Extracting Multiword Sentiment Expressions by Using a Domain-Specific Corpus and a Seed Lexicon

Kong-Joo Lee, Jee-Eun Kim, and Bo-Hyun Yun

This paper presents a novel approach to automatically generate Korean multiword sentiment expressions by using a seed sentiment lexicon and a large-scale domain-specific corpus. A multiword sentiment expression consists of a seed sentiment word and its contextual words occurring adjacent to the seed word. The multiword sentiment expressions that are the focus of our study have a different polarity from that of the seed sentiment word. The automatically extracted multiword sentiment expressions show that 1) the contextual words should be defined as a part of a multiword sentiment expression in addition to their corresponding seed sentiment word, 2) the identified multiword sentiment expressions contain various indicators for polarity shift that have rarely been recognized before, and 3) the newly recognized shifters contribute to assigning a more accurate polarity value. The empirical result shows that the proposed approach achieves improved performance of the sentiment analysis system that uses an automatically generated lexicon.

Keywords: Sentiment analysis, multiword sentiment expression, seed lexicon, domain-specific corpus, polarity shift, contextual words.

Manuscript received Jan. 25, 2013; revised Mar. 4, 2013; accepted Mar. 21, 2013.

This work was supported by the National Research Foundation of Korea (No. 2012-0004132) & Hankuk University of Foreign Studies Research Fund of 2011.

Kong-Joo Lee (phone: +82 42 821 5662, kjoolee@cnu.ac.kr) is with the Department of Information & Communication Engineering, Chungnam National University, Daejeon, Rep. of Korea.

Jee-Eun Kim (corresponding author, jeeeunk@hufs.ac.kr) is with the Department of English Linguistics, Hankuk University of Foreign Studies, Seoul, Rep. of Korea.

Bo-Hyun Yun (ybh@mokwon.ac.kr) is with the Department of Computer Education, Mokwon University, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.13.0113.0093

I. Introduction

Sentiment analysis is one of the most popular research topics to which NLP techniques can be applied. As search techniques improve, more advanced search types are available. In particular, users have shown interest in searching the Internet for various opinions on a specific target. An opinion is a private state that is not open to objective observation or verification [9] and includes beliefs, emotions, speculations, and so on. People consider opinions significant and want to hear opinions from others when they make a decision. Sentiment analysis involves determining the opinions and private states of a speaker or a writer.

A sentiment lexicon plays an essential role in identifying an opinion on an entity, a sentence, or a text. A sentiment lexicon in general contains sentiment words and their prior polarity, which is context-independent and represents positive, negative, and neutral semantic orientation of words. The prior polarity is usually decided based on lexicographers' common sense regarding a word. In lexicon-based sentiment analysis, lexicon lookup is performed after analyzing the sentences of the text to find a sentiment word. When a sentiment word is identified from a sentence, the polarity of the word is considered the semantic orientation of the sentence. Its polarity, however, must be changed when the sentiment word is influenced by a neighboring component, such as a negation word. In addition to negation, we can find other types of examples in which the prior polarity should be changed.

Example[1] (1) includes the verb "*hwuhoyha*," meaning

1) The sentences are romanized according to the Yale system, which is the standard to transcribe Korean language in linguistics.

Example (1)

(1.1) [스카프를 산 것을 [후회했다]$^{-}$]$^{-}$
"*scarf-lul* *san kes-ul* *hwuhoyha-ess-ta*"
scarf-ACC having bought-ACC regret-PAST-DECL
Regretted having bought the scarf

(1.2) [스카프를 한 개만 산 것을 [후회했다]$^{-}$]$^{+}$
"*scarf-lul* *han kay-man* *san kes-ul* *hwuhoyha-ess-ta*"
scarf-ACC One CL-only having bought-ACC regret-PAST-DECL
Regretted having bought only one scarf

Example (2)

(2.1) [묻히기에 [아깝다]$^{?}$]$^{+}$
"*mwuthiki-ey* *akkap-ta*"
to be buried-GOL wasteful/precious-DECL
Too precious to be kept in the dark/to be buried

(2.1) [이름이 [아깝다]$^{?}$]$^{-}$
"*ilum-i* *akkap-ta*"
name-NOM wasteful/precious-DECL
The name is wasted

"regret," whose prior polarity was negative. Since the word is not negated, the sentence appears to express a negative sentiment about the entity. Sentence (1.1) carries a negative sentiment, but the sentiment of (1.2) is positive because the speaker regretted not having bought more. The only difference between the two examples is whether or not there is a numeral classifier attached to the auxiliary postposition "*man*" expressing "only." However, usually, the postposition "*man*" determines the scope and does not function as a negator to reverse the polarity.

A more interesting phenomenon can be found in Example (2)[2]. The adjective "*akkap*" is polysemous; its meanings include "precious," "wasteful," and "regrettable," each of which qualifies as a sentiment word. A polysemous word such as "*akkap*" can cause a problem for sentiment analysis because the polarities of the meanings do not conform to one another; one meaning conveys positive polarity, whereas the rest express negative sentiments. Sentence (2.1) means "(something is) too precious to be buried," in which the word "*akkap*" is used as a positive sentiment word. As shown in (2.2), "*akkap*" expresses "wasteful" when used as a predicate adjective (in this case, the noun being "*ilum*," meaning "name"). The literal meaning of the sentence is "The name is wasted," and it implies that "(something is) worthy of nothing, not even having a name." The contextual words can determine the polarity of the ambiguous sentiment word as in Example (2). The examples prove that identifying a sentiment word alone leads to an incorrect conclusion.

The polarity of an expression should be determined by considering the context, as shown in the examples above. The contextual words, however, cannot be predicted, nor be identified with a set of rules. A multiword sentiment expression is non-composable from the polarity point of view. In this paper, we present a multiword sentiment expression, a sequence of words whose polarity can be correctly determined only when it is considered a single lexical entry [9] composed of a seed sentiment word and its contextual words.

