KOSAC: A Full-fledged Korean Sentiment Analysis Corpus

Hayeon Jang, Munhyong Kim, and Hyopil Shin

Department of Linguistics, Seoul National University
Gwanak-no Gwanak-gu, Seoul 151-741
{hyan05, likerainsun, hpshin}@snu.ac.kr

Abstract

This paper aims to introduce the Korean Sentiment Analysis Corpus named KOSAC. KOSAC is a corpus consisting of 332 news articles taken from the Sejong Syntactic Parsed Corpus. These sentences have been manually-tagged for sentimental features. The corpus includes 7,713 sentence subjectivity tags and 17,615 opinionated expression tags based on the annotation scheme called KSML which reflects the characteristics of the Korean language. The results of sentence subjectivity and polarity classification experiements using the corpus show the wide possibilities of application the KSML scheme and the tagged information of the KOSAC comprehensively to other corpus. What is innovative about our work is that it pulls together both the concept of private states and nested-sources into one linguistic annotation scheme. We believe that this corpus could be used by researchers as a gold standard for various NLP tasks related to sentiment analysis.

1 Introduction

There has been much research on the automatic identification and extraction of opinions and sentiments in text. Researchers from many subareas of Artificial Intelligence and Natural Language Processing (NLP) have been working on the automatic identification of opinions and related tasks. To date, most such work has focused on opinion, sentiment or subjectivity classification at the document or sentence level. A common sentiment analysis task is to classify documents or sentences by whether they are subjective or objective, and, if the target text is subjective, to classify it as positive or negative (Pang et al., 2002; Wiebe et al., 2005).

Along with these lines of research, a need for corpora annotated with rich information about opinions and emotions has emerged. In particular, statistical and machine learning approaches have become the method of choice for constructing a wide variety of practical NLP applications. These methods, however, typically require training and test corpora that have been manually annotated with respect to each language-processing task to be acquired. As such a resource, the Multi-perspective Question Answering (MPQA) Opinion Corpus plays an important role in sentiment analysis.

The goal of this paper is to introduce the Korean Sentiment Analysis Corpus, KOSAC[1]. We received two years of support (May, 2011-April, 2013) in this corpus construction project from the Korean Research Foundation (KRF). In the first year of the project, we focused on a fine-grained annotation scheme called KSML (Shin et al., 2012) that identifies key components and properties of sentiments based on solid theoretical background. The annotation scheme has been employed in the manual annotation of a 7,713-sentence corpus of 332 news articles from the Sejong syntactic parsed corpus. This manually-tagged corpus includes 17,615 opinionated expression tags.

The remainder of this paper is organized as follows. Section 2 gives an overview of KSML focused on differences with the annotation scheme of the MPQA. Section 3 describes observations about KOSAC. Section 4 presents the results of subjectivity and polarity classification experiments using the corpus. Section 5 presents conclusions and discusses future work.

[1] http://word.snu.ac.kr/kosac

27th Pacific Asia Conference on Language, Information, and Computation pages 366－373

2 Markup Language: KSML

The MPQA Opinion Corpus began with the conceptual structure for private states in Wiebe (2002) and developed manual annotation instructions (Wiebe et al., 2005; Wilson, 2008). Documents contained in the MPQA version 2.0 corpus are mostly news articles. It contains 461 documents spanning 80,706 sentences, 216,080 tokens, and 10,315 subjective expressions annotated with links. These subjective expressions are annotated with "attitude types" indicating what type of subjectivity they invoked. 5,127 of these subjective expressions convey sentiment. Since this corpus provides rich annotated expressions based on a fine-grained annotation scheme, it is widely used as a source for training data in machine learning approaches and serves as the gold standard in many sentiment analysis tasks. Since, we took advantage of the MPQA as a fundamental resource for sentiment corpus construction in Korean.

In the first year of the project constructing the Korean Sentiment Analysis Corpus, we focused on the theoretical background for the annotation scheme named the Korean Subjectivity Markup Language (KSML). Shin et al. (2012) provides a solid theoretical background for the corpus and described the results of inter-annotator agreement test with a view to improving the annotation scheme. Our work essentially follows the idea of the annotation scheme of the MPQA, but we have modified the existing framework and attributes in order to address the characteristics of Korean. In this section, we give an overview of KSML focused on differences with the annotation scheme of the MPQA.

2.1 Annotation Framework

First of all, the annotation frame of the MPQA is classified as six types by functions and meanings of the expressions regardless of the tagging unit: *type-agent*, *expressive-subjectivity*, *direct-subjective*, *objective-speech-event*, *attitude*, and *target*. Each unit could connect by various links such as target-links or attitude-links.

