Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining

Nozomi Kobayashi [*] **Kentaro Inui, and Yuji Matsumoto**
Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
{nozomi-k,inui,matsu}@is.naist.jp

Abstract

The technology of opinion extraction allows users to retrieve and analyze people's opinions scattered over Web documents. We define an opinion unit as a quadruple consisting of the opinion holder, the subject being evaluated, the part or the attribute in which the subject is evaluated, and the value of the evaluation that expresses a positive or negative assessment. We use this definition as the basis for our opinion extraction task. We focus on two important subtasks of opinion extraction: (a) extracting aspect-evaluation relations, and (b) extracting aspect-of relations, and we approach each task using methods which combine contextual and statistical clues. Our experiments on Japanese weblog posts show that the use of contextual clues improve the performance for both tasks.

1 Introduction

The explosive increase in Web communication has attracted increasing interest in technologies for automatically mining personal opinions from Web documents such as product reviews and weblogs. Such technologies would benefit users who seek reviews on certain consumer products of interest.

Previous approaches to the task of mining a large-scale document collection of customer opinions (or reviews) can be classified into two approaches: Document classification and information extraction. The former is the task of classifying documents or passages according to their semantic orientation such as positive vs. negative. This direction has been forming the mainstream of research on opinion-sensitive text processing (Pang et al., 2002; Turney, 2002, etc.). The latter, on the other hand, focuses on the task of extracting opinions consisting of information about, for example, ⟨*who* feels *how* about *which aspect* of *what product*⟩ from unstructured text data. In this paper, we refer to this information extraction-oriented task as *opinion extraction*. In contrast to sentiment classification, opinion extraction aims at producing richer information and requires an in-depth analysis of opinions, which has only recently been attempted by a growing but still relatively small research community (Yi et al., 2003; Hu and Liu, 2004; Popescu and Etzioni, 2005, etc.).

Most previous work on customer opinion extraction assumes the source of information to be customer reviews collected from customer review sites (Popescu and Etzioni, 2005; Hu and Liu, 2004; Liu et al., 2005). In contrast, in this paper, we consider the task of extracting customer opinions from unstructured weblog posts. Compared with extraction from review articles, extraction from weblogs is more challenging because weblog posts tend to exhibit greater diversity in topics, goals, vocabulary, style, etc. and are much more likely to include descriptions irrelevant to the subject in question. In this paper, we first describe our task setting of opinion extraction. We conducted a corpus study and investigated the feasibility of the task def-

[*] Currently, NTT Cyber Space Laboratories,
1-1, Hikarinooka, Yokosuka, Kanagawa, 239-0847 Japan

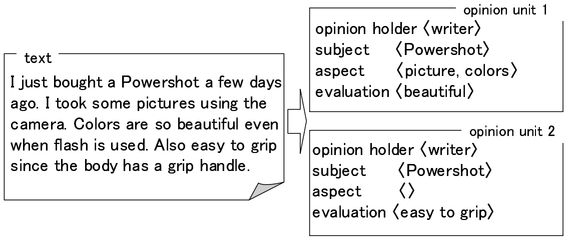
Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1065–1074, Prague, June 2007.

inition by showing the statistics and inter-annotator agreement of our corpus annotation. Next, we show that the crucial body of the above opinion extraction task can be decomposed into two kinds of relation extraction, i.e. aspect-evaluation relation extraction and aspect-of relation extraction. For example, the passage "*I went out for lunch at the Deli and ordered a curry with chicken. It was pretty good*" has an aspect-evaluation relation ⟨*curry with chicken, was good*⟩ and an aspect-of relation ⟨*The Deli, curry with the chicken*⟩. The former task can be regarded as a special type of predicate-argument structure analysis or semantic role labeling. The latter, on the other hand, can be regarded as bridging reference resolution (Clark, 1977), which is the task of identifying relations between definite noun phrases and discourse-new entities implicitly related to some previously mentioned entities.

Most of the previous work on customer opinion extraction, however, does not adopt the state-of-the-art techniques in those fields, relying only on simple proximity-based or pattern-based methods. In this context, this paper empirically shows that incorporating machine learning-based techniques devised for predicate-argument structure analysis and bridging reference resolution improve the performance of both aspect-evaluation and aspect-of relation extraction. Furthermore, we also show that combining contextual clues with a common co-occurrence statistics-based technique for bridging reference resolution makes a significant improvement on aspect-of relation extraction.

