Opinion Extraction based on Syntactic Pieces[*]

Suguru Aoki and Kazuhide Yamamoto

Nagaoka University of Technology,
1603-1, Kamitomioka, Nagaoka, Niigata 940-2188 Japan
{aoki, ykaz}@nlp.nagaokaut.ac.jp

Abstract. This paper addresses a task of opinion extraction from given documents and its positive/negative classification. We propose a sentence classification method using a notion of *syntactic piece*. Syntactic piece is a minimum unit of structure, and is used as an alternative processing unit of n-gram and whole tree structure. We compute its semantic orientation, and classify opinion sentences into positive or negative. We have conducted an experiment on more than 5000 opinion sentences of multiple domains, and have proven that our approach attains high performance at 91% precision.

Keywords: syntactic piece, opinion extraction, sentence classification

1. Introduction

One can easily disseminate information through the Internet, that include their personal opinions, such as reputation and dissatisfaction with products, complaints about services, and so on. Weblogs and message boards in particular have attracted a great deal of attention as a new information source, since they enable us to obtain subjective opinions easily. In order to automatically extract useful information from these sources, various approaches have been proposed. (Inui and Okumura, 2006)

Researchers have been exploring techniques for classifying documents according to sentiment orientation, or positive/negative (p/n) in particular. Turney (2002) extracts phrases containing adjectives or adverbs, and determines their semantic orientation. Further, p/n of a document is judged by computing the average of the semantic orientations. The *NEAR* operator in the Alta Vista search engine is used in the method. If you search a query like *A NEAR B* in Alta Vista, the search engine shows pages containing within near words of each other. However, this operator is not provided in Japanese version.

Wang and Araki (2007) extend Turney's method into Japanese by collecting and using a set of words that contribute significantly to p/n orientation. Fujimura et al. (2004) use corpora divided into p/n words, and statistically classify a document by extracting opinions. These methods provide the semantic orientation only with bag-of-words. A word, however, contains little and partial information. Therefore, we assert that the scope of the process needs to be expanded. Moreover, these methods do not identify reasons for p/n judgment, that is, expressions that cause semantic orientation. For a commercial application, this function is essential in marketing research.

Besides these, other document classification approaches have also been proposed, which prepare a semantic orientation dictionary in advance. Tateishi et al. (2004) construct this

[*] Copyright 2007 by Natural Language Processing Laboratory, Department of Electrical Engineering, Nagaoka University of Technology.

dictionary in advance by extracting opinion triplets that consist of `object name, attribute expression, evaluative expression', and classify documents by using the triplets. The method extracts only expressions that appear in a definite pattern, and it is thus difficult to obtain satisfactory coverage and accuracy. Kobayashi et al. (2005) first extract a few opinion pairs of attribute and value with an anaphora resolution technique, and construct a semantic orientation dictionary of the pairs for a target domain. They then gradually expand the entries of the dictionary from the pair seed. However, if the accuracy of the primary pairs is not adequate, the quality of the dictionary becomes gradually poorer due to gradually involving noises. In addition, making the dictionary domain by domain is very expensive.

In this paper we do not use a word as an unit of tagging information to extract opinions. As other work does, we also think that a document can be classified when we only extract partial sentiment expressions. However, we do not think that bag-of-words approach is not suitable for this task, and we need something else instead. One can see this fact that, for example, the semantic orientation of a word can vary according to a given domain. Let us consider this example; a word `big' is positive when used in sentence such as `this LCD monitor is *big.*', while the word should be judged negative in a sentence such as `that portable audio player is too *big* to me.'

Consequently, we assert that it is necessary to use a longer unit as a sentiment expression instead of using an unit of word in which conventional works do so. In this paper, we propose an opinion-mining method that utilize a new notion of *syntactic piece*. Syntactic piece is a unit of sentiment expression that is suitable for keeping semantic orientation. More explanation of syntactic piece is described later.

2. Syntactic Piece

Syntactic piece and other units are illustrated in Figure 1.

