence extraction module is applied for eliminating noisy sentences and then extracted sentences are sent to the question conversion.

Note that a relationship holds that the input opinion sentence is the answer to the output question. Nio and Murakami (2018) reported a method that achieves state-of-the-art performance by using a user-review data set prepared purely for evaluation. Unfortunately, however, real review data is very noisy, so measures against such noisy data are required.

In this paper, we propose a novel sentence extraction model that eliminates noisy sentences and extracts sentences suitable for question conversion. The proposed model works as a preprocessing module for question conversion, as shown in Figure 1. Here, note that each sentence to be extracted needs to include opinion(s) like (**s2**) and (**s6**). Therefore, the proposed model is formulated as a maximum coverage problem of opinions, which makes it possible to exclude sentences including no opinions like (**s1**) and (**s4**). Naturally, the formulation also makes it possible to exclude sentences that have redundant content like (**s5**). Moreover, the basic formulation is extended to exclude sentences having sentence structures that are too complex for question conversion like (**s3**). The extended model enables us to control the number of opinions in each output sentence in order to extract opinion sentences that have simple structures. Details on the proposed model will be given in Section 3.

Through experiments done for evaluation, it is found that the proposed method achieved a precision of 0.880. Furthermore, we revealed the characteristics of the extracted opinion sentences in terms of length and the number of types of opinions. We also give details on the optimal combination of model parameters.

2 Related Work

2.1 QGSTEC

The automatic generation of questions is essential to various applications such as dialog systems and quiz generation in educational E-learning systems. The question generation shared task and evaluation challenge (QGSTEC) is a shared task for automatically generating questions for those applications. In QGSTEC, given a text segment, the goal of a system is to generate questions whose answers are included in the input segment. There have been many successful studies based on QGSTEC (Mannem et al., 2010; Ali et al., 2010; Agarwal et al., 2011). Nevertheless, our final goal is to generate questions that enable an AOAS to elicit user opinions, quite different from QGSTEC.

2.2 Neural Question Generation

Zhang et al. (2018) proposed a question generation model that uses a neural network. On a news web site, if the headline of an article is a question, the click through rate increases; thus, a question headline is generated by using an encoder-decoder model. This model requires correct answer data because it involves supervised learning. Our study differs from this study in that correct answer data is not required because our study involves unsupervised learning with only reviews and question examples are created instead of question headlines.

2.3 ILP-based Sentence Extraction

Sentence extraction has been widely studied as a form of document summarization (Kupiec et al., 1995; Hirao et al., 2002). Among the methods of extraction proposed so far, integer linear programming (ILP) formulation provides better solutions because of its flexibility and extensibility. Given a set of sentences $D = \{s_1, \ldots, s_N\}$ as an input, ILP-based sentence extraction aims at constructing an appropriate subset $S \subseteq D$. Here, suppose D is represented by an N-dimensional 0/1 vector $\mathbf{y} = \{y_1, \ldots, y_N\}$. When a sentence s_i in D is $s_i \in S$, $\mathbf{y}$ represents the result of sentence extraction as $y_i = 1$; otherwise, $y_i = 0$.

The most fundamental model of ILP-based sentence extraction is formulated as Figure 2.

$$
\begin{aligned}
\mathbf{y}^* &= \arg\max_{\mathbf{y}} f(\mathbf{y}) \\
\text{s.t.} \quad & \sum_{i=1}^{N} l_i y_i \leq L_{max} \\
& \forall i, \quad y_i \in \{0, 1\}
\end{aligned}
$$

Figure 2: Fundamental model for ILP-based sentence extraction

Here, L_{max} represents the maximum output length, and l_i represents the length of a sentence s_i. The function $f(\mathbf{y})$ is an objective function that measures the quality of an output candidate $\mathbf{y}$. The model outputs the candidate holding a maximum value of $f(\mathbf{y})$ while satisfying all constraints.