2 Opinion extraction: Task design

Our present goal is to build a computational model to extract opinions from Web documents in such a form as: *Who* feels *how* on *which aspects* of *which subjects*. Given the passage presented in Figure 1, for example, the opinion we want to extract is: "*the writer* feels that *the colors* of *pictures* taken with *Powershot* (product) are *beautiful*." As suggested by this example, we consider it reasonable to start with an assumption that most evaluative opinions can be structured as a frame composed of the following constituents:



I just bought a Powershot a few days ago. I took some pictures using the camera. Colors are so beautiful even when flash is used. Also easy to grip since the body has a grip handle.

opinion holder ⟨writer⟩
subject ⟨Powershot⟩
aspect ⟨picture, colors⟩
evaluation ⟨beautiful⟩

opinion holder ⟨writer⟩
subject ⟨Powershot⟩
aspect ⟨⟩
evaluation ⟨easy to grip⟩

Figure 1: Extraction of opinion units

Opinion holder The person who is making an evaluation. An opinion holder is typically the first person (the author). We say the opinion holder is unspecified if the opinion is mentioned as a rumor.

Subject A named entity (product or company) of a given particular class of interest (e.g. a car model name in the automobile domain).

Aspect A part, member or related object, or an attribute (of a part) of the subject on which the evaluation is made (*engine*, *size*, etc.)

Evaluation An evaluative or subjective phrase used to express an evaluation or the opinion holder's mental/emotional attitude (*good, poor, powerful, stylish, (I) like, (I) am satisfied*, etc.)

According to this typology, the example in Figure 1 has six constituents, *the writer* (opinion holder), *Powershot* (subject), *pictures* (aspect), *colors* (aspect), *beautiful* (evaluation), *easy to grip* (evaluation), and constitute two units of opinions as presented in the right half of the figure. We call such a unit an *opinion unit*. In this paper, we only consider explicitly mentioned evaluative opinions as our targets of extraction, excluding opinions indirectly expressed through, for example, style or language choice from our scope.

Under this assumption, opinion extraction can be defined as a task of filling a fixed number of slots as above for each of the evaluations expressed in a given text collection. Two issues then immediately arise. First, it is necessary to make sure that the definition of the opinion units is clear enough for human annotators to be able to carry out the task with sufficient accuracy. Second, all the slots might not consist of simple expressions in that the filler of an aspect slot may have a hierarchical structure in itself. For example, "*the leather cover of the seats (of a car)*" refers to a part of a part of a car. In theory, such a hierarchical chain can be of any length, which may affect the feasibility of the task. For tackling these issues, we built a corpus annotated with the above sort of information and investigated the feasibility of the task.

2.1 Corpus study

We first collected 116 Japanese weblog posts in the restaurant domain by randomly sampling from a collection of posts classified under the “gourmet” category on a major blog site: http://blog.livedoor.com/.

We asked two annotators to annotate them independently of each other following the above specification. The annotators first identified evaluative phrases, and then for each evaluative phrase judged whether it was concerning a particular subject (i.e. a restaurant) in the given domain. If judged yes, the annotators filled the opinion holder and subject slots obligatorily. The annotators filled the aspect slot only when its filler appeared in the document and identified the hierarchical relations between aspects if any (e.g. *noodle* and its *volume*). Note that, if a sentence has two or more evaluations, they have to make one opinion unit for each.

2.1.1 Inter-annotator agreement

We investigated the degree of inter-annotator agreement. In the task of identifying evaluations, one annotator A_1 identified 450 evaluations while the other A_2 identified 392, and 329 cases of them coincided. The two annotators did not identify the same number of evaluations, so instead of using $kappa$ statistics, we use the following metric for measuring agreement as Wiebe et al. (2005) do:

$$agr(A_1||A_2) = \frac{\text{\# of tags agreed by } A_1 \text{ and } A_2}{\text{\# of tags annotated by } A_1}$$

$agr(A_1||A_2)$ was 0.73 and $agr(A_2||A_1)$ was 0.83. The F1 measure of the agreement between the two was therefore 0.79, which indicate that humans can identify evaluation at a reasonable level.

Next, we investigated the inter-annotator agreement of the aspect-evaluation and subject-evaluation relations. Annotator A_1 identified 328 relations, and A_2 identified 346 relations. 295 cases coincided, and $agr(A_2||A_1)$ was 0.90 and $agr(A_1||A_2)$ was 0.86 (F1 measure was 0.88). This shows that we obtained high consistency. Finally, for the subject-aspect and aspect-aspect relations, annotator A_1 identified 296 relations, while A_2 identified 293, 233 cases of which got agreement. $agr(A_2||A_1)$ was 0.79 and $agr(A_1||A_2)$ was 0.80 (F1 measure was 0.79), which show that the human annotators can carry out the task at a reasonable accuracy. Based on this corpus study, we believe that our definitions of two relations are clear enough for constructing annotated corpus.