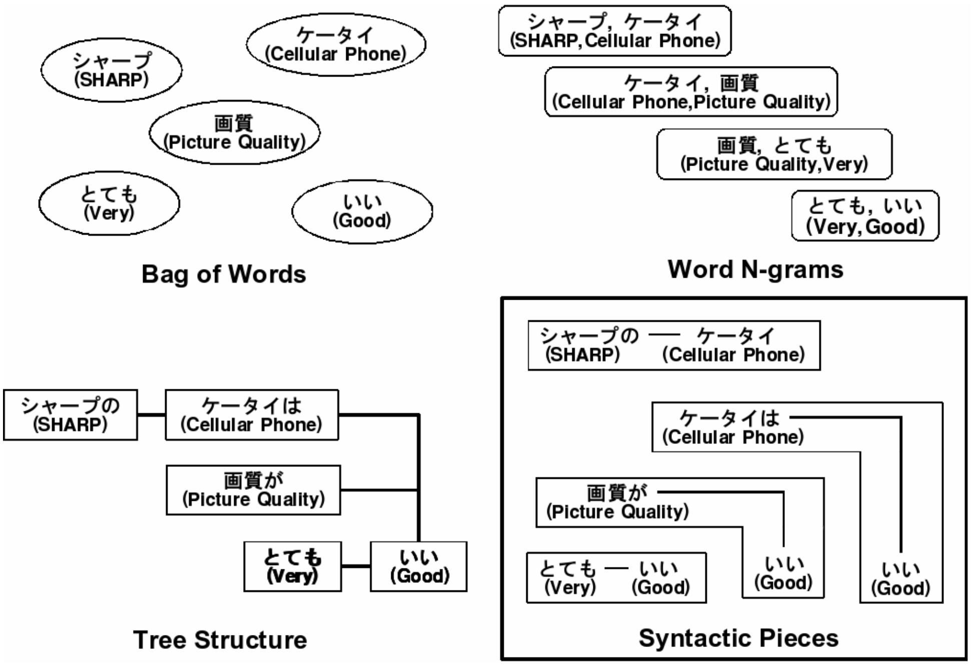

Figure 1: Idea of syntactic piece compared with other units

Syntactic piece is a minimum unit of syntactic structure of an expression. It is defined as a pair consisting of a modifier and a modifiee (modified entity) from dependency analysis result. This pair is expressed as follows.

syntactic piece : modifier ⇒ modifiee

Syntactic piece has several characteristics:

- it is very simple; it is easy, just like n-gram statistics, to extract pieces from any given expression since it requires only (partial) parsing result. In contrast, using the whole syntactic structure for opinion extraction is computationally very expensive. Consequently, we think the notion of syntactic piece is considered to have advantages of both n-gram and (whole) tree structure.
- it contains far more information than n-gram. N-gram is a consecutive sequence that keeps information of local context. It is observed that agglutinative languages such as Japanese and Korean has enormous combination of word sequence, since the word order is relatively free. In such languages n-gram-based model is expected not to work well, and some kinds of syntax should be dealt with.
- it can deal with a chunk of meaning, such as a phrasal idiom; e.g. a Japanese phrase `気-に ⇒ なる' that means `feel uneasy.' Conventional methods have avoided this problem by (1)ignoring them, or (2) importing an idiom dictionary from outside. In contrast, our method gives them the same treatment, hence we do not need such idiom dictionary or we do not need to recognize idioms as they are.

Here we present Japanese patterns of the syntactic piece below:

continuous modification

- **case frame** : noun(-particle) ⇒ predicate
  e.g. 画面-が ⇒ きれい (clear screen)
- **adverbial modification** : adverb ⇒ predicate
  e.g. とても ⇒ おいしい (delicious)

adnominal modification

- **noun modification** : noun-no ⇒ noun
  e.g. キャノン-の ⇒ カメラ (Canon's camera)
- **verbal modification** : verb ⇒ noun
  e.g. くつろげる ⇒ 店 (comfortable shop)
- **adjectival modification** : adjective ⇒ noun
  e.g. おいしい ⇒ ケーキ (delicious cake)
- **compound noun** : noun-noun
  e.g. 携帯-電話 (cellular phone)
- **prefix** : adverb ⇒ prefix-noun
  e.g. 高-画質 (high picture quality)

3. Method

Our opinion extraction model is illustrated in Figure 2.