2.4 Maximum Coverage Model

The maximum coverage model (MCM) is an instance of an ILP-based sentence extraction model, that is known to be suitable for multi-document summarization (Yih et al., 2007). MCM prefers to create a summary output that has as many varieties of *concepts*, typically words, as possible. As a result, this model is naturally able to exclude redundant concepts from the output.

Multi-document summarization based on the MCM is formulated as Figure 3. Here, the objective function $f_{mcm}(\mathbf{y})$ is defined as follows:

$$f_{mcm}(\mathbf{y}) = \lambda \sum_i r_i y_i + (1 - \lambda) \sum_k w_k z_k$$

The w_k in $f_{mcm}(\mathbf{y})$ represents the weight of the word k. The r_i represents the similarity score between a sentence s_i and entire input documents. The

$$
\begin{aligned}
\mathbf{y}^* &= \arg\max_{\mathbf{y}} f_{mcm}(\mathbf{y}) \\
\text{s.t.} \quad & \sum_{i=1}^{N} l_i y_i \leq L_{max} \\
& \forall k, \quad \sum_i o_{ik} y_i \geq z_k \\
& \forall i, \quad y_i \in \{0, 1\} \\
& \forall k, \quad z_k \in \{0, 1\}
\end{aligned}
$$

Figure 3: Maximum coverage model for multi-document summarization

z_k is a 0/1 variable that is 1 when a word k is included in an output candidate, and 0 otherwise. Also, o_{ik} in Figure 3 is a constant that becomes 1 when s_i contains k, 0 otherwise. The model guarantees consistency between y_i and z_k through the constraint $\sum_i o_{ik} y_i \geq z_k$. Nishikawa et al. (2010) proposed a variation of the MCM for multi-document opinion summarization. This model adopts an opinion as the concept e_k instead of a word to create a summary that has as many varieties of opinions as possible. The objective function $f_{nishikawa}(\mathbf{y})$ is defined as follows:

$$f_{nishikawa}(\mathbf{y}) = \lambda \sum_k w_k z_k + (1 - \lambda) \sum_{i,j} c_{i,j} x_{i,j} \tag{1}$$

The first term is the same as the second term of $f_{mcm}(\mathbf{y})$. In the second term of $f_{nishikawa}(\mathbf{y})$, $x_{i,j}$ is a decision variable that indicates the sentence order, and $c_{i,j}$ is a weight related to the naturalness of the sentence order. This makes it possible to select sentences so that important concepts are included in the summary and arrange those sentences as naturally as possible.

This is similar to our model proposed in the next section. However, its focal point is different from ours. The model of (Nishikawa et al., 2010) does not care how many opinions are included in each sentence in the output, while the proposed model controls the number of opinions in each output sentence in order to extract opinion sentences that have simple structures. The details will be given in the next section.

3 Proposed Method

In this section, we describe our novel sentence-extraction model based on the MCM formulation. Given a set of user review sentences, the model is expected to extract sentences suitable for question conversion, as mentioned in Section 1.

Suppose again that, given the six sentences shown in Figure 1 as input, only (**s2**) and (**s6**) should be extracted and sent to the question conversion process. Sentences (**s1**) and (**s4**) should not be extracted because they include no opinions at all. (**s3**) and (**s5**) are not worth extracting despite both sentences including opinions. (**s5**) is redundant because it has almost the same meaning as (**s2**) [1]. In addition, (**s3**) has too complex of a sentence structure for question conversion.

From these observations, it was found that each sentence output from the proposed model should satisfy the following requirements.

> **Requirement I:** include opinion(s),
>
> **Requirement II:** have a simple sentence structure, and
>
> **Requirement III:** exclude redundant content appearing in other output sentences.

Among these three, the first and third requirements can be achieved by applying a MCM framework, as mentioned in the previous section. In this paper, we propose an extension of the basic MCM to satisfy the second requirement. First, we propose additional constraints to control the number of opinions in each output sentence, and we then describe a novel objective function for estimating how much standard the expression of opinion is.