The proposed technique in our work can automatically extract multiword sentiment expressions from a large-scale corpus, using a seed sentiment lexicon that lists single sentiment words only. A single sentiment word is expanded to a multiword sentiment expression including the word itself and its contextual words, which are to be used as an indicator to correctly assign sentiment polarity. For the extraction process, an unsupervised method is employed, which does not require pre-tagging on the corpus with sentiment polarity. A multiword sentiment expression is extracted from a corpus by identifying a seed sentiment word and its context. If the extracted expression is regarded as useful, it is assigned its final polarity value. The resultant lexicon is updated with the list of multiword sentiment expressions and their polarity.

This paper is organized as follows. Section II introduces related studies. Section III describes how to extract sentiment expressions from a corpus. Section IV presents the evaluation result. Finally, section V concludes the paper.

II. Related Studies

Many different techniques have been adopted in various sentiment analysis applications. A significant resource for these applications is a sentiment lexicon with broad coverage and high accuracy.

Lu and others [7] presented an experiment on constructing a sentiment lexicon that was domain and aspect dependent. The domains of the tested data included hotel reviews and customer feedback on printers. They used the following for the process: 1) prior sentiment from general purpose sentiment lexicon, 2) overall sentiment value, which was obtained from the words used in the reviews, 3) similar and opposite sentiments, and 4) linguistic heuristics using "and," "but," and "not." The experiment was performed using an unannotated corpus. Their approach successfully identified domain-specific sentiment words and determined the aspect-dependent polarity.

Choi and Cardie [3] presented a machine learning algorithm that determined the polarity of sentiment-bearing expressions. Initially, they implemented heuristic-based methods that evaluated the polarity of each constituent of an expression using a polarity lexicon, and they identified negators from the expression using a negator lexicon. Through a process called "voting," they counted the number of polarity words for both

2) CL: a numeral classifier; GOL: an abbreviation of GOAL

positive and negative and determined the overall polarity of the expression. The polarity was determined by applying a set of handwritten rules, adopting compositional semantics to the expression that had undergone a process of identifying its syntactic pattern. In doing so, the polarity of the expression was determined, considering the polarity of each constituent. This heuristic was then integrated with the learning algorithm, expecting to complement the rigidness inherent in general heuristic methods.

Taboada and others [10] presented a lexicon-based method for a deep level of analysis by which sentiment was extracted from text while contextual valence shifters were taken into account. The lexicon was built manually, and it listed sentiment-bearing words, which were classified into different parts of speech. The lexical items included multiword expressions as well as single-word expressions. The expressions were tagged with their semantic orientation, polarity, and strength. The semantic orientation was used by the Semantic Orientation CALculator (SO-CAL), which also incorporated intensification and negation. SO-CAL determined sentiment value, assuming that each word had its prior polarity and that semantic orientation could be specified using hand-ranked multiple POS dictionaries. Semantic orientation was determined by adopting a polarity shift model that was expected to better capture pragmatic intuitions regarding negation since affirmation and negation were not in symmetrical relation.

In this study, we explore an integrated method that adopts the approaches considered as new and advantageous from the previous studies. The sentiment is analyzed at a context level to identify multiword sentiment expressions. The identified expressions include not only a single sentiment word and its contextual words but also various polarity shifters that have rarely been recognized before.

III. Corpus-Based Automatic Lexicon Expansion

1. Basic Idea

The objective of our research is to automatically build an expanded sentiment lexicon by using a large corpus and a seed lexicon containing single sentiment words only. The resultant lexicon contains an extended list of entries including multiword sentiment expressions and their polarity. A multiword sentiment expression is identified, considering the context in which a single sentiment word occurs within a sentence. The polarity of a sentiment multiword expression, on the other hand, may not conform to the prior polarity value of a sentiment word, depending on the context. When a multiword sentiment expression includes an indicator to change the prior polarity of the sentiment word, the polarity of the expression must be interpreted correctly and determined carefully for an improved result. Misinterpretation of the polarity imposes a negative effect on the accuracy of the sentiment analysis, whereas multiword sentiment expressions are valuable resources for identifying the correct polarity of a target object.

2. Korean Corpus and Seed Lexicon Entries

Our corpus is collected from the material found on various Korean drama websites. When a drama gains popularity, its website's bulletin board overflows with opinions and reviews. The postings vary in topic, expressions, and length. The audiences express diverse direct opinions regarding various entities, such as plots, actors and actresses, authors, directors, and so on. We find that the postings make up one of the richest sentiment corpora. We scan the bulletin pages of 15 different websites representing popular dramas and TV variety shows. We collect 370,693 reviews from 15 different sites, which are further segmented into 2,173,210 sentences or 18,711,095 words.

The seed lexicon used for this work consists of 2,110 entries. A seed word plays a role in defining its context, as it is employed to identify a set of multiword sentiment expressions including the word itself. The entries in the seed lexicon are extracted from an existing sentiment lexicon created for a previous work [2], which listed 21,235 sentiment words with their polarity. The entries of previous work used to analyze sentiments of general domain text are manually selected. As the lexicon uses eight emotional features, including "joy," "excitement," "sadness," "fear," "anger," "shame," "disgust," and "surprise," to specify the affect of each entry, we implement a simple process to map the eight emotional features to positive or negative polarity. The mapping process is performed without complication, aside from the entries specified with "surprise" not showing any positive or negative sentiment. Accordingly, we remove the entries marked with "surprise" from the lexicon when it is confirmed as the only feature for the word. We then extract the entries consisting of a single word only, which adds up to 8,077 out of 21,235 in total. Finally, after removing the entries that never occur in our target corpus, 2,010 entries remain, as shown in (A) of Table 1. Additionally, we collect the words that often co-occur with a sentiment word to form a multiword sentiment entry but fail to become a single sentiment entry when used alone. Among these words, the 50 most frequently used words having neutral polarity are added to the lexicon, as shown in (B). The list includes "*noph*" ("high"), "*ppalu*" ("fast"), "*nukki*" ("feel"), and "*kiph*" ("deep"), and each may express a sentiment when used with its contextual words. In addition, 50 topic-related words having neutral polarity are listed, as described in (C).