To begin with, our system extracts syntactic pieces[1] from a training corpus. We then compute a semantic orientation score for each piece, and construct a seed dictionary of the pieces. We then generalize the dictionary by increasing the entries of the pieces that are labeled according to information of the existing pieces. We extend the dictionary employing large texts

[1] We may just call *piece* for short hereafter.

of general domains. Finally, we classify a sentence with the dictionary into three types: positive, negative or other.

In this paper we assume that we have a training corpus in which semantic orientation of positive or negative tags are labeled for all sentences. There are no chance that both positive and negative tags are labeled to the same sentence. We also assume that the large amount of texts of general domains, such as newspaper corpus, is available.

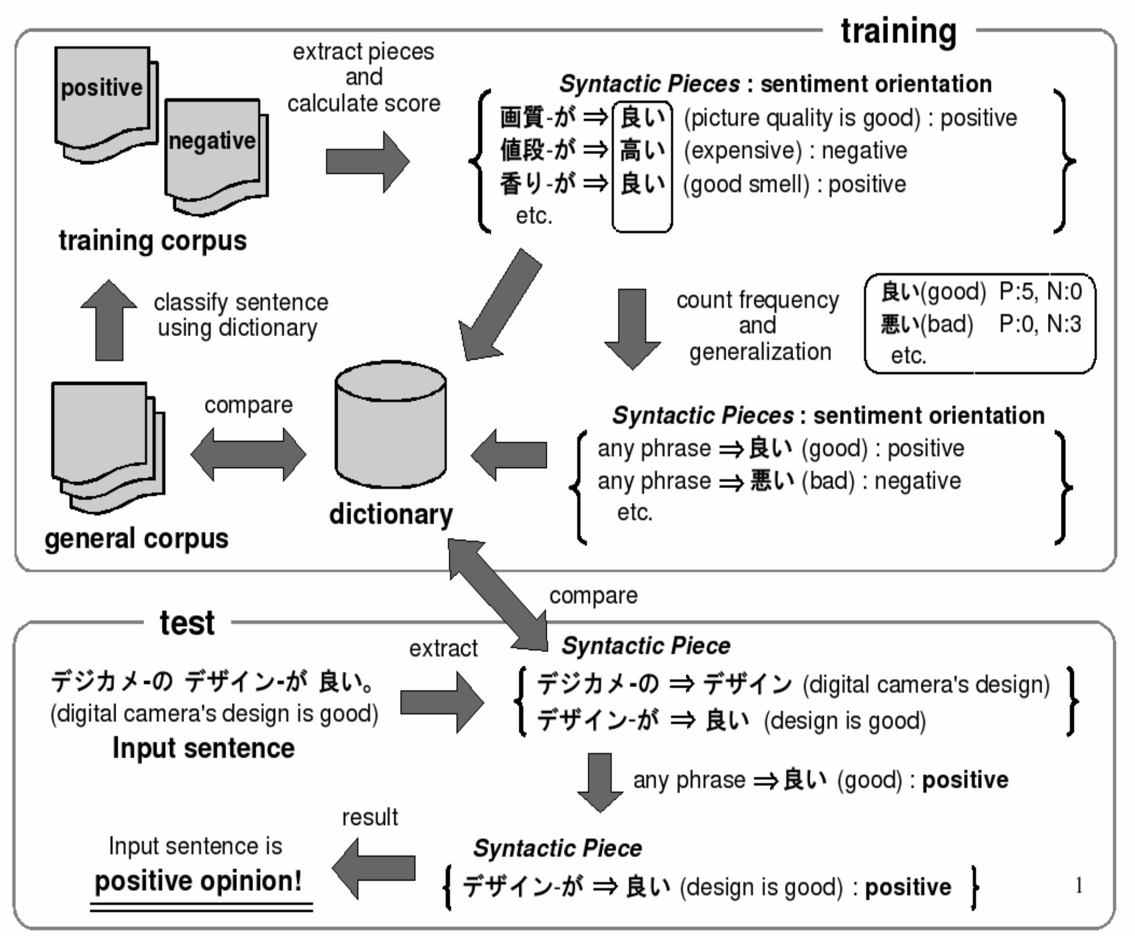
training
positive
negative
training corpus
extract pieces and calculate score
Syntactic Pieces : sentiment orientation
画質-が ⇒ 良い (picture quality is good) : positive
値段-が ⇒ 高い (expensive) : negative
香り-が ⇒ 良い (good smell) : positive
etc.
classify sentence using dictionary
count frequency and generalization
良い(good) P:5, N:0
悪い(bad) P:0, N:3
etc.
compare
general corpus
dictionary
Syntactic Pieces : sentiment orientation
any phrase ⇒ 良い (good) : positive
any phrase ⇒ 悪い (bad) : negative
etc.
compare
test
extract
デジカメ-の デザイン-が 良い。
(digital camera's design is good)
Input sentence
Syntactic Piece
デジカメ-の ⇒ デザイン (digital camera's design)
デザイン-が ⇒ 良い (design is good)
any phrase ⇒ 良い (good) : positive
result
Input sentence is
positive opinion!
Syntactic Piece
デザイン-が ⇒ 良い (design is good) : positive
1