Figure 4 shows the formulation of the proposed model. Note that an opinion $\langle a_j, e_k \rangle$ is assigned as the *concept* in the MCM framework. Here, $a_j (\in Q_a)$ is an aspect word such as "*aftertaste*," $e_k (\in Q_e)$ is a sentiment word such as "*bitter*," and Q_a and Q_e represent a pre-defined set of aspect words and sentiment words, respectively.

Two constraints, Equation (2) and (3) in Figure 4, are added to control the number of opinions in an output sentence. A_{max} and E_{max} are constants representing the maximum number of aspect and sentiment words included in an output sentence, respectively. The function $c_a(\mathbf{y}_i, a_j)$ in Equation (2) indicates the number of sentences that contain a_j in $\mathbf{y}_i$ and is defined as follows.

$$
\begin{aligned}
\mathbf{y}^* &= \arg\max_{\mathbf{y}} f_{prop}(\mathbf{y}) \\
\text{s.t.} \quad & \sum_{i=1}^{N} l_i y_i \le L_{max} \\
& \forall i, \quad \sum_{j=1}^{|Q_a|} c_a(\mathbf{y}_i, a_j) \le A_{max} \quad (2) \\
& \forall i, \quad \sum_{k=1}^{|Q_e|} c_e(\mathbf{y}_i, e_k) \le E_{max} \quad (3) \\
& \forall j,k, \quad \sum_{i=1}^{N} o_{ijk} y_i \ge z_{jk} \quad (4) \\
& \forall i, \quad y_i \in \{0,1\} \\
& \forall j,k, \quad z_{jk} \in \{0,1\}
\end{aligned}
$$

Figure 4: Proposed model. It enables control of number of opinions in each output sentence through additional constraints.

$$\sum_{i=1}^{N} h_{ij} y_i$$

The h_{ij} takes 1 if a sentence s_i contains the aspect word a_j and 0 otherwise. Here, $\mathbf{y}_i$ is a vector for which the *i*-th element is the same value as that of $\mathbf{y}$, and the others are 0. As a result, $c_a(\mathbf{y}_i, a_j)$ takes 1 if s_i contains a_j and 0 otherwise, and the function $c_e(\mathbf{y}_i, e_k)$ in Equation (3) is similarly defined as $c_a(\mathbf{y}_i, a_j)$ for sentiment words. The constraint of Equation (4) has the same role as the original MCM in Figure 3. It is modified slightly from the original model due to the *concept* (opinion) structure. Here, z_{jk} is a variable that has 1 when an opinion $\langle a_j, e_k \rangle$ is included in the output and 0 otherwise.

The objective function $f_{prop}(\mathbf{y})$ for the proposed model is defined as follows.

$$f_{prop}(\mathbf{y}) = \sum_{j=1}^{|Q_a|} \sum_{k=1}^{|Q_e|} w_{jk} z_{jk} \quad (5)$$

[1] On the contrary, (**s2**) is redundant if the model outputs (**s5**).

It forms a simple version of $f_{nishikawa}(\mathbf{y})$. The value of $f_{prop}(\mathbf{y})$ becomes larger when the output includes many different types of opinions. We use half of $f_{nishikawa}(\mathbf{y})$ because our model does not need to consider the order of sentences unlike (Nishikawa et al., 2010).

When asking a user a question, the model prefers standard expressions frequently used. From this fact, the weight w_{jk} of the variable z_{jk} is defined as:

$$w_{jk} = \frac{w_{jk}^{word}}{w_{jk}^{syn}} \tag{6}$$

Here, w_{jk}^{word} represents the co-occurrence probability of an aspect word a_j and a sentiment word e_k in an input document. w_{jk}^{syn} represents the average syntactic distance between a_j and e_k, which increases the weight of syntactically concise opinions in which aspect words and sentiment words appear close to each other. These values are calculated separately from a large review data set.

