

Opinion Mining from Web Documents: Extraction and Structurization

Nozomi Kobayashi Nara Institute of Science and Technology
nozomi-k@is.naist.jp

Kentaro Inui (affiliation as previous author)
inui@is.naist.jp

Yuji Matsumoto (affiliation as previous author)
matsu@is.naist.jp

keywords: opinion mining, sentiment analysis, information extraction, relation extraction

Summary

The task of opinion extraction and structurization is the key component of opinion mining, which allow Web users to retrieve and summarize people’s opinions scattered over the Internet. Our aim is to develop a method for extracting opinions that represent evaluation of consumer products in a structured form. To achieve the goal, we need to consider some issues that are relevant to the extraction task: How the task of opinion extraction and structurization should be designed, and how to extract the opinions which we defined. We define an opinion unit consisting of a quadruple, that is, the opinion holder, the subject being evaluated, the part or the attribute in which it is evaluated, and the evaluation that expresses positive or negative assessment. In this task, we focus on two subtasks (a) extracting subject/aspect-evaluation relations, and (b) extracting subject/aspect-aspect relations, we approach each extraction task using a machine learning-based method. In this paper, we discuss how customer reviews in web documents can be best structured. We also report on the results of our experiments and discuss future directions.

1. Introduction

The explosive increase in Web communication has attracted increasing interests in technologies for automatically mining personal opinions from Web documents such as posts on message board and weblogs. Such technologies can be an alternative means for traditional questionnaire-based social or customer research and would also benefit Web 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. In the former, researchers have been exploring techniques for classifying documents or passages according to semantic/sentiment orientation such as positive vs. negative [Dave 03, Pang 04, Turney 02, etc.].

The latter, on the other hand, focuses on the task of extracting opinions consisting of information about particular aspects of interest and the corresponding

sentiment orientation in a structured form from unstructured text data. In this paper, we refer to the information extraction-based task as *opinion extraction*. In contrast to sentiment classification, opinion extraction in general aims at producing richer information useful for in-depth analysis of opinions, which has recently been challenged by a growing research community [Hu 04, Kanayama 04, Kobayashi 05, Popescu 05, etc.].

This task can be decomposed into the following series of subtasks:

- (1) Extract opinions in a structured form
- (2) Determine the semantic orientations (positive, negative or neutral) for each extracted opinion
- (3) Classify the extracted opinions into pre-defined categories (for example, “*delicious noodle*” and “*mild curry*” may be classified into the same category “positive taste”)
- (4) Visualize the classified opinions on, for example, radar charts [Tateishi 04], bar charts [Liu 05]

and so on.

In this paper, we focus on the first subtask and address the following two issues: First, previous work does not sufficiently discuss how customer reviews can be best structured. We reconsider this issue and set up an opinion extraction task based on our corpus study in Section 2. Second, existing methods for opinion extraction tend to rely on relatively simple proximity-based or pattern-based techniques. However, naive methods do not work well, as we demonstrate below, which we consider will become important source for opinion mining.

In section 3, we propose a machine learning-based method, which are portable across domains. We then report the results of our experiments on a domain-restricted set of Japanese weblog posts in Section 4.

2. Opinion extraction: Task design

2.1 Constituents of an opinion

Our present goal is to build a computational model to extract opinions from Web documents in such a form as:

Who feels *what* on *which aspects* of *which subjects*
Here we assume that the *subject* of an opinion is either a consumer product (e.g. a cellular phone model) or a corporate body (e.g. a restaurant, manufacturer, etc.) in a given domain of interest. Given the passage presented in Figure 1, for example, one of the opinions we want to extract is the information that *the writer* feels that *the colors* of *pictures* taken with *Powershot* (product) are *great*.

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:

Opinion holder A person who is making an evaluation (usually, either the author or an unspecified person)

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

Part A part, member or related object of the subject on which the evaluation is made (*engine, interior, etc.*)

Attribute An attribute (of a part) of the subject on which the evaluation is made (*size, color, design, etc.*)

Evaluation An evaluative or subjective phrase used to express an evaluation or the opinion holder's

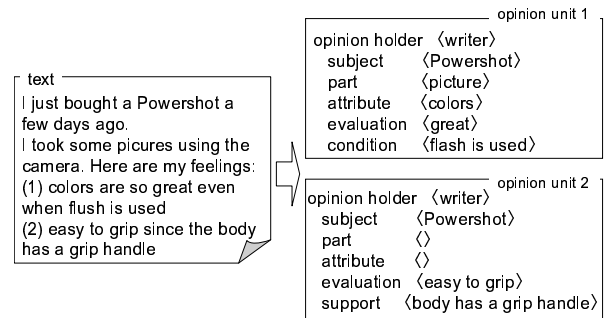


Fig. 1 Extraction of opinion units

mental/emotional attitude (*good, poor, powerful, stylish, (I) like, (I) am satisfied, etc.*)

Condition A condition under which the evaluation applies (*driving on winding roads, when traveling with a family, etc.*)

Support An objective fact or experience described as a supporting factor of the evaluation (*weights nearly 1,500 kg, etc.*)

According to this typology, the example text given in Figure 1 has eight constituents, *the writer* (opinion holder), *Powershot* (subject), *pictures* (part), *colors* (attribute), *great* (evaluation), *easy to grip* (evaluation), *when flash is used* (condition), and *body has a grip handle* (support), which we consider to constitute two units of opinion as illustrated in Figure 1. We call each unit an *opinion unit*.