Figure 2: Opinion extraction model

3.1. Syntactic Piece Extraction

First of all, a sentence in the training corpus is analyzed by a dependency parser that creates the dependency structure. We then pick up pairs of modified item and a modifier, for all modifications in the dependency structure. Each pair constitutes a syntactic piece. Sometimes two modifiers, say A and B, modify the same expression, say C. In this case two syntactic pieces, i.e. (A ⇒ C) and (B ⇒ C), are created.

When a given sentence has a positive tag, all syntactic pieces extracted from a given sentence are tagged positive. The same is done for negative case.

As mentioned, we do not need sentences which have no semantic orientation since we use syntactic pieces to classify the sentence into p/n. Moreover, although compound noun is regarded as noun modification of noun, that is the target of syntactic piece extraction, they are not extracted since compound noun seems to have no semantic orientation.

3.2. Semantic Orientation Score

We then compute the semantic orientation score for each syntactic piece extracted from the corpus using a formula seen in Fujimura et al. (2004).

According to Fujimura, a word that has positive semantic orientation should appear in a positive opinion. The same can be said to syntactic piece. Based on this hypothesis, we compute the differences of frequency between positive opinions and negative opinions. If the syntactic piece does not have semantic orientation, the frequency of positive opinions should be the same as the frequency of negative opinions. A positive piece is expected to have a positive value.

Semantic orientation scores for all syntactic pieces are given in the following procedure.

1. prepare text collection each of which has either p or n,
2. extract all syntactic pieces from all texts,
3. count the frequency of p/n opinion for each piece, and
4. calculate the semantic orientation score with the following equation:

$$score(piece_i) = \frac{P(piece_i) - N(piece_i)}{P(piece_i) + N(piece_i)} \qquad (1)$$

$$\left(-1 \leq score(piece_i) \leq 1 \right)$$

where *piece*$_i$ is a syntactic piece, *score(piece*$_i$*)* is sentiment orientation score of *piece*$_i$, *P(piece*$_i$*)* is frequency of *piece*$_i$ appeared in positive opinions, and *N(piece*$_i$*)* is frequency of appeared *piece*$_i$ in negative opinions. By conducting this process we can automatically construct a syntactic-piece dictionary that has semantic orientation score.

In the conventional methods such as Turney (2002) and Fujimura et al. (2004), document classification is based on statistically extracted keywords, as well as tagged semantic orientation for each word n-gram. Hence, the method depends on the domain of the learning corpus, that is, semantic orientation of a word may change with the domain. For example, a word `高い' in `値段-が ⇒ 高い' (expensive in terms of money) is considered to be negative in general, while the same word in `画素数-が ⇒ 高い' (high picture quality) is considered to be positive.

In contrast, the semantic orientation of syntactic piece is expected to be independent of the domain of learning corpus. This is because, semantic orientation of a syntactic piece lengthens the context than using an unit of word, that enables us clearer semantic orientation. This feature is important since it is not required for our proposed method to consider (i.e. classify and/or identify) domain; it is enough that we prepare pieces with semantic orientation in any domain as many as possible.