Table 1. Number of entries in seed lexicon.

(A) No. of sentiment entries with single word only		No. of non-sentiment entries		Total
		(B) Co-occurring words	(C) Topic words	
Positive 1,108	Negative 902	Neutral 50	Neutral 50	2,110

These words include "*yenki*" ("acting"), "*tulama*" ("drama"), "*paywu*" ("actor" or "actress"), "*cakka*" ("writer"), and "*yenchwul*" ("directing"). In total, the seed lexicon contains 2,110 entries, and Table 1 presents the number of entries in the final version according to their polarity.

3. Extraction Procedure

The extraction procedure involves five steps: 1) preprocessing, 2) basic sentiment analysis, 3) expansion of sentiment expressions, 4) normalizing and filtering, and 5) decision of lexicon entries.

A. Preprocessing

The text posted by reviewers contains numerous errors, including misspellings, incorrect usage of spaces, and ungrammatical expressions. The posted text also contains slang, jargon, argot, taboo expressions, expletives, and misused words. Accordingly, the reviews need to be cleaned up through the process of normalizing the texts. Since the corpus is a collection of reviews on dramas and variety shows, there are plenty of proper nouns, including the names of the actors, actresses, and characters and the titles of the show and episode. To properly handle the proper nouns in the remaining process, named entity recognition has to be employed first. The person names and the titles are identified and replaced by the tag "PNAME" and "TITLE," respectively. With the tagging process completed, each sentence in the review has undergone morphological analysis to find the stem of each word. Additional analysis is conducted using a Korean dependency parser [6], [8] to identify relations between the words in a sentence.

B. Basic Sentiment Analysis

We have a sentiment analyzer [2] that adopts a sentiment lexicon containing 21,235 entries. This lexicon, however, is replaced with the seed lexicon described in Table 1 of subsection III.2 to find multiword sentiment expressions from the text at a later point in the processing.

When the sentiment analyzer detects a sentiment word that is accompanied by an explicit negator, its polarity value should be reversed. The explicit negators considered in the research include three different types [1]. The first type is a negative adverb such as "*an*" or "*mos*," which means "no," "not," "never," or "not able," and reverses the polarity of the predicate that immediately follows. The second type is a negative auxiliary[3] such as "*anh*," "*aniha*," "*mosha*," or "*mal*," which negates the main verb if positioned after the main verb. The last type is a negative adjective[4] such as "*eps*" or "*ani*," which can form a predicate by itself, negating the subject. They are all explicit negators that reverse the polarity.

The polarity values are marked with an integer. The integer value +1 is assigned to the positive sentiment, the value −1 is for the negative, and zero is for the neutral. When a sentiment word is detected in a sentence, the sentiment of the sentence is copied from the polarity of the word. If more than one sentiment word is found in a sentence, the overall polarity of the sentence is decided by the sum of the polarity values. The neutral value is assigned to a sentence with no sentiment word. As an output of the process, every sentence in a review document is assigned a polarity value.

Two adjacent sentences are likely to have the same polarity unless they are connected by an adversative connector [5]. For the further processing, we define a relationship between every pair of adjacent sentences in a review. If the pair is not connected by an adversative conjunction, the two sentences are regarded as sharing the same sentiment polarity. Korean adversative conjunctions used in this work are as follows: "*kurena*" ("however"), "*kurehciman*" ("however"), "*haciman*" ("but"), "*kuremeyto_pwulkwuhako*" ("nevertheless"), "*hana*" ("but"), and "*kurayto*" ("and yet").

C. Expansion of Expressions

We expand a sentiment entry by adding the contextual words that hold dependency relations in a sentence. In other words, a sentiment word is expanded with its head and dependent words to generate a set of multiword sentiment expressions.

For example, Fig. 1 depicts a dependency tree of a sentence including a seed sentiment word, "*hwuhoyhal*" ("*hwuhoyha*" — "regret"), identified at the sixth word in the sentence. A list of ten multiword sentiment expressions is extracted from the tree, as shown in Table 2. We use a window with a maximum of four dependency words to expand the expressions. The underlined words are the seed sentiment words from which the expressions are expanded.

One of our assumptions, as described in subsection III.1, is that the polarity of the nearest neighboring sentence is assigned

3) Their meaning is the same as that of a negative adverb.

4) "eps" means "not exist," "there is not"/"there are not," or "not have" and "ani" means "be not."

