Natural Language Engineering **18** (1): 83–107. © Cambridge University Press 2011
doi:10.1017/S1351324911000131

Unsupervised lexicon induction for clause-level detection of evaluations

HIROSHI KANAYAMA and
TETSUYA NASUKAWA

IBM Research – Tokyo, 1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken 242-8502, Japan
e-mail: hkana@jp.ibm.com, nasukawa@jp.ibm.com

(*Received 22 December 2009; revised 18 November 2010; accepted 4 February 2011; first published online 30 March 2011*)

Abstract

This article proposes *clause-level evaluation detection*, which is a fine-grained type of opinion mining, and describes an unsupervised lexicon building method for capturing domain-specific knowledge by leveraging the similar polarities of sentiments between adjacent clauses. The lexical entries to be acquired are called *polar atoms*, the minimum human-understandable syntactic structures that specify the polarity of clauses. As a hint to obtain candidate polar atoms, we use *context coherency*, the tendency for the same polarity to appear successively in a context. Using the overall density and precision of coherency in the corpus, the statistical estimation picks up appropriate polar atoms from among the candidates, without any manual tuning of the threshold values. The experimental results show that the precision of polarity assignment with the automatically acquired lexicon was 83 per cent on average, and our method is robust for corpora in diverse domains and for the size of the initial lexicon.

1 Introduction

Collecting peoples' opinions, particularly evaluations of products, services or brands, is becoming more important in the commercial activities of both companies and individual consumers, because they want to sort out such opinions from multiple perspectives due to the diversification of items and people's various preferences. Due to the emergence of massive online text resources such as product review articles and messages in bulletin boards on the WWW, now the deeper analysis of people's opinions from text documents is not only feasible, but even necessary, especially for enterprises.

Beyond estimating the overall preference for a product or a service by a writer, detecting the individual utterances that specify each positive or negative property is very important for enterprises that continually seek information to improve their products or services while estimating the future demands for various products.

For example, a customer's review of a specific digital camera may contain the Sentences (1) and (2). In Sentence (1), an evaluation with positive (+) polarity (3) is given, and the two evaluations with opposite polarities (4) and (5) are found in the single sentence (2). The evaluations should be output in more structured forms as

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

at the right side of (3)–(5) for the deeper analysis that can help users aggregate the frequently appearing features or target products.

(1) I think the contrast is sharp.
(2) While the new model doesn't have a zoom lens, the price is low.

(3) [+] the contrast is sharp	sharp (contrast)
(4) [−] the new model doesn't have a zoom lens	have^{+neg} (zoom lens)
(5) [+] the price is low	low (price)

Therefore, the task we address here is the *clause-level* detection of expressions that indicate positive and negative evaluations, in other words, a fine-grained type of opinion mining or sentiment analysis. The clause-level approach is desirable not only for the high-precision detection of evaluation expressions relying on predicate–argument structures, but also for the construction of contextual structures that enable the automatic acquisition of a new lexicon of evaluation expressions. An unsupervised lexicon induction method is the main contribution of this article.

The clause-level detection of semantic expressions requires multiple types of syntactic and semantic knowledge. First, we need lexical entries such as 'to be sharp' and 'to have a zoom lens' that describe favourable features in the digital camera domain. Formally, we describe a lexicon for the evaluation expressions with the set of minimal syntactic structures specifying the polarity in an expression, as in examples (6)–(8). In the rest of this article, we call them *polar atoms*.

(6)	[+]	'be sharp'
(7)	[+]	'have ← zoom lens-NOM'
(8)	[+]	'be low ← price-NOM'

Also we need systematic ways to appropriately handle such syntactic phenomena as the negation in (4). A previous work (Kanayama, Nasukawa and Watanabe 2004) showed that this approach has good precision in the detection of evaluation expressions when using a domain-independent lexicon and syntactic pattern.

Now the challenge is to increase the recall, namely, to cover new domains and extract informative knowledge. In light of this, building of a domain-dependent lexicon is very important, but it is difficult because the possible lexical entries are too numerous and they may differ in each domain. For example, the adjective 'predictable' has a positive polarity in the software domain, while it has a negative polarity in the domain of movie reviews. Even within a single domain, the required knowledge to assess polarity tends to be very complex: while 'the face was in focus' is a good result when taking a picture, 'the fence (behind my friend) was in focus' expresses a bad experience.

To build a domain-dependent lexicon without tedious manual work, we have devised an unsupervised method where a user only needs to collect unannotated corpora in the target domain. Making use of a domain-independent lexicon and syntactic patterns, which are already available, our method exploits *context coherency* in terms of polarity, an assumption that *polar clauses* with the same polarity will appear successively unless the context is changed with an adversative expression.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

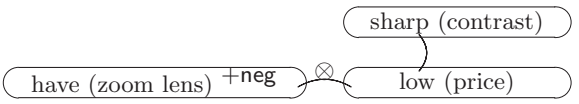


Fig. 1. An example of the context coherency observed in the two Sentences (1) and (2). The main clauses of the sentences are connected vertically, and the subordinate relation is represented by the horizontal connection. ⊗ indicates that the polarity was changed by an adversative conjunction.

For example, suppose the two Sentences (1) and (2) appear in this order. These two sentences contain three clauses and they are semantically connected as in Figure 1.

Exploiting this tendency for context coherency, we can collect candidate expressions with their tentative polarities as those adjacent to the polar clauses that have been identified by the initial lexicon that consists of domain-independent polar atoms as seeds for learning. Though there are intrinsic noises in such candidates, appropriate polar atoms are selected through statistical tests, and high precision is achieved. In addition, our learning method does not require any manual tuning of the threshold values.

In Section 2, we review related work, and Section 3 describes our runtime detection systems for clause-level evaluations in Japanese documents and the structure of the lexical entries for evaluation, the polar atoms. Section 4 explains our primary assumption for unsupervised learning, *context coherency*. To assess the strength of our assumption, we introduce the key metrics of *coherent precision* and *coherent density*. Section 5 describes our unsupervised learning method. Experimental results are shown in Section 6, and we conclude in Section 7.

2 Related work

The detection and classification of evaluations and opinions have been extensively studied in recent years as reviewed in Pang and Lee (2008), and these technologies are actually desired in the business world (Ahmad *et al.* 2004; Wright 2009). This section reviews previous work related to two of the main features of this article: clause-level methods for extracting evaluation expressions and unsupervised learning methods for lexical acquisition.