Under this assumption, opinion extraction can be defined as the 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 human annotators can carry out the task with sufficient accuracy. Second, the filler of a part or attribute slot may have hierarchical structure in itself. For example, “*the leather cover of the steering wheel*” 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 these issues, we built a corpus annotated with the above information and investigated the feasibility of the task. In what follows, we report on the results of our corpus study and design an opinion extraction task based on them.

2.2 Corpus study and task definition

We first collected 116 Japanese weblog posts in the restaurant domain by randomly sampling from the posts classified under the “gourmet” category on the

livedoor blog site ^{*1}. A majority of the sampled posts included descriptions about the writer’s experience and evaluation regarding certain restaurants.

We asked two annotators to annotate them independently of each other according to the above definition. One annotator (S) was a doctoral program student engaged in research on opinion extraction, while the other was an adult person (A) who did not have expertise in natural language processing.

In the annotation process, every evaluative or subjective phrase was considered as a *candidate* evaluation phrase and, for each candidate evaluation phrase, each annotator was asked to judge whether it constituted an opinion unit or not. If judged yes, a candidate evaluation phrase was associated with a new opinion unit whose slots were to be filled. For each opinion unit, the annotators were asked to identify the opinion holder and the subject while being allowed to leave other slots open if there are nothing appropriate.

Here, we slightly simplified the structure of an opinion unit — we merged the part and attribute slots together. We call the merged slot the *aspect* slot. We did it because we had found, in our preliminary trial, that it is considerably difficult to make a clear distinction between parts and attributes. For example, the phrase *buffet* is used to refer to a physical object belonging to a restaurant, while it may also be used to refer to a function of a restaurant. In the former case, the phrase *buffet* should fill the part slot, while, in the latter, it may be interpreted as an attribute. However, this kind of judgment is sometimes extremely hard.

Consequently, the annotators filled the opinion holder, subject and evaluation slots obligatorily, while filling the aspect, condition and support slots optionally. They were also asked to identify hierarchical relations between aspects (e.g. *noodle* and its *volume*), if any. The detailed specifications of the annotation scheme are available at http://cl.naist.jp/~nozomi-k/op-tagged_corpus/.

§ 1 Inter-annotator agreement

We then investigated the degree of the inter-annotator agreement. For the task of identifying evaluations, one annotator (A) identified 450 evaluations, while the other (S) 392 evaluations, 329 cases of which got agreement. Two annotators did not identify the same number of evaluation, so we use the following metric for measuring agreement as [Wiebe 05] does:

$$agr(A||B) = \frac{\# \text{ of tags agreed by A and B}}{\# \text{ of tags annotated only by A}}$$

This metric corresponds to the recall if A’s annotation is always correct, and to precision, if they are reversed. $agr(A||S)$ was 0.73 and $agr(S||A)$ was 0.83, which indicate that the human can identify evaluation at some reasonable level. Next, we investigated the inter-annotator agreement of the subject-aspect and aspect-aspect relations whose evaluation slot had agreement. Annotator A identified 296 relations, while the other 293, 233 cases of which got agreement. $agr(S||A)$ was 0.79 and $agr(A||S)$ was 0.80. From these results, we consider that the human annotators can carry out the task at a certain level of accuracy.

§ 2 Opinion-tagged corpus

Based on these results, we next collected a larger set of weblog posts for four domains, restaurant, automobile, cellular phone and video game, and asked annotator A to annotate them in the same annotation scheme as above.

We collected Japanese weblog posts from the restaurant domain by randomly sampling from the posts classified under the “gourmet” category on the livedoor blog site, and for the automobile, cellular phone, and video game, we collected weblog pages by issuing subject names as queries to a weblog search engine. The results are summarized in Table 1. One observation is that, for all the domains, the hierarchical chain of the aspects is longer than two (Subj-Asp-Asp-Eval) in only less than 10% of all the opinion units. From this, we can conclude that hierarchical chains of aspects are unlikely to be too complicated to handle.

The row of “Opinion holder” in Table 1 shows the number of opinion units whose opinion holder is *not* the writer of the document. The results indicate that when an evaluative description is found, its opinion holder is highly likely to be the writer of the document, which suggests that identification of opinion holder is not a hard problem.

Table 1 also shows that the occurrence of supports and conditions is not as frequent as one may expect. While we are aware that supports and conditions, if any, may well provide important information for opinion analysis, we should conclude from the statistics that it is practical to put a higher priority of research on the task of filling the other four slots: opinion holder, subject, aspect and evaluation.

*1 <http://blog.livedoor.com/>

Table 1 Statistics of opinion-tagged corpus (Rest: restaurant, Auto: automobile, Phone: cellular phone and Game: video game)

	Rest	Auto	Phone	Game
articles	1,356	564	481	361
sentences	21,666	14,005	11,638	6,448
Asp-Eval	3,692	943	965	521
Asp-Asp	1,426	280	296	221
Subj-Asp	2,632	877	850	451
Opinion holder	95	17	22	2
Support	68	86	80	95
Condition	113	86	76	41
# of opinion units	4,267	1,519	1,518	775
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

§3 Task definition

Based on this corpus study, we consider an opinion extraction task as follows:

Given a text collection, extract opinions and structure them in the form of quadruple $\langle \textit{Opinion holder}, \textit{Subject}, \textit{Aspect}, \textit{Evaluation} \rangle$, where *Subject* and *Evaluation* are obligatory while *Aspect* is optional and may have a hierarchical chain.