3.3. Dictionary Generalization

We first provide syntactic pieces that are extracted from a training corpus. Next we use this seed to label other pieces. Compared with labeling p/n information to words, we need to tag more since the number of syntactic pieces are far more than number of words in general. Therefore it is necessary to extend given information to other unlabeled pieces. Here we will explain how we do that.

As we have mentioned before, semantic orientation of a word may change with a domain in many cases. However, we all know that some words always show only p or n. For example, we see that sentiment orientation of a word `良い(good)' is always positive, even if we do not know what is good. Hence, we generalize the dictionary by identifying such syntactic pieces as `良い'.

We automatically label p/n as follows. When we collect pieces that constitute the same first element (i.e. modifier), and only positive tags are observed in the set, we tag all pieces that has the same first element positive. The similar procedure is conducted for negative tags, and also for the same second element (i.e. modified item) set.

We show how to do this process. Suppose that there are six syntactic pieces that has semantic orientation in the dictionary.

`画質-が⇒良い’	`味-が⇒良い’	`画面-が⇒大きい’	: tagged **positive**
(picture quality is good)	(taste good)	(big screen)	
`騒音-が⇒大きい’	`デザイン-が⇒悪い’	`印象-が⇒悪い’	: tagged **negative**
(very noisy)	(poor design)	(bad impression)	

	positive	negative	
`*any phrase* ⇒ 良い’	2	0	⇒ tagged **positive**
`*any phrase* ⇒ 大きい’	1	1	⇒ not extracted
`*any phrase* ⇒ 悪い’	0	2	⇒ tagged **negative**

3.4. Sentence Classification

The semantic orientation of a sentence is determined between p/n by extracting syntactic pieces in the given sentence. If syntactic piece appears more than once, the sentence score is calculated by summing up the score of each syntactic piece in the sentence.

$$sentence\_score(S) = \sum_{piece_i \subset S} score(piece_i) \tag{2}$$

$$\begin{cases} if \quad sentence\_score(S) > 0 \Rightarrow positive \\ \quad sentence\_score(S) = 0 \Rightarrow not \quad opinions \\ \quad sentence\_score(S) < 0 \Rightarrow negative \end{cases} \tag{3}$$

Here, $piece_i$ is a syntactic piece in a sentence *S*, and *sentence_score(S)* is its sentence score. The final judgment of semantic orientation is either positive, negative, or neutral (i.e., the sentence is not an opinion sentence). If there is no syntactic piece extracted in the given sentence, the sentence is judged not to be an opinion sentence. And if, the syntactic piece in the sentence is not in the dictionary, the syntactic piece score is 0.

3.5. Dictionary Extension

We have two types of dictionaries: the seed dictionary and the generalized seed dictionary. Although the size of the seed dictionary is small, it is not easy in general to increase learning corpus, when we want to extend the amount of the seed dictionary. Hence, what we do here is to classify other large corpus using the dictionaries to make learning data.

First, we provide a large corpus. This corpus is a collection of raw texts and do not require any tags such as positive or negative. We classify sentences in the corpus using the seed dictionary and the the generalized seed dictionary into either positive, negative or other. We use the classes of positive and negative as tagged data, and the other class is not used. We extract syntactic pieces out of the tagged data, and calculate semantic orientation score from the method explained in section 3.1 and 3.2 using this positive and negative tagged corpus. This is what we call the extended dictionary. Finally, we generalize this extended dictionary.

4. Experiment

In this experiment we have prepared a training corpus that consists of a variety of 13 domains and 5,608 sentences. Statistics of the corpus are presented in Table 1. And we use a general corpus that consists of Weblog texts and million sentences to extend dictionary.

As pre-processing, we analyzed the corpus using CaboCha[1)], a Japanese dependency analyzer. Evaluation was performed by 13-fold cross validation, a way of unseen input evaluation, using all domains. The corpus was divided into thirteen domains. We classified sentences in the test data using the dictionary, and evaluated the results by precision and recall values.