The technologies to analyse people's opinions are roughly classified into document-level classification and mention-level extraction. Typical document-level approaches classify a document into positive and negative (Pang, Lee and Vaithyanathan 2002; Turney 2002; Dave, Lawrence and Pennock 2003; Gamon 2004), or finer ratings expressed by a number of 'stars' (Pang and Lee 2005; Goldberg and Zhu 2006) where a document is assumed to have a single sentiment or a level of preference. These approaches have merits in that predefined evaluations of documents can be used for the supervised learning and the performance can be evaluated relatively easily. However, as mentioned in Section 1, our goal is to grasp individual favourable or unfavourable features of the products or services in question. Even sentence-level classification (Hu and Liu 2004; Meena and Prabhakar 2007) is not sufficient because a single sentence may contain multiple positive or negative opinions, as pointed out in Nasukawa and Yi (2003) and Wilson *et al.* (2006). Therefore, in this

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

paper, we focus on the extraction of each utterance at a finer level as phrases or clauses (Popescu and Etzioni 2005; Wilson *et al.* 2006). The clause-level sentiment detection relies on syntactic and semantic operations analogous to those in machine translation (Kanayama *et al.* 2004) for high precision. This method is the basis of this work described in Section 3. The expressions to be detected vary depending on the task. For example, Yu and Hatzivassiloglou (2003) separated facts from opinions and assigned polarities only to opinions. This distinction is important for the analysis of news articles and leads to the task of subjectivity analysis (Riloff and Wiebe 2003; Pang and Lee 2004; Wiebe *et al.* 2004). In contrast, in documents such as product reviews, even rather objective opinions tend to have polarities. They are called evaluative factuals (Nigam and Hurst 2004) and provide tangible information on the properties.

Unsupervised learning for sentiment analysis is being studied. For example, Hatzivassiloglou and McKeown (1997) labelled adjectives as positive or negative, relying on *semantic orientation*. Turney (2002) used collocations with 'excellent' or 'poor' within ten words from the given polar words, to obtain positive and negative clues for document classification.

Many other learning methods for sentiment analysis have been proposed, and they will be compared with the proposed method in Section 6.5 after a detailed description of our method.

3 Methodology for clause-level evaluation detection

Our approach towards clause-level analysis is aimed at fine-grain, high-precision sentiment analysis. This concept was originally discussed as a transfer-based extraction method analogous to machine translation (Kanayama *et al.* 2004). The MT method introduced there detected evaluation expressions with high precision using syntactic and semantic operations, and also constructed more informative semantic structures for opinion mining than the surface expressions which were extracted by the previous methods.

This section describes the first of the two advantages of the MT method: the use of deeper syntactic knowledge for the precise detection of evaluation expressions. Figure 2 illustrates the flow of our evaluation detection system, which involves two major steps. The first step is *proposition detection*. A proposition is a clause that can form a polar clause, and is identified by applying patterns to the syntactic tree corresponding to the input sentence. The next step is *polarity assignment*: the polarity of each proposition is examined by considering polar atoms, the lexicon for this task.

The remainder of this section describes these two processes, the definitions of the polar atoms, and the accuracy of the clause-level evaluation detection.

3.1 Proposition detection

'Clause-level' means that we focus on the predicative verbs and adjectives, rather than handling all of the words or phrases that may convey the sentiment of the writer as word-level or phrase-level approaches have done. In the Japanese implementation, the predicative verbs and adjectives such as in (9) are detected, and adnominal (attributive) usages of verbs and adjectives as in (10) are ignored, because '美しい'

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

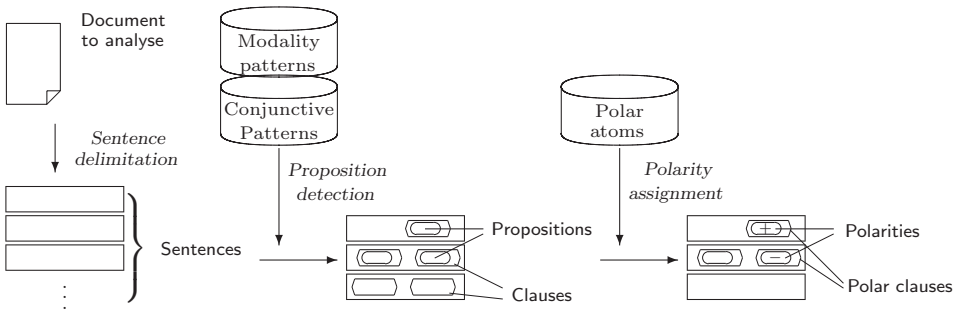


Fig. 2. The flow of the clause-level evaluation detection.

('beautiful') in (10) does not convey a positive polarity.

(9) この絵が美しい。
'The picture is beautiful'.

(10) 美しい人に会いたい。
'I want to meet a beautiful person'.

Here, we use the notion of a *proposition* as a clause without modality, led by a predicative verb or a predicative adjective. The propositions clarify the syntactic units which are subject to the assignment of polarities.

Basically, we detect a proposition only at the head of a syntactic tree[1]. However, this limitation reduces the recall of evaluation detection to a very low level. In the example (9) above, the adjective '美しい' ('beautiful') is the head of the tree, while those initial clauses in (10) and (11)–(15) below are not. To achieve higher recall while maintaining high precision, we apply two types of syntactic patterns, *modality patterns* and *conjunctive patterns*[2], to the tree structures from the full-parsing.

(11) この写真は美しいと思う。
'I think this picture is beautiful'.

(12) この写真は美しくない。
'This picture is not beautiful'.

(13) 写真が美しいとよい。
'It will be good if the picture is beautiful'.

(14) 美しいものは高い。
'Something beautiful is expensive'.

(15) 写真が美しいし、値段が安い。
'The picture is beautiful, and the price is low'.

Modality patterns, exemplified in Table 1, match some auxiliary verbs or corresponding sentence-final expressions, to allow for specific kinds of modality

[1] This is the rightmost part of the sentence since all Japanese modification is from left to right.

[2] These two types of patterns correspond to *auxiliary patterns* in the MT method, and can be applied independent of the domains.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 1. *Japanese modality patterns used to identify propositions. '+neg' denotes the proposition has the opposite polarity*

[v] -と-思う ('I think [v]')	→	[v] = proposition
[v] -と-感じる ('I feel [v]')	→	[v] = proposition
[v] -と-言わ-ざる-を-得-ない ('I have to say [v]')	→	[v] = proposition
[v] -ない ('not [v]')	→	[v] = proposition +neg
[v] -こと-は-ない ('never [v]')	→	[v] = proposition +neg

Table 2. *Japanese conjunctions used for conjunctive patterns*

coordinative (roughly 'and')
"-て", "-し", "-うえに", "-だけでなく", "-のみならず"
causal (roughly 'because')
"-ため", "-から", "-ので"
adversative (roughly 'but')
"-が", "-けど", "-けれど", "-ものの"

and negation. One of the typical patterns is [[v] -と-思う] ('I think [v]')[3], which allows the clause led by '美しい' ('beautiful') in (11) to be a proposition. Note that the expression 'I think [v]' can be regarded as the same meaning as just saying '[v]'. This treatment is specific to opinionated types of documents, but it can be applied to any domain, so we can use these patterns as domain-independent knowledge.

Negation is also handled with a modality pattern, such as [[v] -ない] ('not [v]'). In this case, a neg feature is attached to the proposition to identify '美しい' in (12) as a negated proposition. In contrast, no proposition is identified in (13) due to the deliberate absence of a pattern [[v] -と-よい] ('It is good if [v]') and more complex expressions such as in (14). We used a total of 103 domain-independent modality patterns, most of which are derived with the MT method and selected from a set of frequently used sentence-end expressions which come after verbs and adjectives.