The followings are examples.

- (1) *I hear that the ipod is very good.*
 $\rightarrow \langle \textit{unspecified person}, \textit{ipod}, \phi, \textit{good} \rangle$
- (2) *I got Canon G3 and am amazed at the quality of photos.*
 $\rightarrow \langle \textit{the writer}, \textit{Canon G3}, \langle \textit{photos}, \textit{quality} \rangle, \textit{be amazed} \rangle$
- (3) *Nokia 6800 has a nice color screen.*
 $\rightarrow \langle \textit{the writer}, \textit{Nokia 6800}, \textit{color screen}, \textit{nice} \rangle$

2.3 Related work

One of the early work taking the information extraction approach to opinion extraction is reported by [Hu 04]. The task they consider is extraction of $\langle \textit{Subject}, \textit{Aspect}, \textit{Semantic-orientation} \rangle$ triples in our terms, where *Semantic-orientation* is binary-valued, either *positive* or *negative*. Therefore, our task setting can be considered as a refinement of theirs in that we consider hierarchical chains of aspects, which may be filled with phrases even from separate sentences, and we also consider the evaluation slot to be filled with an evaluation phrase rather than a binary value of sentiment orientation. In addition, more importantly, while Hu et al. [Hu 04] discusses little about the appropriateness and feasibility of their task, our task definition is grounded on a corpus study.

Perhaps, our task setting is closest to what is considered by [Popescu 05]. They additionally annotate [Hu 04]’s corpus with tags corresponding to our evaluation slots. Their paper, however, also lacks discussion based on corpus analysis.

To our best knowledge, one of the most extensive corpus studies in this field is being conducted in the MPQA project [Wiebe 05]; however, their concerns are not focused on the types of opinions we consider, and they annotate newspaper articles, which presumably exhibit a quite different distributions from weblog posts.

3. Method for opinion extraction

3.1 Task redefined and resource availability

As discussed in 2.2, our opinion extraction task is now recast as the task of filling the slots of $\langle \textit{Opinion holder}, \textit{Subject}, \textit{Aspect}, \textit{Evaluation} \rangle$. Among these slots, we put aside the task of filling the opinion holder slot in this paper because the filler of this slot is highly likely to be the writer of the document as noted in 2.2.2. Furthermore, we consider identification of candidate subjects (e.g. product names) as a separate task, which has been intensively studied over a decade as the task of named entity recognition. Assuming the availability of state-of-the-art models of named entity recognition, in the experiments we report on in the next section, we manually labeled all the product names appearing in our opinion-tagged corpus as candidate subjects.

Aspect phrases are open-class words and tend to be heavily domain-dependent. In fact, according to our investigation, we found that only 10 expressions, such as “*balance, color, price*”, appeared across four domains, accounted for only 1% of the whole aspect expressions in each domain. Given this, it is not realistic to assume the availability of any list of aspect expressions applicable to a wide range of domains with a broad coverage. One important issue, therefore, is how to identify aspects without any predefined list of candidate aspect expressions.

Evaluation phrases, on the other hand, are much more likely to be used commonly across different domains. To prove this assumption, we actually constructed a dictionary of evaluation expressions from automobile reviews (230,000 sentences in total) using the semi-automatic method proposed in [Kobayashi 04]. We expanded the dictionary to include entities by hand from external resources such as publically avail-

able ordinal thesauri. As a result, we collected 5,550 entries, which is now available from http://cl.naist.jp/~nozomi-k/evaluative_expressions.html.

According to our investigation of the coverage for the dictionary, approximately 80% of the evaluation phrases annotated in the restaurant corpus are covered by the dictionary. Note that to build the dictionary, we used the reviews from the automobile domain but not any document from the restaurant domain. This result supports our assumption about the availability of an open-domain lexicon of evaluation expression. In our experiments, we used the aforementioned evaluation dictionary.

Given these considerations about the resource availability, we design the process of extracting ⟨Subject, Aspect, Evaluation⟩ as follows:

1. **Aspect-evaluation relation extraction:** For each of the candidate evaluation that are selected from a given document by dictionary look-up, identify the object of the evaluation. Here the identified object may be an aspect of an opinion subject (e.g. *the quality (is amazing)*) but may also be an opinion subject itself (e.g. *Canon G3 (is well-designed)*). Hereafter, we use the term *aspect* to refer to both an aspect of an opinion subject and an opinion subject itself.
2. **Opinion-hood determination:** Judge whether the obtained pair ⟨Aspect, Evaluation⟩ is an expression of an opinion or not by considering the given context. If it is, go to step 3; return to step 1, otherwise.
3. **Aspect-of relation extraction:** If the identified aspect is not an opinion subject, search for its parent, i.e. the object whose part or attribute is the current aspect. Repeat step 3 until reaching an opinion subject or no parent is found.

3.2 Existing techniques for relation extraction

Approaches to the aspect–evaluation extraction task mainly use simple proximity- or pattern-based techniques. Popescu *et al.* [Popescu 05] approached this problem by filtering out the non-aspect candidates using automatically extracted aspect expressions. However, as we show in Section 4, these naive methods do not work well on weblog texts since such patterns may also match non-aspect candidates in a given domain. In light of these problems, we apply machine learning to this task using syntactic and statistical information instead of a dictionary to provide useful clues for

estimating the *aspect-hood* of candidates.