Table 1: Statistics of opinion-annotated corpus (Restaurant, Automobile, cellular phone and video game)

		Rest	Auto	Phone	Game
	articles	1,356	564	481	361
	sentences	21,666	14,005	11,638	6,448
	# of opinion units	4,267	1,519	1,518	775
I	Asp-Eval	3,692	943	965	521
	Asp-Asp	1,426	280	296	221
	Subj-Asp	2,632	877	850	451
II	Subj-Eval	575	576	553	243
	Subj-Asp-Eval	2,314	736	768	351
	Subj-Asp-Asp-Eval	1,065	175	172	127
	other	313	32	25	54
	Non-writer op. holder	95	17	22	2

2.1.2 Opinion-annotated corpus

Based on these results, we collected a larger set of weblog posts in four domains: restaurant, automobile, cellular phone and video game. We then asked annotator A_1 to annotate them in the same annotation scheme as above. The results are summarized in Table 1. I in the table shows the number of the identified opinion units and relations, and II shows the number of hierarchical chains of aspects. For example, “*Nokia 6800 has a nice color screen*” is counted as “Subj-Asp-Eval” since this example includes a subject “Nokia 6800”, an aspect “color screen” and an evaluation “nice”. “Other” indicates the number of the case where the length of hierarchical chains of aspects is three or more. One observation is that, for all the domains, 90 % of all the opinion units have a hierarchical chain of aspects whose length is two or less. From this, we can conclude that hierarchical chains longer than two are rare, and the problem is not so complicated, though they can be of any length in theory.

The row of “Non-writer op(inion) holder” at the bottom of Table 1 shows the number of opinion units whose opinion holder is *not* the writer of the weblog. This result indicates that when an evaluative expression is found, its opinion holder is highly likely to be the writer of the blogs. Therefore, we put aside the task of filling the opinion holder slot in this paper.

2.2 Related work on task settings of opinion extraction

There are several researches on customer opinion extraction. Hu and Liu (2004) considered the task of extracting ⟨*Aspect, Sentence, Semantic-orientation*⟩ triples in our terminology, where *Sentence* is the one that includes the *Aspect*, and *Semantic-orientation* is either positive or negative.

The notion of Evaluation in our term has also been introduced by previous work (Popescu and Etzioni, 2005; Tateishi et al., 2004; Suzuki et al., 2006; Kobayashi et al., 2005, etc.). For example, our previous paper (Kobayashi et al., 2005) addresses the task of extracting ⟨Subject,Aspect,Evaluation⟩. However, none of those papers reports on such an extensive corpus study as what we report in this paper. In addition, in this paper, we consider not only aspect-evaluation relations but also hierarchical chains of subject-aspect and aspect-aspect relations, which has never been addressed in previous work.

Open-domain opinion extraction is another trend of research on opinion extraction, which aims to extract a wider range of opinions from such texts as newspaper articles (Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Wiebe et al., 2005; Choi et al., 2006). To the best of our knowledge, one of the most extensive corpus studies in this field has been conducted in the MPQA project (Wiebe et al., 2005); while their concerns include the types of opinions we consider, they annotate newspaper articles, which presumably exhibit considerably different characteristics from customer-generated texts.

Though we do not discuss the problem of determining semantic orientation, we assume availability of state-of-the-art methods that perform this task (Suzuki et al., 2006; Takamura et al., 2006, etc.). The problem of determining semantic orientation will be solved by using these techniques, so we focus on the main issue: Extracting opinion units from given texts.

3 Method for opinion extraction

Before designing a model for our opinion extraction task, it is important to note that aspect phrases are open-class expressions and tend to be heavily domain-dependent. In fact, according to our investigation on our opinion-annotated corpus, the number of aspect types is nearly 3,200, and only 3% of them appear in two or more domains as shown in Figure 2. For evaluation expressions, on the other hand, the number of types is much smaller than that of aspect expressions, and 27% of them appear in multiple domains. This indicates that evaluation expressions are more likely to be used commonly across different domains compared with aspects.