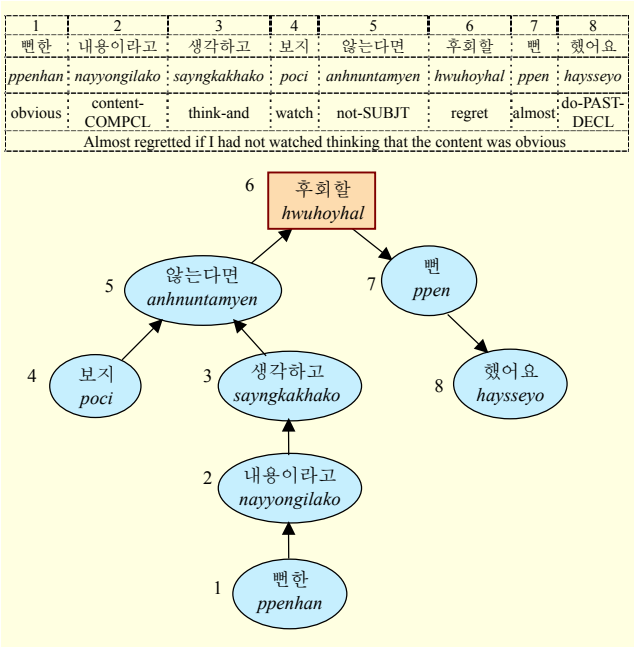

Fig. 1. Dependency tree of example sentence.

to a multiword sentiment expression when the polarity of the expression is different from that of the seed sentiment word. The example sentence in Fig. 1 contains a sentiment word with negative orientation, but its correctly interpreted polarity is positive. In this case, the polarity of the nearest neighboring sentence can be assigned to the multiword expressions listed in Table 2.

The procedure to find the nearest neighboring sentence to the target multiword expression is as follows. The notation **BUT** denotes an adversative conjunction that connects two sentences, while **AND** signifies a relation between a pair of sentences without an adversative conjunction. When a multiword expression, *ME*, is extracted from the *i*-th sentence S_i in a document,

Polarity(*ME*) =
- if S_{i-1} has non-zero polarity and **AND** (S_{i-1}, S_i)
  - then Polarity(S_{i-1})
- else if S_{i-1} has non-zero polarity and **BUT** (S_{i-1}, S_i)
  - then ¬ Polarity(S_{i-1})
- else if S_{i+1} has non-zero polarity and **AND** (S_i, S_{i+1})
  - then Polarity(S_{i+1})
- else if S_{i+1} has non-zero polarity and **BUT** (S_i, S_{i+1})
  - then ¬ Polarity(S_{i+1})
- else Polarity(S_j) such that absolute value of (i–j) is minimum and S_j has non-zero polarity.

D. Normalization and Filtering

We collect all the multiword sentiment expressions for each seed entry. The column headed "Surface form" in Table 3 presents several examples of the expressions that are expanded with the seed word "*hwuhoyha*" ("regret").

Even though expressions 3), 7), 11), 12), and 13) specify the same meaning, their surface forms are slightly different from one another due to the differences in suffixation. Korean is very

Table 2. Examples of multiword sentiment expressions.

Length 2	1. [*hwuhoyhal*$_6$ *ppen*$_7$] (almost regret)
	2. [*anhnuntamyen*$_5$ *hwuhoyhal*$_6$] (if not, will regret)
Length 3	3. [*hwuhoyhal*$_6$ *ppen*$_7$ *haysseyo*$_8$] (almost regretted)
	4. [*poci*$_4$ *anhnuntamyen*$_5$ *hwuhoyhal*$_6$] (if you do not watch/see, will regret)
	5. [*sayngkakhako*$_3$ *anhnuntamyen*$_5$ *hwuhoyhal*$_6$] (if thinking that the content is about something and not doing, will regret)
	6. [*anhnuntamyen*$_5$ *hwuhoyhal*$_6$ *ppen*$_7$] (if not, almost regret)
Length 4	7. [*anhnuntamyen*$_5$ *hwuhoyhal*$_6$ *ppen*$_7$ *haysseyo*$_8$] (if not, almost regretted)
	8. [*poci*$_4$ *anhnuntamyen*$_5$ *hwuhoyhal*$_6$ *ppen*$_7$] (if not watching, almost regret)
	9. [*sayngkakhako*$_3$ *anhnuntamyen*$_5$ *hwuhoyhal*$_6$ *ppen*$_7$] (if thinking that the content is about something and not doing, almost regret)
	10. [*nayyongilako*$_2$ *sayngkakhako*$_3$ *anhnuntamyen*$_5$ *hwuhoyhal*$_6$] (if thinking that the content is about something and not doing, will regret)

Table 3. Surface form and normalized form of multiword sentiment expressions for seed entry "*hwuhoyha*."

	Surface form	Normalized form
1)	*hwuhoyhal ppen*	*hwuhoy*
2)	*anhnuntamyen hwuhoyhal*	NEG *hwuhoy*
3)	*poci anhnuntamyen hwuhoyhal*	*pota* NEG *hwuhoy*
4)	*sayngkakhako anhnuntamyen hwuhoyhal*	*sayngkak* NEG *hwuhoy*
5)	*anhnuntamyen hwuhoyhal ppen*	NEG *hwuhoy*
6)	*anhnuntamyen hwuhoyhal ppen haysseyo*	NEG *hwuhoy*
7)	*poci anhnuntamyen hwuhoyhal ppen*	*pota* NEG *hwuhoy*
8)	*hwuhoyhal ppen haysseyo*	*hwuhoy*
9)	*sayngkakhako anhnuntamyen hwuhoyhal ppen*	*sayngkak* NEG *hwuhoy*
10)	*nayyongilako sayngkakhako anhnuntamyen hwuhoyhal*	*nayyong sayngkak* NEG *hwuhoy*
11)	*poci anhnuntamyen hwuhoyhalcito*	*pota* NEG *hwuhoy*
12)	*poci anhasstamyen hwuhoy hal*	*pota* NEG *hwuhoy*
13)	*poci anhnuntamyen hwuhoy hal*	*pota* NEG *hwuhoy*
14)	*hwuhoyhal kes*	*hwuhoy*

rich in the variety of suffixes that are, in their use, agglutinated to the end of a content word. Even when expressions 3) and 13) have exactly the same meaning, the space between "*hwuhoy*" and "*hal*" in 13) makes the surface form distinctive from 3). These two different surface forms are possible because spacing is optional for some cases according to the Korean orthography. The rest of the surface forms also mean the same while presenting a difference in nuance. The various surface forms having the same meaning can make the later processing more complicated due to data sparseness. To alleviate this problem, we define normalized forms, as listed in Table 3. A normalized form is created by replacing all the content words of the expression with their stem while discarding the rest, such as a pronoun, a bound noun, a function word, and so on. An explicit negator (described in part B of subsection III.3) is replaced with the tag "NEG" if it is detected within the expression. After the normalization, entries 3), 7), 11), 12), and 13) share an identical normalized form.