[Kanayama 04] considers an opinion as a tuple of a semantic orientation, a predicate, and its argument, and apply the idea of transfer-based machine translation to the task. We also consider the aspect–evaluation extraction as the task of finding an evaluation’s arguments. While we agree with Kanayama *et al.*’s framing of this task, an important unresolved issue is zero-anaphora resolution, since predicate arguments are often not explicitly expressed in Japanese sentences. Our work includes aspect-of relation extraction even when aspect and evaluation appear in different sentences – both cases that are not addressed in [Kanayama 04]. As we explain in Section 4, aspect–evaluation occur 75% in the intra-sentential cases, and 62% of the cases have syntactic dependency relations. In the aspect-of relations, on the other hand, only 22% of the relations are intra-sentential, and 42% of the cases have syntactic dependency relations. This observation suggest that including inter-sentential relations significantly increases our coverage.

Aspect-of relations can be regarded as a subtype of bridging reference [Clark 77]. Bridging reference is the referent of a definite description implicitly related to some previously mentioned entity. For example, we can see a relation of bridging reference between “*the door*” and “*the room*” in the sentences “*She entered the room. The door closed automatically.*” In recent work on the bridging reference, [Bunescu 03] and [Poesio 04] use the number of web pages which contain both the referring expression (e.g. “*the door*” in the above example) and the mentioned entity (e.g. “*the room*”) being queried as a measure of the strength of association. In our work, one important clue is whether the candidate is relevant to the domain. Therefore, we consider not only the strength of association for the aspect-aspect relations, but also the strength of the co-occurrence between aspects and the domain.

3.3 Our approach

The key idea for our relation extraction subtasks is to combine the following two kinds of information.

- Contextual clues: Syntactic patterns such as “*I ⟨Evaluation⟩ ⟨Aspect_B⟩ of the ⟨Aspect_A⟩ (I ⟨like⟩ ⟨the color⟩ of ⟨the phone⟩)*”, are considered to be useful for extracting relations between slot fillers when they appear in a single sentence. We employ a supervised learning technique to search for useful contextual clues.

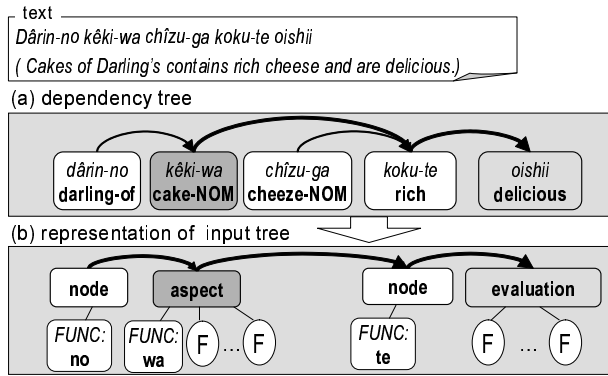


Fig. 2 Representation of input data

- Context-independent statistical clues: For example, the statistics of aspect-aspect and aspect-evaluation co-occurrences which are extracted to be useful clues. We obtain such statistical clue automatically from a large document collection.

We use a machine-learning technique also to learn how to combine these different kinds of clues.

§ 1 Supervised learning of contextual clues

We use supervised learning to obtain useful syntactic patterns. There are various methods of incorporating syntactic patterns into machine learning models, such as dependency kernels, tree kernels and so on. We have so far examined the boosting-based algorithm proposed by [Kudo 04]. This algorithm (implemented as the package *BACT*) outputs a list of decision stumps from training examples represented as labeled ordered trees. The score is calculated by the sum of the score of the subtrees included in the input instance. In what follows, we explain how to represent an example as a labeled ordered tree.

First, we use a dependency parser to obtain dependency parse trees (Figure 2 (a), in which “*kêki*” and “*oishii*” are an aspect–evaluation pair). Next, we extract the path from an evaluation to its aspect, and extract their daughter nodes as shown in Figure 2 (b). In the example, the path between the evaluation “*oishii*” and the aspect “*kêki*” includes the node “*koku-te*” on the path as well as the daughter node of the aspect, i.e. “*dârin-no*”. We delete the content words included in the nodes while function words are remained. For example, in the node “*kêki-wa*” in Figure 2 (b), the content word “*kêki*” is removed, and a functional word “*wa*” is remained. Besides those functional words, we add the features summarized in Table 2, for the aspect or the evaluation node shown as *F* in Figure 2.

§ 2 Unsupervised learning of context-independent statistical clues

We consider four kinds of statistical clues:

- Aspect-evaluation co-occurrences
- Aspect-aspect co-occurrences
- Aspect-hood of candidate aspects
- Statistical inference of aspect-aspect relation classes

i. Aspect-evaluation/aspect-aspect

co-occurrences Among various ways to estimate co-occurrence (e.g. the number of hits returned from a search engine), in the experiments we report below, we counted aspect-aspect and aspect-evaluation co-occurrences in 1.7 million weblog posts using the patterns

- “{aspect} *ga/wa/mo* {evaluation} (*is* {evaluation})”
- “{aspect_A} *no* {aspect_B} *ga/wa*” (*of* {aspect_A} *is*)”