Now, we explain how to determine which pairs of aspect words and sentiment words are regarded as opinions in a sentence. Given a sentence S, V_a represents a subset of Q_a, whose elements are aspect words in S. Also, V_e represents a subset of Q_e. The opinion $\langle a_j, e_k \rangle$ is determined in S immediately when $(|V_a|, |V_e|) = (1, 1), a_j \in V_a$ and $e_k \in V_e$. However, we need to discover meaningful word pairs when several aspect words and sentiment words are included in S, such as $(|V_a|, |V_e|) = (2, 3)$. We solved this problem by performing maximum weight matching on a weighted complete bipartite graph (Korte et al., 2012), where $G(V_a \cup V_e, E)$ is a complete bipartite graph, in other words, every combination of a_j and e_k in S becomes a candidate of opinions. Each candidate $\langle a_j, e_k \rangle$ is weighted by Equation (5).

Table 1 shows examples of opinions with higher weights that were calculated by using the same data used in the experiments in Section 4.1. Similarly, Table 2 shows the case of lower weights. One can see that plausible opinions are included in Table 1 while meaningless aspect/sentiment word pairs are included in Table 2.

Table 1: Examples of high weight opinions

⟨*balance, good*⟩
⟨*taste, long*⟩
⟨*taste, rich*⟩
⟨*aroma, spread*⟩
⟨*cost-performance, excellent*⟩

Table 2: Examples of low weight opinions

⟨*cork, strong*⟩
⟨*taste, hero*⟩
⟨*label, soft*⟩
⟨*bottle, long*⟩
⟨*price, beautiful*⟩

4 Experiments

4.1 Experimental Settings

The following two experiments were conducted.

Experiment I We conducted a series of experiments where combinations of model parameters (A_{max} and E_{max}) were changed to investigate the relationship between the performance and the parameters of the proposed model. Hereafter, we refer to the proposed model as **ILP+C**$_{(A_{max},E_{max})}$ when showing the parameters of the model clearly.

Experiment II We compared a simple version of the ILP-based sentence extraction model, namely **ILP-only**, with a non-ILP-based method to verify the effectiveness of ILP-based formulation. **ILP-only** is equivalent to the proposed model without the additional constraints [Equations (2) and (3)]. Additionally, the proposed model is compared with ILP-only to evaluate the effectiveness of the additional constraints.

We used a set of Japanese user review sentences posted on Rakuten Japan[2], which is one of the major E-commerce web sites in Japan. First, we crawled the sentences in the wine category and randomly selected 1,000 sentences from 19,160 sentences. Then, two annotators independently judged

[2]https://www.rakuten.co.jp

Table 3: Data set for evaluation

#Sentences(Positive/Negative)	715(367/348)
$\overline{\text{aspect}}$	1.97
$\overline{\text{sentiment}}$	1.61
$\overline{\text{length}}$	53.6

whether sentences satisfied the requirements shown in Section 3. Details on the data set are given in Table 3. Here, the symbol "Positive" indicates that a sentence can be converted into relevant questions, that is, it should be extracted, and "Negative" the opposite. $\overline{\text{Aspect}}$ and $\overline{\text{sentiment}}$ indicate the average number of aspect lexicons and sentiment lexicons per sentence, respectively, and $\overline{\text{length}}$ indicates the average number of characters per sentence. Cohen's Kappa, which means the degree of inter-annotator agreement, was 0.765 (Cohen, 1960).

We handcrafted a set of aspect lexicons Q_a and a set of sentiment lexicons Q_e by collecting opinions that appeared in the data set for evaluation because no Japanese aspect/sentiment lexicons suitable for our data set exist. As a result, we determined that $|Q_a| = 81$ and $|Q_e| = 835$. Here, we collected only sentiment lexicons with a positive polarity according to the findings of (Hamashita et al., 2018); it is suitable that questions used in an AOAS include contents with positive polarity.

In Experiment I, A_{max} and E_{max} in the proposed model are changed from 1 to 5, respectively. The non-ILP-based method used in Experiment Ⅱ is a weight-based method that extracts sentences with higher weights until the total size of the extracted opinion sentences is over L_{max}. The weight of sentence s_i is calculated by summing up the weights of the opinions w_{jk} defined in Equation (5), included in s_i. We refer to this method as **w/oILP** hereafter. For each run of all experiments, the ILP solution was obtained by using Python's PuLP library (Mitchell et al., 2011), and L_{max} was set to hold a summarization rate of 5%.