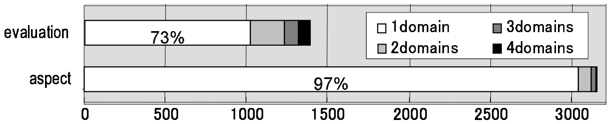


Figure 2: The distributions of evaluation and aspect expressions in the four domains

To prove this assumption, we created a dictionary of evaluation expressions from customer reviews of automobiles (230,000 sentences in total) using the semi-automatic method proposed by Kobayashi et al. (2004). We expanded the dictionary by hand with external resources including publicly available ordinal thesauri. As a result, we collected 5,550 entries. According to our investigation of the coverage by the dictionary, 0.84 (restaurant), 0.88 (cellular phone), 0.91 (automobile), and 0.93 (video game) of the evaluations annotated in our corpus are covered by the dictionary. From this observation, we consider that it is reasonable to start opinion extraction with the identification of evaluation expressions. We therefore design the process of extracting ⟨Subject, Aspect, Evaluation⟩ as follows:

1. **Aspect-evaluation relation extraction**: For each of the candidate evaluations that are selected from a given document by dictionary look-up, identify the target of the evaluation. Here the identified target may be a subject (e.g. *IXY (is well-designed)*) or an aspect of a subject (e.g. *the quality (is amazing)*). Hereafter, we use the term *aspect* to refer to both an aspect and a subject itself, since the subject can be regarded as the top element in the hierarchical chain of aspects.
2. **Opinion-hood determination**: Judge whether or not the obtained pair ⟨aspect, evaluation⟩ is an expression of an opinion by considering the given context. If it is judged yes, go to step3; otherwise, return to step 1 with a new candidate

evaluation expression.

3. **Aspect-of relation extraction**: If the identified aspect is not a subject, search for its antecedent, i.e. an expression that is a higher aspect or a subject of the current aspect. Repeat step 3 until reaching a subject or no parent is found.

3.1 Related work on opinion extraction

A common approach to the customer opinion extraction task mainly uses simple proximity- or pattern-based techniques. For example, Tateishi et al. (2004) implement five syntactic patterns and Popescu et al. (2005) use ten syntactic patterns. Such an approach is limited in two respects. First, it assumes the availability of a list of potential aspect expressions as well as evaluation expressions; however creating such a list of aspects for a variety of domains can be prohibitively expensive because of the domain dependency of aspect expressions. In contrast, our method does not require any aspect lexicon.

Second, their approach lacks the perspective of viewing aspect-evaluation extraction as a specific type of predicate-argument structure analysis, i.e. the task of identifying the arguments of a given predicate in a given text, and fails to benefit from the state-of-the-art techniques of this rapidly growing field. The syntactic patterns used in their research are analyzed by a dependency parser, however, aspect-evaluation relations appear in diverse syntactic patterns, which cannot be easily captured by a handful of manually devised rules.

An exception is the model reported by Kanayama et al.(2004), which uses a component of an existing MT system to identify the "aspect" argument of a given "evaluation" predicate. However, the MT component they use is not publicly available, and even if it were, it would be difficult to apply it to tasks in hand due of the opaqueness of its mechanism. Our approach aims to develop a more generally applicable model of aspect-evaluation extraction.

In open-domain opinion extraction, some approaches use syntactic features obtained from parsed input sentences (Choi et al., 2006; Kim and Hovy, 2006), as is commonly done in semantic role labeling. Choi et al. (2006) address the task of extracting opinion entities and their relations, and incorporate syntactic features to their relation extraction model. Kim and Hovy (2006) proposed a method for extracting opinion holders, topics and opinion words, in which they use semantic role labeling as an intermediate step to label opinion holders and topics. However, these approaches do not address the task of extracting aspect-of relations and make use of syntactic features only for labeling opinion holders and topics. In contrast, as we describe below, we find the significant overlap between aspect-evaluation relation extraction and aspect-of relation extraction and apply the same approach to both tasks, gaining the generality of the model.

Aspect-of relations can be regarded as a sub-type of bridging reference (Clark, 1977), which is a common linguistic phenomenon where the referent of a definite noun phrase refers to a discourse-new entity implicitly related to some previously mentioned entity. For example, we can see a relation of bridging reference between "*the door*" and "*the room*" in "*She entered the room. The door closed automatically.*" A common approach is to use co-occurrence statistics between the referring expression (e.g. "*the door*" in the above example) and the related entity ("*the room*") (Bunescu, 2003; Poesio et al., 2004). Our approach newly incorporates automatically induced syntactic patterns as contextual clues into such a co-occurrence model, producing significant improvements of accuracy.

3.2 Our approach

Now we describe our approach to aspect-evaluation and aspect-of relation extraction. The key idea is to combine the following two kinds of information using a machine learning technique for both tasks.