To avoid the data sparseness problem, we use the Probabilistic Latent Semantic Indexing (PLSI) [Hofmann 99]. We can calculate the joint probability $P(A, B)$ even if A and B do not directly co-occur, since PLSI assumes a set of latent class of co-occurrence:

$$P(A, B) = \sum_{z \in Z} P(A|z)P(B|z)P(z)$$

where Z denotes a set of latent class of co-occurrence. We can calculate pointwise mutual information, conditional probabilities, etc. from the estimated distribution $P(A, B)$.

ii. Aspect-hood of candidate aspects Aspect-hood is an index of the degree to which the term is used as an aspect within a given domain. First, we extract the expression X which appear in the form “subject *no* X ”, and then extract the expression Y which appear in the form “ X *no* Y ”. We calculate the aspect-hood of the expressions X and Y based on pointwise mutual information [Manning 99].

$$PMI(X, Y) = \log_2 \frac{N \times count(X, Y)}{count(Y) \times count(X)}$$

where $count(X, Y)$ is the number of times X occurred in Y , and N is the total number of times all pairs occurred in the form “ X *no* Y ”. $count(X)$ (or $count(Y)$) is the number of X (or Y) occurred.

iii. Statistical inference of aspect-aspect relation classes Aspect-aspect co-occurrences are good clues for extracting aspect-of relations. However, many other types of relations can hold between two nouns which appear in “ A *no* B ” form. It is not clear whether the two nouns have aspect-of relation or not. For example, “*watashi no kuruma* (my car)” and “*kuro no*

seifuku (the black uniform)” appear in the form of “*A no B*”, however, there are no aspect-of relations since the former relation indicates possession and the latter represents a property (color) of the uniform. Therefore, it is important to estimate the relation classes between two expressions.

For this problem, we create the model to estimate the aspect-of relation using the maximum entropy model*2. We created labeled data, which consists of pair of nouns, annotated with ‘aspect-of’ and ‘other’ relation tags, and learned the model with the features: verbs or adjectives co-occurred with *A* or *B* and the semantic classes of *A* or *B* derived from the Japanese thesaurus “Nihongo Goi Taikei” [Ikehara 97]. The size of the labeled data is nearly 5,300, half of the data is “aspect-of” and the remains is “other”.

3.4 Extraction of aspect-evaluation relations

Syntactic patterns can only apply if the aspect and evaluation both appear in the same sentence. We therefore build separate components for intra-sentential and inter-sentential extraction tasks:

- 1) Intra-sentential aspect identification: For a given evaluation, select the most likely candidate aspect *c* within the evaluation sentence using the intra-sentential model. When we learn this model, we encode the above co-occurrence informations as the features. If the identifier decide that none of the candidates are not aspect, search for the aspect to preceding sentences.
- 2) Inter-sentential aspect identification: For a given evaluation, select the most likely candidate aspect from the sentences preceding the evaluation using the inter-sentential model. This task can be recast as a zero-anaphora resolution problem. For this purpose, we employ the supervised learning model for zero-anaphora resolution proposed by [Iida 03]. When we learn this model, we also use the above co-occurrence informations as the features.

The specific features we used in the experiments are summarized in Table 2.

3.5 Opinion-hood determination

Evaluation phrases does not always express opinion holders’ evaluation. Consider an example from the digital camera domain, “*The weather was good, so I went to the park to take some pictures of roses*”.

The evaluation phrase “*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 which we aim to extract. We can consider that the task of judging whether the obtained opinion unit is a real opinion or not in a given domain is a binary classification task. We introduce the opinion-hood determination model learned by Support Vector machines. The specific features we used in the experiments are summarized in Table 3.

3.6 Extraction of aspect-of relations

We also approach the aspect-of relation extraction by decomposing it into two subtasks (explained in Section 3.4), and build a separate component proposed in aspect-evaluation pair extraction problem.

4. Experiments

We conducted experiments with our Japanese opinion-tagged corpus in the restaurant domain to empirically evaluate the performance of our approach. In these experiments, we evaluated the models of aspect-evaluation relation extraction, opinion-hood determination, and aspect-of relation extraction, separately. Evaluating the performance of the entire task is our future work.

4.1 Common settings

§1 Training/Evaluation corpus

We chose 395 weblog posts in the restaurant domain from our opinion-tagged corpus we described in Section 2.2. As preprocessing, we analyzed the opinion-tagged corpus using the Japanese morphological analyzer *ChaSen**3 and the Japanese dependency structure analyzer *CaboCha**4.

§2 Classifier

We used *BACT**5 for the the intra-sentential models, and Support Vector Machines with 2nd order polynomial kernel for the inter-sentential, and opinion-hood determination models.

§3 Features

We summarize the features used for train and test the models in Table 2. We used these features to intra-sentential and inter-sentential models for both aspect-evaluation and aspect-of cases. For the aspect-of relation extraction, we used the result of the statis-

*2 We use a package MaxEnt
<http://maxent.sourceforge.net/>

*3 <http://chasen.naist.jp/>

*4 <http://chasen.org/~taku/software/cabocho/>

*5 <http://chasen.org/~taku/software/bact/>

Table 2 Features: t denotes the target (evaluation or aspect) and c denotes the candidate

Feature type	Description
Contextual clues	Distance between t and c Existence of a direct dependency relation between t and c Position of t and c in the sentence Whether t precedes c or not Part-of-speech of c and t Whether c appears in the quoted text
Context-independent	Whether the candidate is directly co-occur with the subject The score of the aspect-hood of c Rank of the $P(c t)$ indices of c in the candidate set The label and the score of the relation between c and t mentioned in Section 3.2 (for aspect-of) Suffix of c (e.g., -sei, -sa(-ty, -ity)) Character type of t and c (e.g., <i>katakana</i> , <i>English alphabet</i>) Semantic class of c derived from Nihongo Goi Taikai[Ikehara 97]