We used a precision measure of the extractions, the average length of the extracted sentences (|Sentence|), the number of extracted sentences (#Sentences), and the number of types of opinions included in the extracted sentences (#Opinions) as

Table 4: Precision value for each (A_{max}, E_{max})

		E_{max} 1	2	3	4	5
A_{max}	1	.666	.701	.735	.735	.735
	2	.821	.810	.794	.774	.782
	3	.864	.826	.794	.833	.819
	4	**.880**	.810	.782	.794	.791
	5	.868	.794	.785	.794	.797

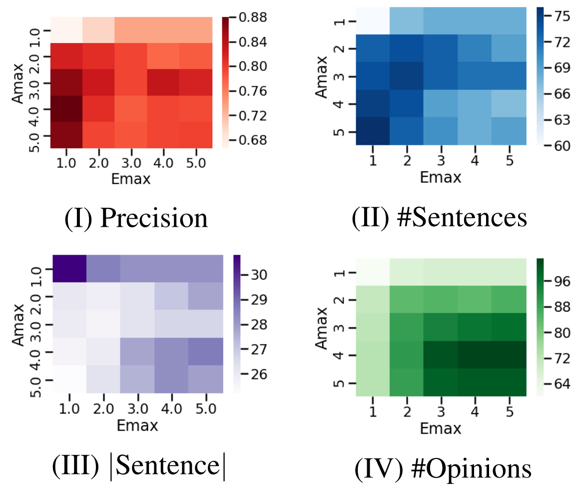

Figure 5: Heat map representations corresponding to results for each evaluation measure. For each map, as metric value becomes larger, cell becomes darker.

the evaluation measures.

4.2 Results

First, Table 4 and Figure 5 show the results of Experiment I. Here, Figure 5 represents heat maps corresponding to the results for each evaluation measure, where the vertical axis indicates A_{max}, and the horizontal axis indicates E_{max}. For each map, the larger the metric value becomes, the darker the color of a cell is.

Table 4 shows the values of the precision measure. It turns out that the precision tended to be large when $E_{max} = 1$. Notably, the best result of 0.880 was achieved for $(A_{max}, E_{max}) = (4, 1)$. We found that almost all opinion sentences extracted by ILP+C$_{(4,1)}$ kept a simple sentence structure. Examples of the extracted sentences are shown in Figure 6(A).

In comparison between Figure 5(I) and

Table 5: Results of Experiments Ⅱ

	w/oILP	ILP-only
Precision	.621	**.803**
|Sentence|	66.2	29.1
#Sentences	29	66
#Opinions	47	102

Figure 5(II), precision and #Sentences show similar results. That is to say, both metric values became larger when $E_{max} = 1$.

Figure 5(III) holds the reversed proportion against Figure 5(II). The reason could be that the value of #Sentences multiplied by that of |Sentence| tends to remain constant due to the constraint of L_{max}. Next, it was found in Figure 5(III) that the sentences extracted by ILP+C$_{(1,1)}$ had a large |Sentence| and also found in Figure 5(II) that the precision rapidly decreased when $(A_{max}, E_{max}) = (1, 1)$. Now, we discuss why the precision decreased. Some of the correct and wrong examples extracted by ILP+C$_{(1,1)}$ are shown in Figure 6(B) and Figure 6(C), respectively. From Figure 6(B), we can see that the correct examples had short lengths and simple structures similar to those of ILP+C$_{(4,1)}$, while the wrong examples in Figure 6(C) tended to be long due to their containing useless words. We also observed that the sentences shown in Figure 6(C) were not extracted when A_{max} increased. From the results, it is expected that inappropriate (long) sentences would be over-extracted due to there being a lack of sentences that satisfy the constraints when $(A_{max}, E_{max}) = (1, 1)$.