Table 1: Number of sentences in training corpus

domain	positive	negative	total
digital camera	533	238	771
PC	112	100	212
soft drink	559	90	649
services	185	271	456
MP3 player	364	231	595
printer	103	177	280
cellular phone	156	73	229
designer goods	221	46	267
shampoo	478	173	651
beer	748	161	909
video game	61	52	113
cosmetics	44	12	56
sweets	322	98	420
total	3886	1722	5608

5. Results and Discussions

Table 2 illustrates the effect of the dictionary extension and generalization. We see from the table that precision is very high, even if we generalize the dictionary. Although we expect before the experiment that the precision sharply or gradually declines with the dictionary generalization, but it has unexpectedly increased. This may imply the robustness of our method against the noise given by the dictionary generalization. One possibility for this surprising result is that there may be low possibility that the syntactic pieces are wrongly tagged thanks to the longer unit.

Other issue is that high precision is attained without domain specification in our method. This should also thank longer processing unit, since it decreases possibility of interference that enables coexistence and no need to switch domains.

On the other hand, the recall is not satisfactory at this time. However, we think that this is an encouraging result since there is an possibility to further extend the dictionary with keeping this precision high, according to the discussion above. Actually, even though we used general corpus consists of weblog, precision and recall have increased.

Table 2: The precision and the recall for sentiment classification

dictionary	precision	recall
seed only	0.85 (752/888)	0.13 (752/5608)
seed + generalization	0.86 (2423/2809)	0.43 (2423/5608)
extended seed	0.82 (1033/1257)	0.18 (1033/5608)
extension + generalization	0.91 (3046/3338)	0.54 (3046/5608)

Table 3 and 4 show an example syntactic piece dictionary. The seed dictionary has about ten thousand pieces. After the dictionary extension, pieces increased to 130,000. We observe that the syntactic piece score was almost `1’ or `-1’. This means that the semantic orientation of syntactic piece is aptly separated into p and n. Therefore, this fact have also proven that a unit of the syntactic piece has little ambiguity in deciding semantic orientation.

Table 3: An example of positive sentiment orientation tagged syntactic piece dictionary

pattern	**syntactic piece**
case frame	コンテンツ−が⇒充実 (contents is enriched) 好感−を⇒持てる (favorable impression) デザイン−が⇒かわいい (design is cute) 動作−が⇒速い (response is quick) 心地⇒良い (feel good)
verbal modification	暖まる⇒エピソード (heart warming episode) 楽しむ⇒方法 (way to enjoy)
adverbial modification	とっても⇒きれい (very beautiful) かなり⇒コンパクト (very compact)
adjectival modification	いい⇒香り (good smell) 高い⇒品質 (high quality) すごい⇒お洒落 (very stylish)
prefix	新−商品 (new product) 省−スペース (small space) 高−機能 (highly functional)

Table 4: An example of negative sentiment orientation tagged syntactic piece dictionary

pattern	**syntactic piece**
case frame	画質−が⇒良い−ない (picture quality is not good) 使い勝手−が⇒悪い (usability is bad) 消耗−が⇒激しい (very waste) サイズ−が⇒小さい (size is small) 気持ち⇒悪い (feel sick)
verbal modification	違う⇒商品 (different item)
adverbial modification	すぐ⇒壊れる (break at once) かなり⇒高額 (very extensive)
adjectival modification	ぬるい⇒ビール (lukewarm beer) 物足りない⇒感じ (not good enough)
prefix	異−音 (noise) 再−起動 (reboot) 非−表示 (no display)

Table 5 shows the result according to patterns of syntactic piece. We see from Table 5 that case frame patterns and adverbial modification pattern almost fill the system output. Weblogs are unstructured data, it is thus not always true that the attribute is written in the sentence; it may also be written before the sentence instead. It is adverbial modification pattern that accounts for a large percentage of system output. Normally, if contextual processing is conducted, or if an anaphora analysis is performed, we have to identify the attribute. However, we are not concerned with this problem here.