Table 3 Features for opinion-hood determination: e denotes the evaluation and c denotes the extracted candidate

Feature type	Description
Context-independent	Whether the candidate is directly co-occur with the subject The score of the aspect-hood of c Suffix of c (e.g., -sei, -sa(-ty, -ity)) Character type of t and c (e.g., <i>katakana</i> , <i>English alphabet</i>) Semantic class of c derived from Nihongo Goi Taikai[Ikehara 97]

tical inference of aspect-aspect relation classes mentioned in Section 3.3.

For opinion-hood determination, we used the some kind of features shown in Table 3, in addition to the contextual clues described in Table 2

4.2 Models

The results are summarized in Table 4, where five models are compared for each of the two subtasks: aspect-evaluation relation extraction and aspect-of relation extraction. The following is a summary of each model for the former subtask:

Baseline A-E model simulates the algorithm proposed by [Tateishi 04]:

1. If there are any candidate aspects which match the following extraction patterns:
-⟨Aspect⟩ *ga/wa/mo/no/ni/wo/de* ⟨Evaluation⟩
-⟨Evaluation⟩ syntactically depends ⟨Aspect⟩
choose the nearest one as the aspect of the evaluation
2. Otherwise, choose the candidate aspect with the highest aspect-evaluation co-occurrence score.

Context-only A-E model: uses contextual pattern-based clues (3.2.1) but not statistical clues (3.2.2) and works in the manner as described in 3.3.

Proposed A-E model: uses both contextual and statistical clues together by encoding the aspect-

evaluation co-occurrence score and the aspect-hood score (3.2.2.i) as a set of additional features to the tree-representation (Fig. 2) of a given input.

Proposed-MI A-E model: uses the same clues as the proposed A-E model except that it uses point-wise mutual information as the aspect-evaluation co-occurrence score instead of the conditional probabilities of aspect-evaluation co-occurrence, which is used in the proposed model.

Proposed-dic A-E model: incorporates an aspect expression dictionary in the Proposed A-E model instead of automatically calculated aspect-hood scores. The aspect expression dictionary was manually created for the restaurant domain containing 6,129 expressions.

Comparing the Baseline model with the Context-only model shows the effects of the supervised learning of contextual pattern features, while a comparison of the Context-only and Proposed models shows the joint effects of combining contextual and statistical clues. The performance of the Proposed-dic model provides an estimation of the upper-bound of the improvements that could be gained by accurate estimation of the aspect-hood of each candidate aspect.

The Baseline model (the Baseline Aspect-of model) we implemented for aspect-of relation extraction relies only on the aspect-aspect co-occurrence score,

which simulates the method for bridging reference resolution proposed by [Bunescu 03]:

1. Select the expression which has highest scores of pointwise mutual information, if there are any candidate in the sentence which the aspect appear.
2. Otherwise, choose the nearest one which co-occur with the aspect.

The other four models for aspect-of relation extraction were created analogously to the above A-E models. The Proposed Aspect-of model uses the information of the statistically estimated aspect-aspect relation class for each candidate aspect in addition to the aspect-aspect co-occurrence score and the aspect-hood score.

4.3 Evaluation

We conducted 5 fold cross validation using all the data, and evaluated the results by recall R and precision P defined as follows

$$R = \frac{\text{correctly extracted relations}}{\text{total number of relations}},$$

$$P = \frac{\text{correctly extracted relations}}{\text{total number of relations found by the system}}.$$

4.4 Results and discussions

Table 4 shows the result of aspect-evaluation and aspect-of relation extraction tasks, and opinion-hood determination. In both aspect-evaluation and aspect-of relation extraction tasks, we can see that our models outperform the baseline models in both recall and precision.

As for the aspect-evaluation relation extraction, no significant improvement was observed among four non-baseline models. However, concerning the inter-sentential cases, we can see constant improvement according to the quantity of information used in the models, showing that the context-independent information of the candidate and co-occurrence statistics are important clues for finding the aspect expressions appearing beyond sentence boundaries.

For the aspect-of relation extraction, our models achieve more than 10% improvement in precision, and 20% improvement in recall over the baseline model. Also, there is significant improvement in the Proposed model compared with the Context-only model. As far as the experiments show, the point-wise mutual information score does not give better performance than that the conditional probability score for co-occurrence measurement. Although there is still much room for improvement, the notable difference

between the Proposed and Proposed-dic models shows that accurate estimation of the aspect-hood of candidate aspect expressions has a potential effect.

One of the reasons of low performance of aspect-of relation extraction is that the evaluation criteria is a bit too strict. The extracted aspect-aspect and subject-aspect relations are evaluated against the human annotated gold-standard in a strict manner. For example, when the gold-standard data includes the chain of aspect-of relations A-B and B-C, and the system extracts aspect-of relation A-C, it is evaluated as incorrect. In some application domains this kind of skipping aspect-of relation may not raise a severe issue. If we assume that A-C is correct, the precision and the recall in our Proposed aspect-of model increase to 0.45. While the Proposed-dic model achieve nearly 10% improvement in both recall and precision.