Another type of pattern is a conjunctive pattern, which allows multiple clauses in a sentence. We used a total of tewnty-two conjunctive patterns, also derived with the MT method, as exemplified in Table 2. In such coordinative clauses and causal clauses, both clauses are regarded as propositions. For example, two propositions are detected in (15) because the coordinative particle '[v]-し-[v]' is found in the conjunctive patterns. On the other hand, no proposition is identified in a conditional clause such as in (13) due to the absence of corresponding conjunctive patterns. In this manner, our clause-level approach does not explicitly split a sentence into clauses, but processes clauses which can potentially be polar clauses by using the conjunctive patterns.

[3] [v] denotes a verb or an adjective.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Both types of patterns are applied recursively and therefore two or more patterns (e.g. [v -と-思う] and [v -ない]) can match the combinations of the expressions (e.g. 'v-と-思わ-ない' ('I don't think v')) and the appropriate propositions are identified.

The entire syntactic trees with full parsing are needed for these operations because the relations between the clauses, especially the roles of subordinate clauses, have to be captured correctly. Also, the full parser increases precision for the detection of evaluations expressed by verb/adjective clauses that may contain embedded clauses, as shown by Nasukawa and Yi (2003). In the Japanese-to-English translation engine, Japanese sentences are parsed by a full parser that exploits hand-crafted grammar and statistics (Kanayama *et al.* 2000). Its current accuracy is about 90 per cent for the dependencies, which is not perfect but is the state-of-the-art level, and this apparently gives richer structures for complex sentences than shallow parsing. Parsing errors seldom affect the precision of the evaluation detection because errors may lead to failure of matching of the syntactic subtree with the patterns, but this merely lowers the recall.

3.2 Lexicon: polar atoms

We exploit *polar atoms* as the lexicon for the assignment of polarities to propositions. A polar atom consists of a polarity, the key declinable word (a verb or adjective) and optionally, its arguments. Example (16) is a *simple polar atom*, where no argument is specified. This atom matches any proposition whose head is '美しい' ('beautiful'). Example (17) is a *complex polar atom*, which assigns a negative polarity to any proposition whose head is the verb '欠く' ('lack') and where the accusative case is '魅力' ('attraction'). This matching method is based on the MT method, in which the roles of each argument are appropriately estimated using case frames.

(16) [+] 美しい
'to be beautiful'

(17) [−] 欠く ← 魅力-を
'to lack ← attraction-ACC'

Compared to surface *n*-grams which tend to be the output of statistical learning as clues of sentiment analysis, the polar atoms introduced here are much more understandable for humans because polar atoms form the predicate–argument structures from which we can easily grasp their meanings by imagining the finite predicative form, and thus the lexicon itself can be a knowledge base for the favourable and unfavourable features. We prepared a total of 3,275 domain-independent polar atoms such as (16) and (17), most of which are originally derived from an English sentiment lexicon (Yi *et al.* 2003). The polar atoms will be outputs of learning for domain adaptation in this study.

3.3 Polarity assignment using polar atoms

To assign a polarity to each proposition, polar atoms in the lexicon are compared to the proposition. A polarity is assigned if there is a polar atom for which the verb/adjective and the arguments coincide with the proposition, and otherwise no

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 3. *Precision and recall for the detection of polar clauses from 200 sentences*

	MT method		Lexicon only	
Weak precision	100%	(31/31)	80%	(41/51)
Strong precision	89%	(31/35)	44%	(41/93)
Recall	43%	(31/72)	57%	(41/72)

polarity is assigned. When the proposition is compared with a complex polar atom, the children nodes that modify the verb/adjectives with Japanese case markers (postpositional particles) are examined, and the polarity is assigned when the noun and the case marker coincide with the polar atom. The opposite polarity of the polar atom is assigned to a proposition that has the neg feature.

Here, the matching process of Japanese case markers (postpositional particles) handles several syntactic phenomena. For example, the proposition in the example (18) matches the polar atom (19) even though the postpositional particles are different (‘は’ versus ‘が’). With the knowledge that a topic marker ‘は’ can behave like the nominative case marker ‘が’, this matching can be done beyond the surface form. This process may be much more complicated when a verb is transformed into special forms, e.g. passive or causative. The details of this process were described using an analogy between sentiment extraction and machine translation (Kanayama *et al.* 2004).

(18) このカメラにズームは付いている。
‘This camera has a zoom lens’.

(19) [+] 付く ← ズーム-が
‘have ← zoom lens-NOM’

3.4 Effects of the clause-level approach

In order to see the reliability of the clause-level evaluation detection using the syntactic patterns and domain-independent polar atoms, the previous study (Kanayama *et al.* 2004) compared the evaluation detection method using syntactic patterns and a more superficial approach relying only on a verb/adjective lexicon. Table 3 shows the experimental results in Section 5 of that paper to review the performance of polarity assignment in the digital camera domain. The results were compared by using these metrics, where $\#ab$ denotes the system's output is a and the manual output is b, and p, n, x stand for positive, negative and nonpolar clauses, respectively:

Weak precision: $\frac{\#pp+\#nn}{\#pp+\#pn+\#np+\#nn}$. The coincidence rate of the polarity between the system output and the manual output when both the system and the human evaluators assigned either a positive or negative clause.

Strong precision: $\frac{\#pp+\#nn}{\#pp+\#pn+\#px+\#np+\#nn+\#nx}$. The coincidence rate of the polarity between the system output and the manual output when the system assigned either a positive or negative clause.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Recall: $\frac{\#pp+\#nn}{\#pp+\#np+\#xp+\#pn+\#nn+\#xn}$. The detection rate of polar clauses within the manual output.

According to the observations, high precision was achieved because the modality patterns appropriately handled the syntactic phenomena such as negation and the polar words in the embedded clauses and the conditional clauses were appropriately excluded from the polar clauses. Also, the complex polar atoms contributed to limit the detection of polar clauses taking the semantics of the child node into consideration.

However, the recall was lowered by a prudent detection method based on syntactic trees and the lack of some domain-dependent expressions. Therefore, we want a method to obtain the domain-dependent knowledge to raise the recall, in a systematic way.

4 Context coherency

This section introduces the intra- and inter-sentential contexts in which we assume *context coherency* in terms of polarity of evaluation, and describes some observations based on this assumption.

4.1 Intra- and inter-sentential context

As some previous work pointed out, sentiment words with the same polarity (positive or negative) tend to collocate (Hatzivassiloglou and McKeown 1997; Turney 2002), and they can be clues to acquire new knowledge for polarities. Here is an example discourse (20) from the review of a digital camera, which contains four positive evaluations and two negative ones.

(20) このカメラは素晴らしいと思う。
'I think this camera is splendid'.
軽いし、ズームも付いている。
'It's light and has a zoom lens'.
液晶が小さいけど、満足です。
'Though the LCD is small, I'm satisfied'.
ただ、値段がちょっと高い。
'But, the price is a little high'.