Contextual clues: Syntactic patterns such as

⟨Aspect⟩-ga *VP-te,* *⟨Evaluation⟩*
⟨Aspect⟩-NOM *VP*-CONJ *⟨Evaluation⟩*

which matches such a sentence as

⟨sekkyaku⟩-ga kunrens-aretei-te ⟨kimochiyoi⟩
⟨service⟩-NOM *be trained*-CONJ *⟨feel comfortable⟩*
(*The waiters were well-trained, so I felt comfortable*.)

are considered to be useful for extracting relations between slot fillers when they appear in a single sentence (Here, ⟨⟩ indicates a slot filler). We employ a supervised learning technique to search for such useful syntactic patterns.

Context-independent statistical clues: Statistics such as aspect-aspect and aspect-evaluation

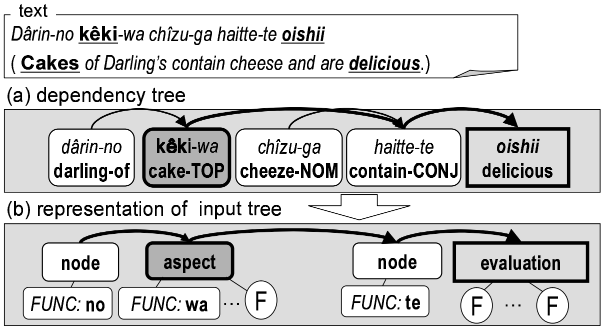


Figure 3: Representation of input data

co-occurrences are expected to be useful. We obtain such statistical clues automatically from a large collection of raw documents.

In what follows, we describe our method for aspect-evaluation. The aspect-of relation extraction is done in an an analogous way.

3.2.1 Supervised learning of contextual clues

Let us consider the problem of searching for the aspect of a given evaluation expression t. This problem can be decomposed into binary classification problems of deciding whether each pair of candidate aspect c and target t is in an aspect-evaluation relation or not. Our goal is to learn a discrimination function for this classification problem. If such a function is obtained, we can identify the most likely candidate aspect simply by selecting the best scored c-t pair and, if its score is negative for all possible candidates, we conclude that t has no corresponding aspect in the candidate set.

For finding syntactic patterns that extract an aspect c starting with an evaluation t, we first represent all the sentences in the annotated corpus that has both an aspect and its evaluation, as shown in Figure 3. A sentence is analyzed by a dependency parser, then the dependency tree is converted so as to represent the relation between content words clearly and to attach other information (such as POS labels and other morphological features of content words and the functional words attached to the content words) as shown in the lower part of Figure 3. Among various classifier induction algorithms for tree-structured data, in our experiments, we have so far examined Kudo and Matsumoto (2004)'s algorithm, packaged as a free software named *BACT*. Given a set of training examples represented as ordered trees labeled either positive or negative class, this algorithm learns a list of weighted decision stumps as a discrimination function with the Boosting algorithm. Each decision stump is associated with tuple $\langle s, l, w \rangle$, where s is a subtree appearing in the training set, l a label, and w a weight of this pattern. The strength of this algorithm is that it automatically acquires structured features and allows us to analyze the utility of features.

Given a c-t pair in an annotated sentence, tree encoding of this sentence is done as follows: First, we use a dependency parser to obtain a dependency tree as in Figure 3 (a). We assume "*kêki* (*cake*)" as the candidate aspect c and "*oishii* (*delicious*)" as the target evaluation t. We then find the path between t and c together with their daughter nodes. For example, the node "*Darling-no* (*Darling's*)" is kept since it is a daughter of c. Then, all the content words are abstracted to either of the class types, evaluation, aspect or node, that is, c is renamed as "aspect", t as "evaluation" and all other content words as "node". Other information of a content word and the information of functional words attaching to the content word are represented as the leaf nodes as shown in Figure 3 (b). The features used in our experiments are summarized in Table 2.

We apply the same method to the aspect-of relation extraction by replacing the "evaluation" label as the second "aspect" label.