Table 5: The precision and the recall for each pattern of syntactic piece

pattern	**precision**	**recall**
case frame	0.82 (417/506)	0.07 (417/5608)
adverbial modification	0.85 (290/340)	0.05 (290/5608)
verbal modification	0.88 (59/67)	0.01 (59/5608)
adjectival modification	0.85 (69/81)	0.01 (69/5608)
prefix	0.67 (16/24)	0.00 (16/5608)

Table 6 shows an example of the generalized dictionary. The generalized seed dictionary has about 7,800 pieces. After the dictionary extension, generalized pieces increased to 33,000. We see that most of the generalized pieces are reasonable, such as positive for beautiful, good taste, and easy-to-use, and negative for no good, bad taste, and troublesome. This observation also supports the high precision.

Table 6: An example of the generalized dictionary

semantic orientation	**syntactic piece**
positive	*any phrase* ⇒キレイ (beautiful) *any phrase* ⇒使い−やすい (easy to use) *any phrase* ⇒美味しい (good taste) 飲み−やすい (easy to drink) ⇒*any phrase*
negative	*any phrase* ⇒良い−ない (no good) *any phrase* ⇒使い−にくい (hard to use) *any phrase* ⇒まずい (bad taste) いまひとつ (unattractive) ⇒*any phrase* 不具合−が (trouble) ⇒*any phrase*

Table 7 shows the results of each domain using extended dictionary. From results shown in Table 7, it is clear that high precision is obtained regardless of domains. It is also important that as the size increases, the precision also increases.

Table 7: The precision and the recall for sentiment classification using extended and generalized dictionary

domain	**precision**	**recall**
digital camera	0.84 (408/484)	0.53 (408/771)
PC	0.90 (109/121)	0.51 (109/212)
soft drink	0.92 (406/441)	0.63 (406/649)
services	0.88 (206/233)	0.45 (206/456)
MP3 player	0.91 (317/350)	0.53 (317/595)
printer	0.91 (117/129)	0.42 (117/280)
cellular phone	0.96 (130/136)	0.57 (130/280)
designer goods	0.95 (156/164)	0.58 (156/267)
shampoo	0.91 (326/358)	0.50 (326/651)
beer	0.96 (544/567)	0.60 (544/909)
video game	0.89 (59/66)	0.52 (59/113)
cosmetics	1.00 (37/37)	0.66 (37/56)
sweets	0.92 (231/252)	0.55 (231/420)

6. Conclusion

This paper presents a new method of opinion extraction. The novel feature of our method is use of its treating unit; *syntactic piece*. Compared with other units such as a single word, n-gram, and a whole tree structure, a notion of syntactic piece collects several advantages in total that others have. Our proposed method achieves high precision, and is domain-independent. Moreover, our approach is able to clearly identify the reasons for positive or negative semantic orientation by observing tags of the pieces.

Although the low recall is observed throughout the experiment this time, several ways are considered to improve recall rate. The biggest issue is that we need to further label pieces by somehow generalizing information in the seed pieces. This is an exciting items and we will tackle this task first as a future work.

List of Tools Used in this Work

1) CaboCha, Ver.0.53, Matsumoto Lab., Nara Institute of Science Technology. http://chasen.org/~taku/software/cabocha/

References

Shigeru Fujimura, Masashi Toyota, and Masaru Kitsuregawa. 2004. A Consideration of Extracting Reputations and Evaluative Expressions from the Web. *Technical Report of the Institute of Electronics, Information and Communication Engineers*, 104:144-146.

Takashi Inui and Manabu Okumura. 2006. Research Trend about Opinions Analysis of the Text. *Journal on Natural Language Processing*, 13(3):201-241.

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2005. Opinion Extraction Using a Learning-Based Anaphora Resolution Technique. In *Proceedings of the Second International Joint Conference on Natural Language* Processing, pp. 175-180.

Kenji Tateishi, Toshikazu Fukushima, Nozomi Kobayashi, Tetsuro Takahashi, Atsushi Fujita, Kentaro Inui, and Yuji Matsumoto. 2004. Web opinion extraction and summarization based on viewpoints of products. *Information Processing Society of Japan SIGNL Note*, 2004(93):1-8.

Peter 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*, pp. 417-424

Guangwei Wang and Kenji Araki. 2007. Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 189-192.