The first clause of the third sentence and the fourth sentence convey the negative evaluations, and the others convey the positive evaluations. The identification of the propositions described in Section 3.1 clarifies this distribution of the polar clauses. Figure 3 illustrates the context coherency in (20) from our viewpoint with two types of contexts: *intra-* and *inter-sentential contexts*.

The intra-sentential context is the link between propositions in a sentence, which are detected as coordinative or causal clauses. If there is an adversative coordinative conjunction such as '[v1]-けど-[v2]' ([v2]-'though'-[v1]) in the third sentence in (20), a flag is attached to the relation, as denoted by '⊗' in Figure 3. Though there

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 4. *Inter-sentential adversative expressions*

しかし (‘however’), でも (‘but’), ただ (‘but’), それなのに (‘even though’), 但し (‘on condition that’), だけど (‘but’), 逆に (‘on the contrary’), とはいえ (‘although’), けれども (‘however’), 一方 (‘in contrast’)

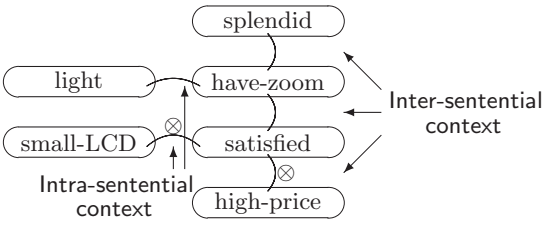


Fig. 3. The concept of the intra- and inter-sentential contexts, where the polarities are perfectly coherent. The symbol ‘⊗’ denotes the existence of an adversative conjunction.

are differences in syntactic phenomena, this is similar to the semantic orientation proposed by Hatzivassiloglou and McKeown (1997).

The inter-sentential context is the link between propositions in the main clauses of pairs of adjacent sentences in a discourse. The polarities are assumed to be the same in the inter-sentential context, unless there is an adversative expression such as those listed in Table 4. If no proposition is detected as in a nominal sentence, the context is split. This means that there is no link between the proposition of the previous sentence and that of the next sentence.

Note that the proposition ‘splendid’ in the first sentence and the proposition ‘light’ are close in the surface strings, but they are not connected in the context based on propositions. Our clause-level operation enables this kind of appropriate handling of the coherency of polarities.

4.2 Observations of context coherency

We believe these two types of context can be used for unsupervised learning as clues to assign a tentative polarity to unknown expressions, and this holds true in various domains. To test our assumption, we studied several corpora.

4.2.1 Corpora

In this section and the next one, we use Japanese corpora from discussion boards in four different domains, whose relevant features are summarised in Table 5. All of the corpora have clues to the boundaries of the posts in the form of HTML tags, so they were suitable to identify the discourses. We did not use any other structural information such as the date, author or human ratings that might have depended on the source documents.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 5. *The corpora from the four domains used in this article. The '#Posts' and '#Sentences' columns denote the numbers of posts and sentences, respectively. 'Length' is the average length of the sentences in Japanese characters*

Domain	#Posts	#Sentences	Length
Digital cameras	263,934	1,757,917	28.3
Movies	163,993	637,054	31.5
Mobile phones	155,130	609,072	25.3
Cars	159,135	959,831	30.9

4.2.2 Coherent precision

How strong is the coherency in the context described in Section 4.1? Using the polar clauses detected by the evaluation detection system with the initial lexicon, we observed the *coherent precision* of domain d with lexicon L, defined as

$$cp(d, L) = \frac{\#(\text{Coherent})}{\#(\text{Coherent}) + \#(\text{Conflict})}, \qquad (21)$$

where #(Coherent) is the occurrence count of the adjacent pairs of polar clauses that have the same polarities, as found in the discourse. Similarly, #(Conflict) is the occurrence count of the adjacent pairs of polar clauses that have opposite polarities.

There are always some errors in the automatic detection of evaluation expressions, but these errors do not affect the value of (21) because the weak precision is quite high, as described in Section 3.4. The strong precision was 89 per cent, so only 79 per cent of the co-occurrences of two polar clauses are estimated to be reliable[4], but this kind of error merely reduces both #(Coherent) and #(Conflict) assuming the errors are independent, so we can use the *cp* value in (21) for the coherent precision.

As strategies to identify the two polar clauses in these relationships, we compared two methods:

Window. A polar clause and the nearest polar clause found in the preceding n sentences in the discourse.

Link. Two polar clauses in the intra- and/or inter-sentential links as described in Section 4.1. This is the viewpoint of the context in our method.

Table 6 shows the frequencies of coherent pairs, conflicting pairs and the coherent precision for half of the digital camera domain corpus. 'Baseline' is the percentage of positive clauses among the polar clauses.

For the 'Window' method, we tested for $n = 0$, 1, 2 and ∞. '0' means two propositions within a sentence. Apparently, the larger the window size, the smaller the *cp* value. The window size '∞' implies that the nearest polar clause is picked up from anywhere within a discourse, where the average length is seven sentences. In this case, the ratio is larger than the baseline by only 2.7 per cent, and thus the

[4] The estimated precision of two adjacent polar clauses is 0.89×0.89, ignoring false negatives.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 6. *Coherent precision with various viewpoints of the contexts*

	Model	Coherent	Conflict	$cp(d, L)$
	Baseline			57.0%
Window	$n = 0$	3,428	1,916	64.1%
	$n = 1$	11,448	6,865	62.5%
	$n = 2$	16,231	10,126	61.6%
	$n = \infty$	26,365	17,831	59.7%
Link	Intra-sentence	2,583	996	72.2%
	Inter-sentence	3,987	1,533	72.2%
	Both	6,570	2,529	72.2%

Table 7. *Coherent precision for each domain*

Domain	$cp(d, L)$
Digital cameras	72.2%
Movies	76.7%
Mobile phones	72.9%
Cars	73.4%

coherency between distant clauses is not reliable even though the number of clues is relatively large.

'Link' shows the coherency of the two types of context that we considered. The *cp* values are much higher than those in the 'Window' methods, because the relationships between adjacent pairs of clauses are handled more appropriately by considering syntactic trees, adversative conjunctions, etc. The *cp* values for inter- and intra-sentential contexts are almost the same, and thus both contexts can be used to obtain 2.5 times more clues for the intra-sentential context. This result has an interesting implication because the inter-sentential context is observed in any of the successive sentences that has a proposition in the main clause, while the intra-sentential context relies on the overt occurrences of coordinative or causal conjunctions. In the rest of this article, we will use both contexts to exploit more clues for learning.

We also observed the coherent precision for each domain corpus. The results in Table 7 indicate that the number varies slightly among the corpora. Although context coherency is an intuitively reasonable assumption, there are still many discordant examples in all of our domains, even if we adopt our viewpoint of context coherency.

We selected one hundred cases of polarity conflicts in the digital camera domain. From our analysis, the causes of the conflicts can be classified into five groups (with the number in parenthesis indicating the percentage):

Errors in sentence delimitation (11%): Typically, the two sentences were not divided correctly due to the omission of a period between the sentences.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 8. *Coherent density for each domain*

Domain	$cd(d, L)$
Digital cameras	7.23%
Movies	18.71%
Mobile phones	7.31%
Cars	7.36%

Undefined adversative expressions (19%): Besides the expressions in Table 4, there were some adversative expressions like 'それでも' ('even so'), with which an opposite sentiment occurred.