Each multiword sentiment expression is characterized by four different types of information: the surface form, the normalized form, its source_site, and the estimated_polarity. The source_site refers to the drama/show site that the expression is extracted from. The estimated_polarity is the sentiment value determined by the process described in part C of subsection III.3.

For the same seed word, we create a cluster of multiword sentiment expressions based on the same normalized form. When all the multiword expressions in a cluster share the same source_site, the cluster is discarded because the expressions are considered to be not general enough; that is, they are inclined to a particular entity or to the writing style of the reviewer who generated the text. Consequently, those expressions are determined as ineligible to be listed in the lexicon.

E. Decision of Entries

Multiword sentiment expressions were extracted and tied in clusters in the previous step. The next step is to decide which multiword sentiment expression is qualified to be listed as a new entry in the sentiment lexicon. In doing so, we evaluate which cluster is significant enough to be registered in the lexicon. The possibility that an expression becomes a candidate for a new headword is high when all multiword expressions in a cluster have the same estimated polarity. If all the expressions in a cluster do not have the same polarity, the cluster is to be discarded. However, the cluster is saved when the classification of the expressions according to their polarity can be confirmed as valid using a different set of features. We adopt a measurement of classifiability [4] from a decision tree pruning technique to decide if a cluster has been classified properly or not. The classifiability is used as a criterion to stop a decision tree

```
Decision_on_candidate_sentiment_entries:
1.  for each seed entry x
2.      get S_x and organize it as a collection of clusters C_xi
3.                          which have the same normalized form;
4.      input: S_x = {C_x1, C_x2, ..., C_xn};
5.      output: R_x = { };
6.      for each C_xi in S_x
7.          if num_of_expressions_in(C_xi) < θ1, then discard C_xi from S_x;
8.          else if impurity(C_xi) < θ2 then put C_xi into R_x;
9.          else if classifiability(C_xi) < θ3 then discard C_xi from S_x;
10.         else
11.             do a best split C_xi into C_xi^+y and C_xi^-y by using a feature y;
12.             add C_xi^+y and C_xi^-y into S_x and discard C_xi from S_x;
```

Fig. 2. Algorithm to decide candidate for sentiment entry.

from growing. A cluster can be thought of as the root node of a growing decision tree, and each expression in the cluster has its polarity value, which is used as the target value of the decision tree. The basic objective of the algorithm described in Fig. 2 is to grow a decision tree to identify which nodes are meaningful for determining the candidates for the sentiment lexicon.

Let S_x be a set of the multiword expressions expanded from the same seed sentiment entry x. S_x can be regarded as a collection of clusters that have the same normalized forms. Therefore, S_x can have a set of clusters C_{x1}, C_{x2}, ..., C_{xn}, where C_{xi} is a cluster of the expressions that not only has the same seed sentiment entry x but also has the same normalized form. The value n is the number of different normalized forms found in S_x. The algorithm determines which cluster can be a candidate for a multiword sentiment entry by constructing a small decision tree for each cluster, and ultimately to classify the expressions according to their polarity. Finally, the expressions that remain in the final output produced by the algorithm become the candidates for a new sentiment entry.

Because we cannot decide whether the expressions in a cluster are useful or not when the number of expressions in the cluster is too small, such a cluster is discarded, as described in line 7 of the algorithm. For the meaningful clusters, the entropy is measured for the function impurity in line 8. When the entropy value indicating the impurity of the cluster is less than the predefined value, the cluster is added to the output R_x. The classifiability in line 9 is used to check if the cluster C_{xi} is valuable enough to be split into smaller clusters in a decision tree while consulting an additional feature.

When all expressions in a cluster have the same normalized form, additional information is necessary to split the cluster further. For the splits, we use the surface forms of an expression since they are different from one another. Each expression is represented as a feature vector that consists of unigrams and bigrams of the surface form. Both content words and function words are used as a feature for further splits since

they equally affect sentiment analysis. We use the information of gain ratio to check the classifiability, as in line 9. If the gain ratio of the cluster is not high enough for further splits, the cluster is discarded. This is interpreted as the cluster that is neither pure nor classifiable in terms of the polarity value. When a cluster is classifiable, as in lines 10 through 12, it is split into smaller ones based on the polarity, by consulting an additional feature, y, which is the best feature to split a cluster into smaller ones with the lowest impurity. The newly formed clusters are then added to S_x.

The algorithm returns R_x, which contains the clusters of the candidates for a new sentiment entry. A cluster is assigned with the representative polarity, which the majority of the expressions in the cluster share. A candidate with the simplest form is selected as a new sentiment entry to be listed in the lexicon with the cluster's representative polarity. Since it is not easy to determine the parameter values using the algorithm in Fig. 2, we determine them through several heuristic experiments to get the best performance, shown in Table 5. The value of θ_1, θ_2, and θ_3 are set to 4, 0.75, and 0.5, respectively.

IV. Resultant Lexicon and Evaluation

The focus of this work lies on multiword sentiment expressions, which have different polarity from that of the seed entries. If a multiword sentiment expression has a reversed polarity from the prior value due to a simple negating process, such as using an explicit negator only, the expression is beyond our interests for this experiment.

1. Resultant Lexicon

This subsection presents various types of resultant multiword sentiment expressions: Examples (3) through (7), which are automatically expanded.