3.3 Context-independent statistical clues

We also introduce the following two kinds of statistical clues.

i. Co-occurrences of aspect-evaluation/aspect-aspect: Among various ways to estimate the strength of association (e.g. the number of hits returned from a search engine), in our experiments, we extracted aspect-aspect and aspect-evaluation co-occurrences in 1.7 million weblog posts using the patterns "⟨aspect⟩ *ga/wa/mo* ⟨evaluation⟩ (⟨aspect⟩ *is (subject-marker)* ⟨evaluation⟩)" and "⟨aspect_A⟩ *no* ⟨aspect_B⟩ ga/wa (⟨aspect_B⟩ *of* ⟨aspect_A⟩ *is*)". To avoid the data sparseness problem, we use Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) to estimate conditional probabilities $P(Aspect|Evaluation)$ and $P(Aspect\_A|Aspect\_B)$. We then incorporate the

information of these probability scores into the learning model described in 3.2 by encoding them as a feature that indicates the relative score rank of each candidate in a given candidate set (see Table 2).
ii. Aspect-hood of candidate aspects: Aspect-hood is an index of the degree that measures how plausible a term is used as an aspect within a given domain. We consider that a phrase directly co-occurred with a subject often is likely to be an aspect of the subject, and extract the expression X which appears in the form "*Subject no X* (*X of Subject*)" and the expression Y which appears in the form "X *no* Y". We calculate the aspect-hood of the expressions X and Y by the pointwise mutual information. This score is also used as a features (see Table 2).

3.4 Intra-/inter-sentential relation extraction

Syntactic pattern induction as described in 3.2.1 can apply only when an aspect-evaluation (or aspect-of) relation appears in a single sentence. We therefore build a separate model for inter-sentential relation extraction, which is carried out after intra-sentential relation extraction.

1) Intra-sentential relation identification: Given a target evaluation (or aspect), select the most likely candidate aspect c^* within the target sentence with the intra-sentential model described in 3.2.1. If the score of c^* is positive, return c^*; otherwise, go to the inter-sentential relation extraction phase.

2) Inter-sentential relation identification: Search for the most likely candidate aspect in the sentences preceding the target evaluation (or aspect). This task can be regarded as a zero-anaphora resolution problem. For this purpose, we employ the supervised learning model for zero-anaphora resolution proposed by (Iida et al., 2003).

3.5 Opinion-hood determination

Evaluation phrases do not always extract correct opinion units in a given domain. Consider an example from the digital camera domain, "*The weather was good. so I went to the park to take some pictures*". "*good*" expresses the evaluation for "*the weather*", but "*the weather*" is not an aspect of digital cameras. Therefore, ⟨*the weather, good*⟩ is not an opinion in the digital camera domain. We can consider a binary classification task of judging whether the obtained opinion unit is a real opinion or not in a given domain. In this paper, we conduct a preliminary experiment which uses the opinion-hood determination model learned by Support Vector Machines. We conduct the model using our opinion-annotated corpus. The positive examples are aspect-evaluation pairs annotated in the corpus. The negative examples are artificially generated as follows: We first identify the expression in the evaluation dictionary that appear in our annotated corpus. We then apply the above aspect-evaluation extraction method and get the most plausible candidate aspect. The result is regarded as a negative example if the extracted aspect is not the true aspect. The features we used in our experiments are summarized in Table 2.

4 Experiments

We conducted experiments with our Japanese opinion-annotated corpus to empirically evaluate the performance of our approach. In these experiments, we separately evaluated the models of aspect-evaluation relation extraction, aspect-of relation extraction, and opinion-hood determination.

4.1 Common settings

We chose 395 weblog posts in the restaurant domain from our opinion-annotated corpus described in 2.1, and conducted 5-fold cross validation on that dataset. As preprocessing, we analyzed this corpus using the Japanese morphological analyzer *ChaSen*[1] and the Japanese dependency structure analyzer *CaboCha*[2].

4.2 Models

The results are summarized in Tables 3 and 4. We evaluated the results by recall R and precision P defined as follows

$$\begin{aligned} R &= \frac{\text{correctly extracted relations}}{\text{total number of relations}}, \\ P &= \frac{\text{correctly extracted relations}}{\text{total number of relations found by the system}}. \end{aligned}$$

Note that, in aspect-of relations, we permit ⟨A,C⟩ to be correct when the data includes the chain of aspect-of relations ⟨A,B⟩ and ⟨B,C⟩. Therefore, we merged the intra- and inter-sentential results as shown in Table 4.

[1] http://chasen.naist.jp/

[2] http://chasen.org/˜taku/software/cabocha/

Table 2: Feature list: t denotes a given target (evaluation or aspect) and c a candidate

Features for contextual clues
- Position of c / t in the sentence (beginning, end, other)
- Base phrase distance between c and t (1, 2, 3, 4, other)
- Whether c and t has a immediate dependency relation
- Whether c precedes t
- Whether c appears in a quoted sentence
- Part-of-speech of c / t
- Suffix of c (*-sei*, *-sa* (-ty), etc.)
- Character type of c (*English*, *Chinese*, *Katakana*, etc.)
- Semantic class of c derived from *Nihongo Goi Taikei* (Ikehara et al., 1997).