Changing topics (18%): Besides the adversative expressions, a topic change may link opposite polarities. For example, 'A was really bad. B has much better quality'.

Postings without discourse (28%): Blunt writing (e.g. 'I like A. I hate B'.), and non-narrative writing such as itemizing.

Others (24%): Other errors including parsing or evaluation detection, but most of these errors are due to typos or extremely informal writing and are too difficult to parse.

The last three types (70 per cent) of the conflicting cases are intrinsic noise and cannot be eliminated easily, therefore we need to exploit the tendency of context coherency in a way that is robust against errors.

4.2.3 Coherent density

Besides the conflicting cases, there are many other cases where a polar clause does not appear in the polar context, because even more sentences do not contain evaluation expressions. We also observed the *coherent density* of the domain d with the lexicon L, defined as

$$cd(d, L) = \frac{\#(\text{Coherent})}{\#(\text{Polar})}. \qquad (22)$$

This shows the ratio of polar clauses that appear in the coherent context, among all of the polar clauses detected by the system. Therefore, when the detected polar clause is isolated and has no adjacent polar clauses with the same polarity, this value decreases.

Table 8 shows the coherent density in each domain. The movie domain has notably higher coherent density than the others. This indicates the sentiment expressions are more frequently used in the movie domain. In non-opinionated documents such as newspaper articles, this value should be much lower.

Note that the coherent density does not reflect the true distribution of the following evaluation expressions because the recall of the automatic evaluation detection is not high enough. This value only indicates the observed ratio of coherence based on the current system with the initial lexicon, but this value can be used for the

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

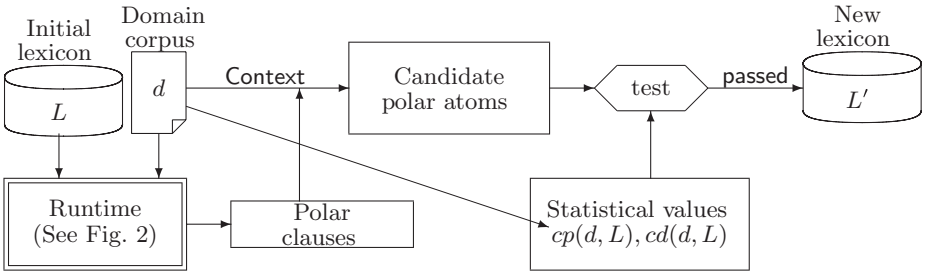


Fig. 4. The flow of the unsupervised learning process.

unsupervised learning method described in the next section. We assume the coherent density equals the probability that a polar clause detected with a newly defined polar atom appears in the polar context, because this value remains approximately unchanged even if a polar atom is added to the lexicon, as long as the lexicon is large enough.

5 Unsupervised learning for acquisition of polar atoms

In Section 4, the polarities tend to be consistent both in the intra- and inter-sentential contexts, but the coherency is far from perfect. This section proposes an unsupervised learning method, which consists of a linguistic approach to acquire candidates as new polar atoms, and a statistical method to pick up appropriate polar atoms from among the candidates.

Figure 4 shows the flow of the method. First, the runtime evaluation detection system identifies the polar clauses, then the candidate polar atoms are collected from the surroundings of the polar clauses. Next, each candidate atom is validated using its frequency count and the two metrics in the previous section, *cp* and *cd*, which are calculated from all of the polar clauses found in the domain corpus.

5.1 Counts of candidate polar atoms

From each proposition to which a polarity is not assigned, a *candidate polar atom* in the form of a simple atom (just a head verb or adjective in the proposition) is extracted. When the proposition has arguments consisting of the pair of a noun and a postpositional, a complex atom (a verb or adjective and its rightmost argument) is also extracted.

For example, from the discourse (23), a simple atom (24) and a complex atom (25) are extracted as candidates when the second sentence in (23) is not detected as a polar clause due to the absence of the corresponding polar atom. Note that no polarities are yet assigned to these atoms at this stage.

(23) このカメラは素晴らしい。ボディーが小さい。
'This camera is splendid. Its body is small'.

(24) 小さい ('to be small')

(25) 小さい ← ボディ-が ('to be small ← body-ACC')

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 9. *Examples of candidate polar atoms and their frequencies. '*' denotes that it should not be added to the lexicon*

ID	Candidate polar atom	$f(a)$	$p(a)$	$n(a)$
1*	小さい 'to be small'	3,014	226	227
2	しっかりする 'to be firm'	246	54	10
3	小さい ← ボディ-が 'to be small ← body-NOM'	11	4	0
4*	届く ← 木曜-に 'to be delivered←on Thursday'	2	0	2

For each candidate polar atom a, the total appearances in the corpus $f(a)$ is counted. Also for each a, its occurrences in positive contexts $p(a)$ and negative contexts $n(a)$ are counted, based on the context of the adjacent clauses, using the method described in Section 4.1. Again, in the example (23), assuming '素晴らしい' ('splendid') has a positive polarity in the lexicon, a positive polar clause is detected in the first sentence and thus the proposition in the second sentence is regarded as the positive context. Therefore, $f(a)$ and $p(a)$ are incremented for the two atoms (24)–(25) due to the appearance of the discourse (23). Naturally, if the proposition has the neg feature, the count of the opposite polarity is incremented.

Table 9 shows examples of candidate polar atoms with their frequencies.

5.2 Determination whether add to the lexicon

How can we assign the polarities to the located candidate polar atoms? In other words, how can we distinguish the *true polar atoms*, which should be added to the lexicon, from the *false polar atoms*, which should be discarded? As shown in Section 4, both the coherent precision (72–77 per cent) and the coherent density (7–19 percent) are so small that we cannot rely on each single appearance of the atom in the polar context.

One possible approach is to set threshold values for the frequency in a polar context, $\max(p(a), n(a))$, and for the ratio of appearances in polar contexts among the total appearances, $\frac{\max(p(a),n(a))}{f(a)}$. However, the optimum threshold values tend to depend on the corpus and the initial lexicon.

In order to set general criteria, here we assume that a true positive polar atom a should have higher $\frac{p(a)}{f(a)}$ than its average throughout the corpus. The average of $\frac{p(a)}{f(a)}$ among the positive polar atoms in the lexicon (L_p) weighted by the frequency of a is calculated as

$$\frac{\sum_{a \in L_p} \frac{p(a)}{f(a)} \cdot f(a)}{\sum_{a \in L_p} f(a)} = \frac{\sum_{a \in L_p} p(a)}{\sum_{a \in L_p} f(a)}$$

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

and assuming the symmetry between positive and negative atoms (let L_n be the negative polar atoms in the lexicon), the average equals

$$\frac{\sum_{a\in L_p} p(a) + \sum_{a\in L_n} n(a)}{\sum_{a\in L_p} f(a) + \sum_{a\in L_n} f(a)} = \frac{\#(\text{Coherent})}{\#(\text{Polar})} = cd(d, L).$$

Thus, it coincides with the coherent density defined in Section 4.2.3.