As shown in Equation (5), the objective function tended to return a larger value when there were a variety of opinions in the output sentences. This relationship immediately lead to the phenomenon that the larger both A_{max} and E_{max} became, the larger #Opinions became. This corresponds to the results shown in Figure 5(IV). Here, we note the results for $(A_{max}, E_{max}) = (5, 5)$. In this case, the precision (0.797) was lower than the best of 0.880 from Table 4. The reason could be that ILP+C$_{(5,5)}$ attempts to extract sentences that include multiple opinions in order to include as many opinions as possible in the output as shown in Figure 6(D).

Next, Table 5 shows the results of Experiment Ⅱ.

Table 6: Correlation coefficients

	correlation coefficient
#Sentences	0.85
#Opinions	0.33
$f_{prop}(\mathbf{y}^*)$	0.42

From the table, we found that (1) ILP-only achieved better precision than w/oILP and that (2) the output obtained by ILP-only included a lot of short sentences with varieties of opinions. Therefore, the ILP-based model was verified to be appropriate for our purpose. The precision of ILP-only was 0.803, confirming that the proposed method had a better extraction precision. ILP-only is an extreme case of the proposed model and strictly equivalent to ILP+C$_{(\infty,\infty)}$. Therefore, ILP-only is considered to be a model similar to ILP+C$_{(5,5)}$. Looking at Table 4 and Table 5, it can be confirmed that the precision of ILP-only and ILP+C$_{(5,5)}$ were similar.

Finally, we discuss how to estimate (A_{max}, E_{max}), which maximizes the precision without seeing it. We mentioned above that the metric #Sentences varies the same as precision. In addition to the findings, we investigated the correlation coefficients between precision and other metrics of each (A_{max}, E_{max}) to find a suitable metric that estimates (A_{max}, E_{max}). The results are shown in Table 6. Since #Sentences and |Sentence| are approximately inversely proportioned, the correlation coefficient with |Sentence| is not included in the table. The function $f_{prop}(\mathbf{y}^*)$ was added to the target metric for the investigation. As a consequence, the correlation coefficient between #Sentences and precision was the largest, while the other correlation coefficients were low. From these results, we can conclude that one can select (A_{max}, E_{max}) with the largest #Sentences. We get $(A_{max}, E_{max}) = (5, 1)$ in the case of our experimental settings if we adopt this strategy. The precision is not optimal but is the second largest when $(A_{max}, E_{max}) = (5, 1)$; thus, we consider that (A_{max}, E_{max}) can be estimated almost exactly by referring to #Sentences.

5 Conclusion

We proposed a novel model for extracting opinion sentences for constructing question DBs. The pro-

(A): sentences extracted by ILP+C$_{(4,1)}$

[c] $_3$果実 $_2$味の $_1$バランスが $_1$素晴らしい． / *The $_1$balance of $_3$fruity $_2$taste is $_1$excellent.*

[c] $_1$インパクトの $_1$強い $_2$味です． / *The $_1$impression of the $_2$taste was $_1$strong.*

[c] $_2$タンニンも $_1$まろやかな $_1$味わいです． / *The $_2$tannin was a $_1$mild $_1$taste.*

(B): sentences extracted by ILP+C$_{(1,1)}$

[c] とても $_1$果実が $_1$豊か． / *It was $_1$very $_1$fruity.*

[c] すべての要素の $_1$バランスが $_1$絶妙です. / *The $_1$balance of all elements was $_1$exquisite.*

(C): sentences extracted by ILP+C$_{(1,1)}$

[w] $_1$ラベルのロゴも $_1$爽やかなライトブルーをあしらってなかなかクールなイメージ. / *The logo of the $_1$bottle label was $_1$refreshing light blue and so cool.*

[w] $_1$程よい$_1$酸味を感じながらスッキリとお飲みいただくことができます. / *You can drink refreshingly with $_1$moderate $_1$acidity.*