Table 5 compares the difference in the cases where the candidate expressions in aspect-evaluation or aspect-of relation are syntactically dependent. In the table, the column “with dependency” means that the pair of expressions have direct modification relation, and the column “no dependency” means that they are syntactically non-dependent. While it is quite natural that the precision is much higher when they are syntactically dependent each other, our models achieves about 40 % precision in the cases where the candidates have no syntactic relation with the other.

Opinion-hood determination also posts a challenging problem. For example, sentence (1) includes the writer’s evaluation on the *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. The wording is, however, so similar that our models have difficulty in learning the difference.

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

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

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

4.5 Domain-portability of the model

To investigate portability of the proposed model, we applied the model learned in the restaurant domain to the cellular phone domain, and vice versa. We used 290 weblog posts in the cellular phone domain from

Table 4 The results of aspect-evaluation (A-E) relation, aspect-of relation and opinion-hood determination

			intra-sentential	inter-sentential	total	
A-E	Baseline	precision	0.56 (432/774)	0.08 (20/235)	0.45 (452/1009)	
		recall	0.53 (432/809)	0.07 (20/274)	0.42 (452/1083)	
	Context-only	precision	0.70 (504/723)	0.13 (46/360)	0.51 (550/1083)	
		recall	0.62 (504/809)	0.17 (46/274)	0.51 (550/1083)	
	Proposed	precision	0.72 (502/694)	0.14 (53/389)	0.51 (555/1083)	
		recall	0.62 (502/809)	0.19 (53/274)	0.51 (555/1083)	
	Proposed-MI	precision	0.70 (505/682)	0.14 (55/401)	0.51 (560/1083)	
		recall	0.62 (505/809)	0.20 (55/274)	0.51 (560/1083)	
	Proposed-dic	precision	0.80 (482/600)	0.17 (83/477)	0.52 (565/1083)	
		recall	0.60 (482/809)	0.30 (83/274)	0.52 (565/1083)	
	aspect-of	Baseline	precision	0.25 (79/312)	0.21 (79/370)	0.23 (158/682)
			recall	0.34 (79/234)	0.10 (79/814)	0.15 (158/1048)
Context-only		precision	0.41 (122/297)	0.30 (222/748)	0.33 (344/1045)	
		recall	0.52 (122/234)	0.27 (222/814)	0.33 (344/1048)	
Proposed		precision	0.43 (139/321)	0.34 (247/814)	0.37 (386/1045)	
		recall	0.59 (139/234)	0.30 (247/814)	0.37 (386/1048)	
Proposed-MI		precision	0.38 (147/387)	0.33 (222/660)	0.35 (369/1047)	
		recall	0.62 (147/234)	0.27 (222/814)	0.35 (369/1048)	
Proposed-dic		precision	0.52 (145/281)	0.42 (319/761)	0.45 (464/1042)	
		recall	0.62 (145/234)	0.39 (319/814)	0.44 (464/1048)	
opinion-hood determination	precision			0.51 (488/949)		
	recall			0.45 (488/1083)		

Table 5 Results of Intra-sentential cases

	with dependency	no dependency
Context-only A-E	0.78 (391/501)	0.37 (113/308)
Proposed A-E	0.77 (385/501)	0.38 (117/308)
Context-only aspect-of	0.69 (68/98)	0.4 (54/136)
Proposed aspect-of	0.71 (70/98)	0.51 (69/136)

our opinion-tagged corpus.

Tables 6 and 7 show the results of the experiment. Comparing with the baseline model, we can see that our proposed method gets better precision without losing the recall, even when the training data is taken from different domain from the test data. This indicates that the contextual clues learned in a domain are effective to another domain, showing the portability of our proposed model.

On the other hand, in the inter-sentential case, the performance drops by nearly 10%. The reason could be attributed to poor performance in the aspect-hood estimation. One of the important key techniques we need to investigate is an effective way of estimating the aspect-hood of terms.

5. Conclusion

In this paper, we discussed the task of structuring opinions, and introduced the opinion units consisting of four constituents: opinion holder, subject, aspect (part or attribute), evaluation. We identified the

task of opinion extraction as the relation extraction tasks and proposed a machine learning-based method which does not use any domain-specific aspect dictionary. Though our experimental results indicate the difficulty of the opinion extraction task, our models outperform the baseline models in both recall and precision. We also showed the contextual clues learned in a domain are effective to another domain.

One of future work is to find the useful information to improve our extraction model. Another task is to identify subject expressions. This is a subclass of named entity recognition, which has been actively studied for a decade. We are planning to incorporate state-of-the-art techniques for named entity recognition in the overall opinion extraction system we are currently developing.