Example (3)

(3.1) [기다리기 [지루하다]$^-$]$^+$
"*kitali-ki* *cilwuha-ta*"
(Waiting is too boring/Too bored to wait)

(3.2) [[지루하기]$^-$ 짝이 [없다]$^\neg$]$^-$
"*cilwuha-ki* *ccak-i* *eps-ta*"
(extremely boring)

(3.3) [[지루할]$^-$ 틈이 [없다]$^\neg$]$^+$
"*cilwuha-l* *thum-i* *eps-ta*"
(No time to be bored/So exciting that one would not feel any sense of boredom)

The sentiment adjective "*cilwuha*" ("bored" or "boring"), shared by the expressions in Example (3), represents negative polarity. Expression (3.1) can be positive or negative depending on the context in which it is used. As the corpus domain is drama reviews, expression (3.1) is interpreted as "Since the TV audience is eager to watch the drama, they are too bored to wait," which is a positive opinion. One of the interesting findings in Example (3) is the difference between the expressions (3.2) and (3.3). The expressions have a similar syntactic structure to each other, and they are also negated by an explicit polarity shifter "*eps*" ("not exist"). However, their polarity values are opposite of each other. The polarity shifter negates the sentiment word "*cilwuha*" in (3.3) while it serves as an intensifier[5] in (3.2), coupled with its preceding word "*ccak-i*."

Example (4)

(4.1) [[믿기]$^+$ [어렵다/힘들다]$^-$]$^-$
"*mit-ki* *elyep-ta/himtul-ta*"
(difficult to believe/hard to believe)

(4.2) [[믿는]$^+$ 도끼에 발등 찍히다]$^-$
"*mit-nun* *tokki-ey* *paltung* *ccikhi-ta*"
(proverb: Get stabbed in the back)

(4.3) [빽만 [믿다]$^+$]$^-$
"*ppayk-man* *mit-ta*"
(Trusting someone only who is backing up)

Example (4) presents a list of expressions that include "*mit*" ("believe" or "trust"), which has positive polarity. In the case of expression (4.1), each of the contextual words, "*elyep*" ("difficult") and "*himtul*" ("hard"), functions as a polarity shifter[6] although they themselves are sentiment words. The expression (4.2), a proverb extracted from the corpus, expresses negative sentiment without an additional sentiment word or a polarity shifter. Example (4.3) is a negative expression frequently used for cynically describing a person with a lot of pull.

Example (5)

(5.1) [시간 가는 게 [아깝다]$^?$]$^+$
"*sikan kanun key akkap-ta*"
(wish to stop time from flying away)

(5.2) [[시간이 [아깝다]$^?$]$^-$
"*sikan-i* *akkap-ta*"
(Time is/was wasted for doing something)

(5.3) [[칭찬이]$^+$ [아깝지]$^?$ [않다]$^\neg$]$^+$
"*chingchan-i* *akkap-ci* *anh-ta*"
(Praise is not wasted/cannot praise too much)

(5.4) [[욕도]$^-$ [아깝다]$^?$]$^-$
"*yok-to* *akkap-ta*"
(even insult is wasted)

5) This will be discussed in detail in relation to Example (7.3).
6) We will discuss more on this when explaining Example (6.2).

The polysemous adjective "*akkap*" of Example (5) translates as follows: "wasteful," "regrettable," "pitiful," "precious" or "good," "worthy," and "be good for." The expressions that include this adjective present interesting phenomena. A noticeable one is the contrast between (5.1) and (5.2). Both examples are "time" related expressions, but their sentiments are opposite. The former means that something is so good/interesting/funny that the speaker would like to stop time from "flying," which expresses high praise. On the other hand, the latter means "to regret having wasted time doing something." The expressions in (5.3) and (5.4) are commonly encountered in everyday use. The expressions have the same syntactic structure except that the adjective "*akkap*" in (5.3) is negated. The adjective "*akkap*" conveys the negative meaning "*wasted*" in both expressions. In (5.3), the negated adjective predicates the noun "*chingchan*" ("praise"), expressing positive polarity, the overall sentiment specifying positive value. The expression (5.4) conveys a negative sentiment while predicating the negative noun "*yok*" ("an insult"). The adjective in (5.4), however, cannot be negated without resulting in a semantic anomaly, whereas (5.3) stays grammatical even when the negation is removed. The restriction is applied to the contextual words of "*akkap*." Occurring with the adjective "*akkap*," a class of nouns allows negation while the others do not.

One of the noteworthy findings from the results is that some expressions intensify the sentiment rather than reverse the polarity value even though they include an explicit negator. In addition, we find several polarity shifters that we did not recognized before. The algorithm in Fig. 2 returns a set of nodes representing the candidates for sentiment entries by constructing a small-scale decision tree. In the process of splitting the nodes in a decision tree, a phenomenon catches our attention. When a specific feature is frequently used in splitting nodes across all the decision trees of the entire collection of seed words, it is considered a general polarity indicator for splitting clusters. We collect the features and the number of times they are used in splitting clusters during the overall process of the algorithm. Some of the results are shown in Examples (6) and (7).

Example (6)

(6.1) (는)커녕 (*nun*)*kenyeng* (far from/instead/anything but at all)
ex) [[재미]+[는커녕]ㄱ]−
"*caymi nunkenyeng*" (far from being fun)

(6.2) 기 어렵/힘들 *ki elyep*/*himtul* (difficult/hard)
ex) [[웃음]+ 주[기 어렵다]ㄱ]−
"*wusum cwuki elyep-ta*" (hard to make it fun)

(6.3) 것과 멀 *keskwa mel* (far from)
ex) [[세련된]+ [것과 멀다]ㄱ]−
"*seylyentoyn keskwa mel-ta*" (far from being polished)

Example (6) enumerates the cases in which a word functions as a polarity shifter that not only reverses a sentiment value but also adjusts the intensity of the sentiment. The postposition "*kenyeng*" or "*nunkenyeng*" in (6.1) reverses the polarity of the sentiment word to which it is attached and triggers another sentiment word with the opposite sentiment value to follow. In addition, it emphasizes the reversed sentiment. The adjectives listed in (6.2) and (6.3) also reverse the polarity value, and, accordingly, they are each classified as a content-word negator [3]. Unlike the polarity shifter in (6.1), they diminish the sentiment intensity, which results in the tone becoming rather moderate.