(D): sentences extracted by ILP+C$_{(5,5)}$

[c] $_1$柔らかな $_1$タンニンが大きく $_2$広がる $_2$味わい. / *The $_2$taste of $_1$soft $_1$tannin was $_2$widespread.*

[c] $_1$酸味が $_1$甘さを抑えた $_2$バランスの $_2$良い $_3$軽快な $_3$味わい. *The $_3$light $_3$flavor with $_2$best $_2$balance between $_1$acidity and $_1$sweetness.*

[w] $_1$香りは $_1$フルーティーな印象ながら、 $_4$穏やかな泡と $_2$すっきりとした $_2$キレのある $_3$味わいは、$_3$しっかりと冷やして お料理と合わせるのがおすすめです. / *This $_3$cool wine matches your special dinner because it has $_1$fruity $_1$aroma, $_2$clear and $_2$sharp $_3$taste, and $_4$mild foam.*

[w] $_5$タンニンと $_4$酸味の $_1$バランスが $_1$よく、$_3$フルーティーで、$_2$口当たりの $_2$優しいワインが多い. / *A lot of wines which have a good balance of $_5$tannin and $_4$acidity, $_3$fruity, and $_2$pleasant $_2$taste.*

Figure 6: Examples of original Japanese sentences and their literal translations into English. The symbols **[c]** and **[w]** indicate correct and wrong extractions, respectively. Underline parts indicate aspect words, and double underline parts indicate sentiment words. A pair of aspect and sentiment words with the same arabic number means an opinion.

posed model was formulated as a maximum coverage problem of opinions. Our model has additional constraints to control the number of opinions in each output sentence and also has an objective function in order to extract opinion sentences that have simple structures. From the experimental results, we found that ILP+C$_{(4,1)}$ achieved a precision of 0.88. We also found that one can achieve promising results when selecting (A_{max}, E_{max}) with the largest #Sentences.

For future work, it is necessary to improve an opinion detection method suitable for our Japanese data set. While we applied a simple dictionary-based detection method in this work, more sophisticated methods (Brody and Elhadad, 2010; He et al., 2018) could be combined with our model. We also plan to develop an AOAS with a QDB constructed with the proposed model and conduct comprehensive evaluations.

References

[Agarwal et al.2011] Manish Agarwal, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Ed-*

ucational Applications, pages 1–9.

[Ali et al.2010] Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 58–67.

[Brody and Elhadad2010] Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812.

[Cohen1960] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Hamashita et al.2018] Masakatsu Hamashita, Takashi Inui, Koji Murakami, and Keiji Shinzato. 2018. Insertion effect of negative affix in question generation for interactive information collection system. (in Japanese). In *The Association for Natural Language Processing, 25(25)*.

[He et al.2018] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1121–1131.

[Hirao et al.2002] Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002. Extracting important sentences with support vector machines. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7.

[Jo and Oh2011] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824.

[Korte et al.2012] Bernhard Korte, Jens Vygen, B Korte, and J Vygen. 2012. *Combinatorial optimization*, volume 2. Springer.

[Kouloumpis et al.2011] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*.

[Kupiec et al.1995] Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 68–73.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Mannem et al.2010] Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at upenn: QGSTEC system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91.

[Mitchell et al.2011] Stuart Mitchell, Michael OSullivan, and Iain Dunning. 2011. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand, http://www. optimization-online. org/DB_FILE/2011/09/3178. pdf*.

[Murao et al.2003] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, and Yasuyoshi Inagaki. 2003. Example-based spoken dialogue system using woz system log. In *Proceedings of the Fourth SIGdial Workshop of discourse and dialogue*.

[Nio and Murakami2018] Lasguido Nio and Koji Murakami. 2018. Intelligence is asking the right question: A study on Japanese question generation. In *IEEE Spoken Language Technology conference*.

[Nishikawa et al.2010] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 910–918.

[Pozzi et al.2016] F. Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment analysis in social networks*. Morgan Kaufmann.

[Yih et al.2007] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 7, pages 1776–1782.

[Zhang et al.2018] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 617–626.