The $p(a)$ and $f(a)$ are observed values, and thus whether the true probability exceeds $cd(d, L)$ can be estimated with 90 per cent confidence using the inequality (26)[5]. We compute the confidence interval using approximation with the F-distribution (Blyth 1986)

$$\begin{aligned} &q > cd(d, L), \\ \text{where} \quad &\sum_{k=0}^{p(a)} {}_{f(a)}\mathrm{C}_k \times q^k (1-q)^{f(a)-k} = 0.9. \end{aligned} \tag{26}$$

Similarly, a true positive polar atom a should have higher $\frac{p(a)}{p(a)+n(a)}$ than its average throughout the corpus. The average of $\frac{p(a)}{p(a)+n(a)}$ among the positive polar atom in the lexicon (L_p) weighted by the frequency of a is calculated as

$$\frac{\sum_{a\in L_p} \frac{p(a)}{p(a)+n(a)} \cdot (p(a)+n(a))}{\sum_{a\in L_p} (p(a)+n(a))} = \frac{\sum_{a\in L_p} p(a)}{\sum_{a\in L_p} (p(a)+n(a))}$$

and assuming the symmetry between positive and negative atoms, the average equals

$$\frac{\sum_{a\in L_p} p(a) + \sum_{a\in L_n} n(a)}{\sum_{a\in L_p} (p(a)+n(a)) + \sum_{a\in L_n} (p(a)+n(a))} = \frac{\#(\text{Coherent})}{\#(\text{Coherent}) + \#(\text{Conflict})} = cp(d, L).$$

Thus, it coincides with the coherent precision defined in Section 4.2.2. Again, we conduct a statistical estimation with 90 per cent confidence. When the two inequalities (26) and (27) are satisfied, a is regarded as a true positive polar atom. Obviously the criteria for the negative polar atoms are analogous

$$\begin{aligned} &r > cp(d, L) \text{ or } n(a) = 0, \\ \text{where} \quad &\sum_{k=0}^{p(a)} {}_{p(a)+n(a)}\mathrm{C}_k \cdot r^k (1-r)^{p(a)+n(a)-k} = 0.9. \end{aligned} \tag{27}$$

These criteria simultaneously solve the problems in setting the threshold values, minimum frequency and scope of the polar atoms. In the example of Table 9, the simple atom '小さい' ('to be small') (ID = 1) is discarded because it does not satisfy (27), while the complex atom '小さい ← ボディ-が' ('body is small') (ID = 3) is

[5] ${}_n\mathrm{C}_k$ notation is used here for combination (n choose k).

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 10. *Agreement of two annotators' judgments of 200 polar atoms ($\kappa = 0.83$)*

		Annotator B		
		Positive	Non-polar	Negative
Annotator A	Positive	65	11	3
	Non-polar	3	72	0
	Negative	1	4	41

adopted as a positive atom. The 'しっかりする' ('to be firm') (ID = 2) is adopted as a positive simple atom, even though ten cases out of sixty four were observed in the negative context. On the other hand, '届く ← 木曜-に ('to be delivered ← on Thursday') (ID = 4) is discarded because it does not satisfy (26), even though $\frac{n(a)}{f(a)} = 1.0$, i.e. it is always observed in negative contexts.

6 Evaluation

6.1 Evaluation using polar atoms

For the evaluation of our learning method, first we propose a method of evaluation for the lexical learning.

It is costly to make consistent and large 'gold standards' in multiple domains, especially for identification tasks such as clause-level evaluation detection (cf. document-level classification tasks). Therefore, we evaluated the learning results by asking human annotators to classify the acquired polar atoms as positive, negative and non-polar, instead of classifying the instances of polar clauses detected with the new lexicon. This can be done because the polar atoms themselves are informative enough to imply to humans whether the expressions hold positive or negative meanings in the domain.

To justify the reliability of this evaluation method, two annotators evaluated 200 randomly selected candidate polar atoms in the digital camera domain. The agreement results are shown in Table 10. The manual classification was agreed upon in 89 per cent of the cases and the Kappa value was 0.83, which is high enough to be considered consistent.

Using manual judgment of the polar atoms, we evaluated the performance with the following three metrics.

Lexicon Precision. The coincidence rate of the polarity between the acquired polar atom (positive/negative) and the human evaluators' judgments (positive/negative/non-polar). It is always false if the evaluators judged it as 'non-polar'.

Runtime Precision. The estimated precision of the detection of polar clauses with newly acquired polar atoms, in the runtime evaluation detection system. The results of the manual judgments are weighted by their frequency in the corpus, and multiplied by the estimated precision when the ideal lexicon is used,

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

assuming that the newly acquired polar atoms which are judged correct by humans are as reliable as those in the initial lexicon.

Relative Recall. The estimated ratio of the number of detected polar clauses with the expanded lexicon to the number of detected polar clauses with the initial lexicon. Relative recall will be one when no new polar atom is acquired. Since the precision was high enough, this metric can be used for approximation of the recall, which is hard to evaluate in extraction tasks such as clause-/phrase-level evaluation detection.

6.2 Effects of determination method of polar atom

In order to see the effects of our method of determination of polar atoms, we conducted a learning experiment using the corpus on digital cameras described in Section 4.2.1. First, we obtained the candidate polar atoms using the domain-independent lexicon, as described in Section 5.1. Next, the polar atoms were tested using our method from Section 5.2.

To measure the two types of precision, the annotators evaluated hundred randomly selected polar atoms that were newly acquired. The lexicon precision is calculated by simply counting the correct polar atoms and dividing the number by hundred. The runtime precision is calculated as

$$\mathrm{RunPrec} \simeq \frac{\sum_a \mathrm{freq}(a)\mathrm{correct}(a)}{\sum_a \mathrm{freq}(a)}\mathrm{Prec}(B),$$

where $\mathrm{freq}(a)$ is the frequency of a polar atom a in the corpus, $\mathrm{correct}(a)$ is 1 if a is judged as correct (or 0 otherwise) and $\mathrm{Prec}(B)$ is the precision of the baseline method. Here, $\mathrm{Prec}(B)$ is set to 89 per cent, derived from the strong precision in the experimental results shown in Section 3.4.

Relative recall is estimated by comparing the numbers of detected polar clauses with and without the acquired polar atoms. Therefore, the relative recall is estimated as

$$\mathrm{RelRec} \simeq 1 + \frac{\mathrm{Num}(A)\mathrm{Prec}(A)}{\mathrm{Num}(B)\mathrm{Prec}(B)},$$

where Prec() and Num() denote the precision and the number of polar clauses detected by the system in the 2,000 sentences, respectively, and A and B denote the use of the newly acquired lexicon and the baseline lexicon, respectively. We used the lexicon precision for $\mathrm{Prec}(A)$ and set $\mathrm{Prec}(B)$ at 89 per cent again.