Features for statistical clues
- Co-occurrence score rank of c (1st, 2nd, 3rd, 4th, other)
- Aspect-hood score rank of c (1st, 2nd, 3rd, 4th, other)

Table 3: The results of aspect-evaluation relation

		intra-sent.	inter-sent.
Patterns	P	0.56 (432/774)	-
	R	0.53 (432/809)	-
Contextual	P	0.70 (504/723)	0.13 (46/360)
	R	0.62 (504/809)	0.17 (46/274)
Contextual +statistics	P	0.72 (502/694)	0.14 (53/389)
	R	0.62 (502/809)	0.19 (53/274)

Table 4: The results of aspect-of relation

	precision	recall
Co-occurrence	0.27 (175/ 682)	0.17 (175/1048)
Contextual	0.44 (458/1047)	0.44 (458/1048)
Contextual+statistics	0.45 (474/1047)	0.45 (474/1048)

The *Contextual* and *Contextual+statistics* models are our proposed models where the former uses only contextual clues (3.2.1) and the latter uses both contextual and statistical clues. We prepared two baseline models, one for each of the above tasks. The *Pattern* model (in Table 3) simulates the pattern-based method proposed by Tateishi at al. (2004), which uses the following patterns: "⟨Aspect⟩ case-particle ⟨Evaluation⟩" and "⟨Evaluation⟩ syntactically depends on ⟨Aspect⟩". The *Co-occurrence* model (in Table 4) simulates the co-occurrence statistics-based model used in bridging reference resolution (Bunescu, 2003): For an aspect expression, we select the nearest candidate that has the highest positive score of the pointwise mutual information regardless of its occurrence (i.e. inter- or intra-sentential). Comparing the *Pattern* (*Co-occurrence*) model with the *Contextual* model shows the effects of the supervised learning with contextual clues, while comparison of the *Contextual* and *Contextual+statistics* models shows the joint effect of combining contextual and statistical clues.

4.3 Results and discussions

As for the aspect-evaluation relation extraction, concerning the intra-sentential cases, we can see that the models using the contextual clues show nearly 10% improvement in both precision and recall. This indicates that the machine learning-based method has a great advantage over the pattern-based approach. Similar results are seen in aspect-of relation extraction. The models using the contextual clues achieved more than 10% improvement in precision and 20% improvement in recall over the co-occurrence statistics-based model. We can say that contextual clues are also useful in aspect-of relation extraction. In comparing the Contextual and Contextual+statistics models, on the other hand, we could get only a slight improvement, which indicates that we need to estimate the statistical clues more precisely. We found that the unsophisticated estimation of the statistical clues was a major source of errors in aspect-of relation extraction, however, this estimation is not so easy since the correct expressions are appeared only once in large data. We are seeking efficient ways to avoid data sparseness problem (e.g. categorize the aspects).

In the aspect-evaluation relation extraction, we evaluated the results against the human annotated gold-standard in a strict manner. However, according to our error analysis, some of the errors can be regarded as correct for some real applications. In the following example, a relation annotated by the human is "*aji (taste)*, *koi-me (strong)*".

misoshiru-wa ⟨*aji*⟩*-ga* ⟨*koi-me*⟩
miso soup-TOP ⟨*taste*⟩-NOM ⟨*strong*⟩
(*The taste of the miso soup is strong*.)

However, there is no harm to consider that "*misoshiru (miso soup)*, *koi-me (strong)*" is also correct. If we judge these cases as correct, the Proposed models achieve nearly 0.8 precision and 0.7 recall, and the baseline model also get 7 % improvement (precision 0.63 and recall 0.6). Based on this result, we consider that we achieved reasonable performance in intra-sentential aspect-evaluation relation extraction.

As Table 3 shows, inter-sentential relation extraction achieved very poorly. In the case of inter-sentential relations, our model tends to rely heavily on the statistical clues, because syntactic pattern features cannot be used. However, our current method for estimating co-occurrence distributions is not sophisticated as we discussed above. We need to seek for more effective use of large scale domain dependent data to obtain better statistics.

We also conducted a preliminary test of the opinion-hood determination model using the features used in aspect-evaluation relation extraction. As a result, we got 0.5 precision and 0.45 recall. Opinion-hood determination problem includes two decisions: whether the evaluation candidate is an opinion or not, and whether the opinion is related to the given domain if the evaluation candidate is an opinion. We plan to use various features known to be effective in the sentence subjectivity recognition task. This task involves challenging problems. For example, sentence (1) includes the writer's evaluation on *shrimps* served at a particular restaurant. In contrast, very similar sentence (2) does not express evaluation since it is a generic description of the writer's taste.