Example (7)

(7.1) 을/를 수밖에 없 *ul/l swupakkey eps* (cannot help doing)
ex) [[두려움이]− 있[을 수밖에 없다]↑]−
"*twulyewumi iss ul swupakkey eps-ta*" (cannot help being scared)

(7.2) 그지 없 *kuci eps* (endless)
ex) [[한심하기]− [그지 없어요]↑]−
"*hansimhaki kuci eps-eyo*" (extremely pathetic)

(7.3) 짝이 없 *ccaki eps* (no comparison)
ex) [[불쌍하기]− [짝이 없다]↑]−
"*pwulssanghaki ccaki eps-ta*" (extremely pitiful)

(7.4) 뿐 아니라 *ppwun anila* (not only)
ex) [[어색할]− [뿐 아니라]↑]−
"*esaykhal ppwun anila*" (not only awkward)

(7.5) 않을 수 없 *anhul swu eps* (no choice but to)
ex) [[의심하지]− [않을 수 없다]↑]−
"*uysimhaci anhul swu eps-ta*" (no choice but to suspect)

The polarity shifters listed in Example (7) contain an explicit negator. Instead of reversing the polarity of the expression, they work as intensifiers. The expressions of (7.1), (7.2), and (7.3) contain the negator "*eps*" ("not exist"). Similarly, (7.4) has the negator "*ani*" ("be not"). The expression in (7.5), however, includes double negation formed with two negators, "*anh*" ("not") and "*eps*" ("not exist"), which is, by convention, expected to retain the polarity value of the sentiment word. Even when all the expressions include an explicit negator, the prior polarity of a seed sentiment word is kept and assigned as the sentiment of the expressions while increasing the intensity of the sentiment. Accordingly, interpreting these negators separately from the context may lead to an incorrect conclusion.

The expressions in Examples (6) and (7) are productive; in theory, they can occur with almost all of the sentiment expressions. Therefore, they should be registered in a negator/intensifier set and checked at runtime rather than listed in the sentiment lexicon, whereas the expressions of Examples (3) through (5) should be registered in the sentiment lexicon with their polarity value.

2. Evaluation

A. Lexicon Accuracy and Coverage

Our main goal is to automatically extract sentiment expressions from the domain-specific corpus. Therefore, an evaluation must be performed to check how useful the multiword sentiment expressions are as lexicon entries. A total of 3,193 entries are extracted from the process described in subsection III.3. We provide a lexicographer with these entries and request verification of whether they are qualified to be sentiment expressions. Table 4 summarizes the evaluation results. Of the total expressions, 23.9% of the total expressions are found to be non-sentiment expressions, which is an unexpected result. The result is due to the following factors. Firstly, the words are homographs with sentiment words. Secondly, the preprocessing steps include the morphological and syntactic analyses. Finally, the seed lexicon includes a set of words whose polarity is neutral. When they are used as seed words, it is very likely that they cause the generation of non-sentiment expressions.

The remaining 76.1% are confirmed as sentiment expressions by the lexicographer. We classify these entries into two groups: a class whose polarity is the same as the prior value of the original seed entries and a class that has a different polarity from the prior value. The classification result is described in Table 4. Even though the 2,103 entries in ① are sentiment expressions, we focus on the entries in ②. The total number of entries listed in ② is 325, which can be thought of as a somewhat small number. However, despite its small size, this group of expressions has a considerable impact on the accuracy of the system, a result addressed further in Table 6.

We examine the 325 sentiment entries more closely. The result is presented in Table 5, in which the second row lists the number of multiword sentiment expressions with the prior polarity of their seed words specified in the first row. The accuracy listed in the third row shows how many entries were assigned the correct polarity according to our approach. Over 53% of the 325 entries come from the neutral seed entries, and their accuracy is higher than that of the other two cases. The neutral seed entries prove to be useful for identifying the expanded sentiment expressions, and they are the main cause of the simultaneous generation of non-sentiment expressions; they are a necessary evil for this work.

We compare the automatically extracted entries with those of the manually-built sentiment lexicon to evaluate the coverage of the automatically-expanded lexicon. While 33.28% of the total entries in the manually-built lexicon are found in the automatically-expanded lexicon, 24.61% of the total entries in the automatically-expanded lexicon are listed in the manually-built lexicon.

Table 4. Evaluation results of lexicon entries.

		No. of entries
Non-sentiment expressions		765 (23.9%)
Sentiment expressions	① Multiword sentiment expressions that have same polarity as original seed entries	2,103 (65.9%)
	② Multiword sentiment expressions that have different polarity than original seed entries	325 (10.2%)
Total		3,193 (100%)

Table 5. Result of assessing 325 entries.

Polarity of seed word	Positive	Negative	Neutral	Total
No. of entries	90 (27.7%)	61 (18.8%)	174 (53.5%)	325 (100%)
Accuracy of estimated polarity (different from prior value)	67.4%	59.1%	85.3%	75.4%

Table 6. Result of sentiment analysis system.