As a control experiment, we also tried preset criteria, where the candidate atom a is adopted when the frequency of the polarity, $\max(p(a), n(a))$, was 3 or more, and the ratio of the polarity, $\frac{\max(p(a),n(a))}{f(a)}$, was higher than the preset threshold θ. Varying θ from 0.05 to 0.8, we evaluated the precisions and the relative recall. Table 11 shows the results, where we call the proposed method the 'Statistical test'.

In our method, a very high runtime precision of 85.9 per cent was achieved. This is extremely high precision for this task, considering the correctness of both the extraction and polarity assignment was evaluated simultaneously. The lower value in the lexicon precision, 65 per cent is not as important because most of the errors were in infrequent polar atoms. When each polar atom is weighted by the frequency,

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 11. *Evaluation results with our method (Statistical test) and preset criteria for the digital camera domain. The column '# of polar atoms' denotes the number of polar atoms acquired in each domain*

Domain	# of Polar atoms	Lexicon precision	Runtime precision	Relative recall
Statistical test	708	65%	**85.9**%	**1.25**
$\theta = 0.05$	1,691	68%	52.0%	1.09
$\theta = 0.1$	1,494	64%	62.7%	1.08
$\theta = 0.2$	981	58%	60.6%	1.05
$\theta = 0.3$	632	55%	63.0%	1.03
$\theta = 0.5$	355	62%	64.0%	1.02
$\theta = 0.8$	144	62%	68.4%	1.02

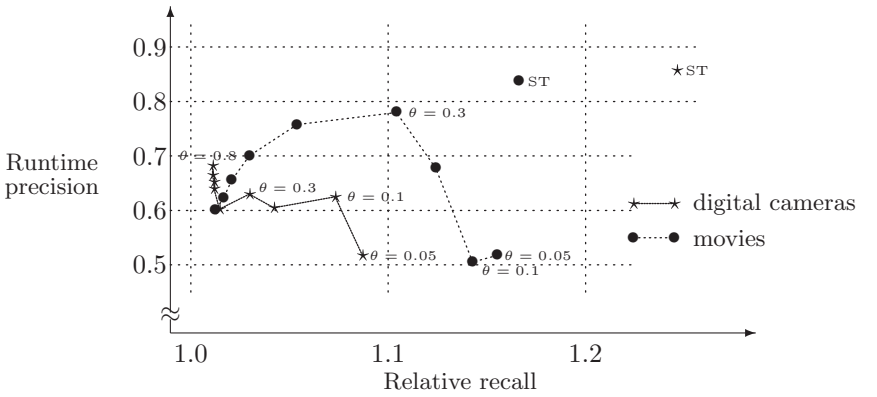


Fig. 5. Relative recall versus runtime precision with various preset threshold values θ for the digital camera and movie domains. The rightmost star and circle (labelled ST) denote the performance of the proposed method (ST; Statistical test).

the precision rises to 96.7 per cent. The relative recall of 1.25 in the digital camera domain means the recall increased from 43 per cent to 54 per cent, compared with the experiment with the domain-independent lexicon in Section 3.4.

The results showed both relative recall and runtime precision by the 'Statistical test' method significantly outperform those with preset threshold values, even for the optimum θ, 0.3.

6.3 Robustness for different conditions

6.3.1 Diversity of corpora

We also conducted the same experiments as in Section 6.2, to see the differences between the digital camera domain and the movie domain. Figure 5 shows the relationships between the runtime precision and the relative recall in the two domains.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 12. *Evaluation results with our method for four domains*

Domain	# of Polar atoms	Lexicon precision	Runtime precision	Relative recall
Digital cameras	708	65%	85.9%	1.25
Movies	462	75%	84.0%	1.16
Mobile phones	228	54%	82.0%	1.11
Cars	487	68%	81.4%	1.16

Table 13. *Evaluation results for various sizes of the initial lexicon (the digital camera domain)*

Lexicon	Coherence density	Runtime precision	Relative recall
L	7.2%	85.9%	1.25
$L_{0.8}$	6.1%	86.8%	1.36
$L_{0.5}$	3.9%	83.8%	1.98
$L_{0.2}$	3.6%	75.5%	3.27

The proposed method also achieved high precision and recall in the movie domain. Interestingly, unlike the normal precision–recall tradeoff, the runtime precision in the movie domain was reduced when the θ was high (i.e. the condition is strict). This is due to the frequent polar atoms that can be acquired at the low ratios of the polarity. In contrast, our method does not discard these important polar atoms.

Table 12 shows the precision and recall in each of the four domains, including mobile phones and cars. High runtime precisions were achieved in all of the domains, 84 per cent on average. Though the relative recall is smaller in the three domains other than the digital camera domain, the domain-dependent polar clauses are much more informative than the general ones, and thus the high-precision detection significantly enhances the system.

6.3.2 Size of the initial lexicon

We also tested the performance while varying the size of the initial lexicon L. We prepared three subsets of the initial lexicon, $L_{0.8}$, $L_{0.5}$ and $L_{0.2}$, removing polar atoms randomly. These lexicons had 0.8, 0.5 and 0.2 times the polar atoms, respectively, compared to L. Table 13 shows the precisions and recalls using these lexicons as the seeds for the learning process. Though the coherence density values vary, the precision was stable, which means that our method was robust even for different sizes of the lexicon. The smaller the initial lexicon, the higher the relative recall, because the polar atoms which were removed from L were recovered in the learning process. For example, 満足する (‘satisfied’), which is naturally included in L but which was

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Table 14. *Examples of acquired polar atoms in four domains*

Domain	
Digital cameras	[+] 明るい ('to be bright') [−] 感じる ← 重さ-を ('to feel ← heaviness-ACC') [+] 無い ← ケラレ-が ('to be free ← from vignetting')
Movies	[+] 出る ← 雰囲気-が ('to be tense ← atmosphere-NOM') [−] 半減する ← 興味-が ('to halve ← fun-ACC') [−] 長い ← オープニング-が ('to be long ← opening')
Mobile phones	[+] 大きい ← 液晶-が ('to be big ← LCD-NOM') [+] 可能だ ← バックアップ-が ('to be possible ← backup') [−] ある ← カメラ-が ('to have ← camera')
Cars	[−] 曇る ('to become fogged') [+] 増す ← 安定感-が ('to increase ← stability-NOM') [+] 無い ← 揺れ-が ('to be free ← from rattles')

removed when $L_{0.5}$ was created, was acquired automatically. This result suggests the possibility of a bootstrapping method from a small initial lexicon. However, the small seed ($L_{0.2}$) did not produce a high-precision lexicon, and the highest runtime precision (from $L_{0.8}$) was mainly due to the general polar atoms that were in the original lexicon L. Therefore, to acquire domain-dependent polar atoms with the highest possible precision, a reliable lexicon as large as possible should be used as the seed.

6.4 Qualitative evaluation

As seen in the agreement study, the polar atoms used in our study were intrinsically meaningful to humans. This is because the atoms are predicate–argument structures derived from predicative clauses, and thus humans can interpret the meaning of a polar atom by generating the corresponding sentence in its predicative form. Table 14 shows examples of the acquired polar atoms in the four domains.