(1) *watashi-wa konomise-no ebi-ga suki-desu*
I the restaurant shrimp like
(*I like shrimps of the restaurant*.)

(2) *watashi-wa ebi-ga suki-desu*
I shrimps like
(*I like shrimps*.)

Thus we need to conduct further investigation in order to resolve this kind of problems.

4.4 Portability of intra-sentential model

We next evaluated effectiveness of the contextual clues learned in the domains to other domains by testing a model trained on the certain domains to other domain. We selected two new domains, cellular phone and automobile, and annotated 290 weblog posts in each domain. For the restaurant domain, we randomly selected 290 posts from the previously mentioned our annotated corpus. We then divide each data set to a training set and a test set so that we had the same amount of training data for each domain. Then we trained a model on the data for each domain, and applied it to each of the three set of data. Table 5 shows the results of the experiment. Compared with the model trained on the same domain, the models trained on different domains exhibited almost comparable performance. This indicates that the contextual clues learned in other domains are effective in another domain, showing the cross-domain portability of our intra-sentential model.

Table 5: Comparing intra-sentential models among three domains (upper: aspect-eval, lower: aspect-of)

test		restaurant	cellular phone	automobile
same	P	0.72 (502/694)	0.75 (522/693)	0.76 (562/738)
dom.	R	0.62 (502/809)	0.63 (522/833)	0.65 (562/870)
other	P	0.73 (468/638)	0.72 (517/710)	0.74 (565/768)
dom	R	0.58 (468/809)	0.62 (517/833)	0.65 (565/870)
same	P	0.43 (139/321)	0.62 (139/224)	0.66 (185/280)
dom.	R	0.59 (139/234)	0.60 (139/230)	0.66 (185/279)
other	P	0.42 (124/293)	0.53 (138/260)	0.59 (195/329)
dom	R	0.52 (124/234)	0.60 (138/230)	0.70 (195/279)

5 Conclusion

In this paper, we described our opinion extraction task, which extract opinion units consisting of four constituents. We showed the feasibility of the task definition based on our corpus study. We consider the task as two kinds of relation extraction tasks, aspect-evaluation relation extraction and aspect-of relation extraction, and proposed a machine learning-based method which combines contextual clues and statistical clues. Our experimental results show that the model using contextual clues improved the performance for both tasks. We also showed domain portability of the contextual clues.

References

R. Bunescu. 2003. Associative anaphora resolution: a web-based approach. In *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, pages 47–52.

Y. Choi, E. Breck, and C. Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 431–439.

H. H. Clark. 1977. *Bridging. Thinking: readings in cognitive science*. Cambridge : Cambridge University Press.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval* (*SIGIR*), pages 50–57.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining* (*KDD*), pages 168–177.

R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

S. Ikehara, M. Miyazaki, S. Shirai A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei (in Japanese)*. Iwanami Shoten.

H. Kanayama and T. Nasukawa. 2004. Deeper sentiment analysis using machine translation technology. In *Proc. of the 20th International Conference on Computational Linguistics*(*COLING*), pages 494–500.

S. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pages 1367–1373.

S. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*.

N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the 1st International Joint Conference on Natural Language Processing* (*IJCNLP*) , pages 584–589.

N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto. 2005. Opinion extraction using a learning-based anaphora resolution technique. In *The Second International Joint Conference on Natural Language Processing* (*IJCNLP*), *Companion Volume to the Proceeding of Conference including Posters/Demos and Tutorial Abstracts*, pages 175–180.

T. Kudo and Y. Matsumoto. 2004. A boosting algorithm for classification of semi-structured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*).

B. Liu, M. Hu, and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International World Wide Web Conference* (*WWW*), pages 342–351.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 79–86.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 339–346.

Y. Suzuki, H. Takamura, and M. Okumura. 2006. Application of semi-supervised learning to evaluative expression classification. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics* (*CICLing*).

H. Takamura, T. Inui, and M. Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*) , pages 201–208.

K. Tateishi, T. Fukushima, N. Kobayashi, T. Takahashi, A. Fujita, K. Inui, and Y. Matsumoto. 2004. Web opinion extraction and summarization based on viewpoints of products. In *IPSJ SIGNL Note 163*, pages 1–8. (in Japanese).

P. D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 417–424.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the third IEEE International Conference on Data Mining* (*ICDM*), pages 427–434.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 129–136.