	Precision	Recall
(A) Manually-built sentiment lexicon (21,235 entries) [2]	62.9%	58.4%
(B) Sentiment lexicon with single entries only (2,110 entries)	51.4%	44.7%
(C) Sentiment lexicon with single entries + automatically expanded expressions (2,110 + 325 entries)	61.5%	58.1%

B. Precision & Recall of Sentiment Analysis System

To evaluate the sentiment analysis system with the multiword sentiment entries, we collect more testing data on drama reviews, which does not overlap with the existing data. The testing corpus is composed of 300 reviews from three different drama websites. We manually annotate them with sentiment polarity. A total of 1,092 expressions are identified while annotating with a polarity value; 654 have positive polarity and 438 have negative polarity.

We compare three sentiment analysis systems that are different from one another in terms of their implemented lexicon. Table 6 presents the comparison result. System (A) adopts a manually-built sentiment lexicon with 21,235 entries [2], whereas system (B) uses single word entries[7] only. System (C) contains automatically expanded expressions[8] in addition to the entries that system (B) uses.

7) They refer to 2,110 entries listed in Table 1.
8) The expressions are presented in ② of Table 4.

As shown in Table 6, a meaningful result is drawn from the comparison. The accuracy of system (C) almost reaches that of (A) even though (C) has a smaller number of lexicon entries. The result proves that we can achieve the same level of performance with a seed lexicon and a large domain-specific corpus of our interest as the performance produced with a manually-built lexicon.

V. Conclusion

In this paper, we presented a novel approach to building a sentiment lexicon that contains multiword sentiment expressions as well as single sentiment entries. Since we adopted an unsupervised model to extract the expressions, we did not have to preprocess the corpus. A large corpus and a sentiment seed lexicon were the only requirements for this approach. Although we must admit that further research is required to improve the performance of the system, this approach has produced significant outcomes. Firstly, multiword sentiment expressions were automatically extracted from a large corpus by using a seed lexicon. A multiword sentiment expression was generated by combining a seed entry with its contextual words. Among the expressions, we selected ones with the different polarity from the prior value of the seed word. Correctly identifying those expressions proved to be essential for improving the performance of the sentiment analysis system. Secondly, a domain-specific sentiment lexicon was built successfully using the proposed approach. When domain-specific topic words were registered in the seed lexicon, domain-specific multiword sentiment expressions were generated and listed in the resultant lexicon. Lastly, we were able to identify various types of polarity shifters in addition to simple negating words. Some of the shifters did not reverse the polarity even though they were explicit negators. Some others, on the other hand, reversed the polarity even when it did not appear to have any clue for the reversal. This was possible only when those shifters occurred with particular contextual words. Our results suggest that the direction of the research is encouraging. Furthermore, the resultant sentiment lexicon is expected to contribute to better performance when it is refined by human experts.

References

[1] S.-J. Chang, *Korean*, Philadelphia, PA: John Benjamins Publishing Company, 1996.

[2] Y. Cho and K. Lee, "Automatic Affect Recognition Using Natural Language Processing Techniques and Manually Built Affect Lexicon," *IEICE Trans. Inf. Syst.*, vol. E89D, no. 12, 2006, pp. 2964-2971.

[3] Y. Choi and C. Cardie, "Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis," *Proc. Conf. Empirical Methods Natural Language Process.*, 2008, pp. 793-801.

[4] M. Dong and R. Kothari, "Classifiability Based Pruning of Decision Trees," *Proc. Neural Netw.*, 2001, pp. 1739-1743.

[5] H. Kanayama and T. Nasukawa, "Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis," *Proc. Conf. Empirical Methods Natural Language Process.*, 2006, pp. 355-363.

[6] S. Lee and J. Seo, "Grammatical Relations Identification of Korean Parsed Texts Using Support Vector Machines," *LNCS*, vol. 3206, 2004, pp. 121-128.

[7] Y. Lu et al., "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach," *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 347-356.

[8] I. Milevskiy and J.-Y. Ha, "A Fast Algorithm for Korean Text Extraction and Segmentation from Subway Signboard Images Utilizing Smartphone Sensors," *J. Comput. Sci. Eng.*, vol. 5, no. 3, Sept. 2011, pp. 161-166.

[9] R. Quirk et al., *A Comprehensive Grammar of the English Language*, New York: Longman, 1985.

[10] M. Taboada et al., "Lexicon-Based Methods for Sentiment Analysis," *Comput. Linguistics*, June 2011, vol. 37, no. 2, pp. 267-307.

Kong-Joo Lee received her B.S. degree in computer science from Sogang University, Rep. of Korea, in 1992. She received her M.S. degree and her Ph.D. degree in computer science from KAIST in 1994 and 1998, respectively. From 1998 through 2002, she worked for Microsoft, Rep. of Korea. In 2003, she was a faculty member of the Department of Computer Science in Ewha Womans University. In 2004, she was a faculty member of the School of Computer Information Technology in KyungIn Women's College. Since 2005, she has been a faculty member of the Department of Information Communications Engineering in ChungNam National University. Her research interests include information retrieval, natural language parsing, and machine translation.

Jee-Eun Kim received her B.S. degree in English from Hankuk University of Foreign Studies, Rep. of Korea, in 1985. She received her M.S. degree and her Ph.D. degree in linguistics from Georgetown University in 1989 and 1993, respectively. She worked for Microsoft Research and Microsoft Korea from 1995 until 2002. She worked from 2006 to 2008 as a freelance researcher on a project to develop an automated English scoring system. She is currently an associate professor in the Department of English Linguistics at Hankuk University of Foreign Studies. Her research interests include sentiment analysis and developing computational grammar of Korean and English for natural language processing applications.

Bo-Hyun Yun received his B.S. from Mokpo University in 1992 and his M.S. and Ph.D. degrees in computer science from Korea University, Seoul, Rep. of Korea, in 1995 and 1999, respectively. Currently, he is a professor in the department of computer education, Mokwon University, Daejeon, Rep. of Korea. His research interests include question answering, Semantic Web, and information retrieval.