In the evaluation process, some interesting results were observed. For example, a negative atom '無い ← ケラレ-が' ('to be free from vignetting') was acquired in the digital camera domain. Even the evaluator who was familiar with digital cameras did not know the term 'ケラレ' ('vignetting'), but after looking it up in the dictionary she was able to label it as negative. Our learning method discovered such technical terms and labelled them appropriately.

Also, there were discoveries in the error analysis. An evaluator assigned positive to 'ある ← カメラ-が' ('to have camera') in the mobile phone domain, but the acquired polar atom had the negative polarity. This was actually an insight from recent opinions that some users want phones without camera functions because cameras

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

tend to consume battery power or because using devices with a camera is prohibited in some buildings.

6.5 Comparison with other learning methods

As reviewed in Section 2, Hatzivassiloglou and McKeown (1997) and Turney (2002) have proposed the use of collocations of semantic orientation. In our work, we use contextual information based on the succession of sentences and multiple clauses in a complex/compound sentence, which is wider and finer than the contexts they used, and address the problem of acquiring lexical entries from the noisy clues.

Qiu *et al.* (2009) proposed a double propagation method exploiting the attributes of products. Their work has the same objective as ours and is evaluated by the newly acquired words. Our algorithm is more general than theirs because we do not require the assumption that the polarity for a specific feature is consistent in the domain corpus.

Kaji and Kitsuregawa (2007) acquired polar phrases from a large HTML corpus by exploiting very strict syntactic patterns and layout information. They successfully detected polar expressions including multiple words which correspond to the complex polar atoms in our work, but, their work can detect only domain-independent expressions, and requires an empirical threshold value to determine the polarity of the phrases. Our work seeks to acquire knowledge for each domain to be analysed with much smaller but domain-aware sets of documents.

We do not need any information on manual ratings either, as required in other methods (Higashinaka, Prasad and Walker 2006; Du and Tan 2009). Another difference in our work is that we aim at acquiring a finer and domain-dependent lexicon by exploiting a relatively large set of lexical entries that are generally used and proven to be reliable, as a seed lexicon, instead of acquiring a general lexicon that can be manually collected from a small seed lexicon. This feature is important from an engineering point of view because it addresses the practical real-world problem of covering more domains.

Learning methods for phrase-level sentiment analysis closely share one of the objectives of our approach. Popescu and Etzioni (2005) achieved high-precision extraction of opinionated phrases by using *relaxation labelling*. Their method iteratively assigns a polarity to a *word* or a combination of a *word* and a *feature*, relying on the semantic orientation of the co-occurring words with specific relationships in a sentence. The unit of expressions to which a polarity is assigned is similar to the *polar atom* in this article, but the scope of their semantic orientation is limited within each sentence. Wilson, Wiebe and Hoffmann (2009) proposed a form of supervised learning, dividing the resources into *prior polarity* and *context polarity*, which are respectively similar to polar atoms and syntactic patterns in our work. Wilson *et al.* prepared prior polarities from existing resources, and learned the context polarities by using prior polarities and annotated corpora. This means the prerequisite data and learned data are reversed in comparison to our approach. Our approach helps

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

acquire more domain-dependent knowledge, and the context polarity is easier to access in Japanese[6]. Our approach and their work can complement each other.

7 Conclusion

We devised an unsupervised method to acquire polar atoms for the clause-level evaluation detection from various domains, and demonstrated its high performance. By exploiting the tendency for context coherency and by using statistical tests, the lexicon can be expanded automatically by using unannotated corpora, and tuning of the threshold values is not required. In addition, not only a single verb or adjective, but also more complex expressions consist of a verb or adjective with an argument can be extracted automatically.

Our tuning-free method enables end-users to run the algorithm to improve the sentiment analysis. These features allow them to do on-demand analysis of more narrow domains, such as the domain of digital cameras from a specific manufacturer, or the domain of mobile phones from the female users' viewpoint.

The acquired polar atoms that consist of a verb or an adjective with an optional case are understandable by humans. This feature allows for various applications. For example, when we want to buy a product about which we do not have any background knowledge, our method allows us to identify important features for selecting products by acquiring the domain-specific polar atoms from the raw domain corpus.

References

Ahmad, K., Taskaya–Temizel, T., Cheng, D., Gillam, L., Ahmad, S., Traboulsi, H., and Nankervis, J. 2004. Financial information grid – an ESRC e-Social science pilot. In *Proceedings of the UK e-Science all Hands Meeting,* Edinburgh, UK.

Blyth, C. R. 1986. Approximate binomial confidence limits. *Journal of the American Statistical Association* **81**(395): 843–55.

Dave, K., Lawrence, S., and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary, pp. 519–28.

Du, W., and Tan, S. 2009. An iterative reinforcement approach for fine-grained opinion mining. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, CO, pp. 486–92.

Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland.

Goldberg, A. B., and Zhu, J. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: HLT/NAACL Workshop on Graph-Based Algorithms for Natural Language Processing*, New York.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational*

[6] For example, indirect negation such as caused by a subject 'nobody' or a modifier 'seldom' is rare in Japanese.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Linguistics and the 8th European Chapter of the Annual Meeting of the Association for Computational Linguistics (ACL-EACL), Madrid, Spain, pp. 174–81.

Higashinaka, R., Prasad, R., and Walker, M. A. 2006. Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp. 265–72.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, Washington, pp. 168–77.

Kaji, N., and Kitsuregawa, M. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, pp. 1075–83.

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

Kanayama, H., Torisawa, K., Mitsuishi, Y., and Tsujii, J. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany, pp. 411–17.

Meena, A., and Prabhakar, T. V. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Proceedings of the 29th European Conference on IR Research (ECIR)*, Rome, Italy, pp. 573–80.

Nasukawa, T., and Yi, J. (2003) Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, Sanibel, FL, pp. 70–7.

Nigam, K., and Hurst, M. 2004. Towards a robust metric of opinion. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: theories and Applications*, Stanford, CA.

Pang, B., and Lee, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 271–8.

Pang, B., and Lee, L. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, pp. 115–24.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval* **2**(1–2): 1–135.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, PA, pp. 79–86.

Popescu, A.-M., and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, British Columbia, Canada, pp. 339–46.

Qiu, G., Liu, B., Bu, J., and Chen, C. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Pasadena, CA, pp. 1199–204.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan, pp. 105–12.

Turney, P. D. 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*, Philadelphia, PA, pp. 417–24.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press

Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., and Martin, M. 2004. Learning subjective language. *Computational Linguistics* **30**(3): 277–308.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009) Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3): 399–433.

Wilson, T., Wiebe, J., and Hwa, R. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence* **2**(2): 73–99.

Wright, A. 2009. Our sentiments, exactly. *Communications of the ACM* **52**: 14–15.

Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. 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*, Melbourne, FL, pp. 427–34.

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan, pp. 129–36.

https://doi.org/10.1017/S1351324911000131 Published online by Cambridge University Press