◇ References ◇

- [Bunescu 03] Bunescu, R.: Associative anaphora resolution: a web-based approach, in *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, pp. 47–52 (2003)

Table 6 Comparison between two domains (aspect-evaluation)

		test		testing on restaurant		testing on cellphone			
				intra	inter	intra	inter		
baseline	precision	0.56	(432/774)	0.08	(20/235)	0.54	(445/826)	0.06	(14/233)
	recall	0.53	(432/809)	0.07	(20/274)	0.53	(445/833)	0.04	(14/361)
training on restaurant	precision	0.72	(502/694)	0.14	(53/389)	0.70	(454/645)	0.06	(36/547)
	recall	0.62	(502/809)	0.19	(53/274)	0.55	(454/833)	0.10	(36/361)
training on cellphone	precision	0.70	(460/659)	0.07	(31/424)	0.75	(522/693)	0.20	(99/499)
	recall	0.57	(460/809)	0.11	(31/274)	0.63	(522/833)	0.27	(99/361)

Table 7 Comparison between two domains (aspect-of)

		testing on restaurant		testing on cellphone					
		intra	inter	intra	inter				
baseline	precision	0.25	(79/312)	0.21	(79/370)	0.30	(83/279)	0.11	(32/298)
	recall	0.34	(79/234)	0.10	(79/814)	0.36	(83/230)	0.06	(32/577)
training on restaurant	precision	0.43	(139/321)	0.34	(247/814)	0.52	(128/244)	0.06	(35/563)
	recall	0.59	(139/234)	0.34	(247/814)	0.56	(128/230)	0.06	(35/577)
training on cellphone	precision	0.53	(136/257)	0.12	(98/798)	0.62	(139/224)	0.30	(172/583)
	recall	0.58	(136/234)	0.12	(98/814)	0.60	(139/230)	0.30	(172/577)

- [Clark 77] Clark, H. H.: *Bridging. Thinking: readings in cognitive science*, Cambridge : Cambridge University Press (1977)
- [Dave 03] Dave, K., Lawrence, S., and Pennock, D. M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in *Proc. of the 12th International World Wide Web Conference*, pp. 519–528 (2003)
- [Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999)
- [Hu 04] Hu, M. and Liu, B.: Mining and summarizing customer reviews, in *Proc. of the Tenth International Conference on Knowledge Discovery and Data Mining*, pp. 168–177 (2004)
- [Iida 03] Iida, R., Inui, K., Takamura, H., and Matsumoto, Y.: Incorporating Contextual Cues in Trainable Models for Coreference Resolution, in *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*, pp. 23–30 (2003)
- [Ikehara 97] Ikehara, S., Miyazaki, M., A. Yokoo, S. S., Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y.: *Nihongo Goi Taikai (in Japanese)*, Iwanami Shoten (1997)
- [Kanayama 04] Kanayama, H. and Nasukawa, T.: Deeper Sentiment Analysis Using Machine Translation Technology, in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 494–500 (2004)
- [Kobayashi 04] Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T.: Collecting Evaluative Expressions for Opinion Extraction, in *Proc. of the 1st International Joint Conference on Natural Language Processing*, pp. 584–589 (2004)
- [Kobayashi 05] Kobayashi, N., Iida, R., Inui, K., and Matsumoto, Y.: Opinion extraction using a learning-based anaphora resolution technique, in *The Second IJCNLP, Companion Volume to the Proceeding of Conference including Posters/Demos and Tutorial Abstracts*, pp. 175–180 (2005)
- [Kudo 04] Kudo, T. and Matsumoto, Y.: A Boosting Algorithm for Classification of Semi-Structured Text, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2004)
- [Liu 05] Liu, B., Hu, M., and Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, in *Proceedings of the 14th International World Wide Web Conference (WWW)*, pp. 342–351 (2005)
- [Manning 99] Manning, C. D. and Schütze, H.: *Foundations of statistical natural language processing*, MIT press (1999)
- [Pang 04] Pang, B. and Lee, L.: A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 271–278 (2004)
- [Poesio 04] Poesio, M., Mehta, R., Maroudas, A., and Hitzenman, J.: Learning to resolve bridging references, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)* (2004)
- [Popescu 05] Popescu, A.-M. and Etzioni, O.: Extracting product features and opinions from reviews, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 339–346 (2005)
- [Tateishi 04] Tateishi, K., Fukushima, T., Kobayashi, N., Takahashi, T., Fujita, A., Inui, K., and Matsumoto, Y.: Web opinion extraction and summarization based on viewpoints of products, in *IPSJ SIGNL Note 163*, pp. 1–8 (2004), (in Japanese)
- [Turney 02] Turney, P. D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424 (2002)
- [Wiebe 05] Wiebe, J., Wilson, T., and Cardie, C.: Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation*, Vol. 39, pp. 165–210 (2005)

〔担当委員：麻生 英樹〕

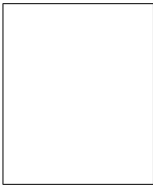
Received August 15, 2006.

Author's Profile

Kobayashi, Nozomi

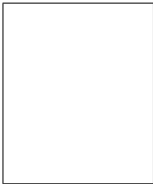
Nozomi Kobayashi received her master degree of engineering from Nara Institute of Science and Technology in 2004. She has been a student of Graduate School of Information Science, Nara Institute of Science and Technology. Her research interests include sentiment analysis and

information extraction.



Inui, Kentaro (Member)

Kentaro Inui received his doctoral degree of engineering from Tokyo Institute of Technology in 1995. He has then experienced an assistant professor at Tokyo Institute of Technology and an associate professor at Kyushu Institute of Technology. He has been an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology since 2001.



Matsumoto, Yuji (Member)

Yuji Matsumoto received his M.S. and Ph.D. degrees in information science from Kyoto University in 1979 and in 1989. He joined Machine Inference Section of Electrotechnical Laboratory in 1979. He has then experienced an academic visitor at Imperial College of Science and Technology, a deputy chief of First Laboratory at ICOT, and an associate professor at Kyoto University.

He has been a professor of Graduate School of Information Science, Nara Institute of Science and Technology since 1993. His main research interests are natural language understanding and linguistic knowledge acquisition.