The KSML, however, divides tagging units as the whole sentences and smaller expressions included in the sentences. The *subjectivity* and *objectivity* present the subjectivity of the whole sentence by reflecting whether an annotator feel the sentence is objectively true or not in terms of the speech event.

```
anchor: morpheme id(s)
id: tag id
nested-source: w-(morpheme id(s)
       |implicit|out)-...-
       (morpheme id(s)|implicit|out)
target: morpheme id(s)
type: direct-explicit,direct-speech,
     direct-action,indirect,
     writing-device
subjectivity-type: emotion-{pos,neg,
     complex,neutral},judgment-{pos,
     neg,complex,neutral},argument-
     {pos,neg,complex,neutral},
     agreement-{pos,neg,neutral},
     intention-{pos,neg},
     speculation-{pos,neg}, others
polarity: positive,negative,complex,
         neutral
intensity: low,medium,high
```

Table 1: The list of SEED tag attributes

In a SEED tag, each individual unit which is smaller than a sentence expresses a private state. The KSML describes information related to subjectivity such as source, target, and subjectivity-type by using attributes of a SEED tag without any links. Table 1 shows the attributes.

2.2 Change of Attributes

`Type` attributes specify either speech events (acts) that express private states or non-speech events. These fit into five subtypes: `direct-explicit`, `direct-speech`, `direct-action`, `indirect`, and `writing-device`. The *expressive-subjectivity* of the MPQA corpus matches the `indirect` type in the KSML. The *attitude* of the MPQA is expressed by `subjectivity-type` in the KSML. The *direct-subjective* of the MPQA corpus classifies `direct-explicit`, `action`, or `speech` types in the KSML depending on the exact nature of the subjectivity. These tags group direct expressions together by the way of express opinions or emotions. Such classification could show different shades of expressed sentiments. The MPQA does not have a specific tag for direct subjective speech events. The *objective-speech-event* of the MPQA is `direct-speech` type expressions of a sentence having an `objectivity` tag in the KSML frame.

The `writing-device` is a newly added attribute to KSML in order to show writers' own subjectivity through non-predicate expressions.

Modal expressions, speaker-oriented adverbials, conjunctive endings, and special functional particles get `writing-device` tags as kinds of devices reflecting sentiments in texts. As a basic annotation unit, we chose a morpheme rather than a word because Korean is an agglutinative language having many meaning-bearing particles and sentence endings which can carry private states. We need to be able to pinpoint precise segments as a basic unit, especially when finding `writing-device` expressions. Since some endings and particles show the subjectivity of a sentence having no direct opinionated expressions, `writing-device` expressions usually have high intensity of subjectivity. Various expressive techniques like *contrast, inferred, repetition*, and *sarcastic* of the MPQA could be classified as `writing-device` in the KSML.

The framework of the MPQA is similar to that of Appraisal Theory by Martin (2002) and White (2002). The Appraisal framework is composed of concepts including *Affect, Judgment, Appreciation, Engagement*, and *Amplification*. *Affect, Judgment*, and *Appreciation* represent different types of positive and negative attitudes. Nonetheless, the MPQA corpus does not distinguish different types of private states like *Affect* and *Judgment*, which can provide useful information in sentiment analysis. On the other hand, the MPQA corpus distinguished different ways that private states may be expressed, such as *directly* or *indirectly*. The KSML, however, not only cover many types of attitudes as in Appraisal theory but also several expressive types as in the MPQA corpus. For example, we added a `Judgment` attribute to the `subjectivity-type` in KSML.

Each attributes of `subjectivity-type` except `others` has directional cues like `positive`, `negative`, `complex`, and `neutral`. Unlike the MPQA, the KSML adds `neutral` and `complex` directional cues. In addition, the `speculation` attribute also has directional cues. Directional cues express semantic orientations of `subjectivity-type` tags. Such detailed classification provides the benefits in the process of sentiment analysis.

2.3 Sentence Tagging Examples

So far we describe the KSML as an annotation scheme for the Korean Sentiment Analysis Corpus with a focus on the differences with the MPQA annotation scheme.

On Saturday he met representatives of two warlords who clashed violently last week over who should be governor in eastern Paktia province.
The MPQA annotation scheme
GATE_objective-speech-event nested-source=w implicit=true GATE_direct-subjective: *clashed violently* nested-souce=w,*warlords* polarity=negative expression-intensity=high intensity=high GATE_agent: *two warlords* id=*warlords* nested-source=w,*warlords*
The KSML annotation scheme
Objectivity tag SEED: *clashed over* nested-souce=w,*warlords* type=dir-explicit subjectivity-type=agreement-negative polarity=negative intensity=high target=*who should be governor in eastern Paktia province* SEED: *violently* nested-souce=w type=indirect subjectivity-type=judgment-negative polarity=negative intensity=high target=*clashed over*

Table 2: Tagging examples of the MPQA and KSML

As an end of this section, the sentence tagging examples in Table 2 show the different tagging aspects according to the annotation schemes. The sample sentence and the example tags of the MPQA are brought from the existing MPQA corpus, and the tagging example of the KSML is made by an annotator who participated in the project constructing the Korean Sentiment Analysis Corpus. Compared to the MPQA scheme, the frame of the KSML is simpler and easier to understand in terms of subjectivity included in the sentence because the KSML grabs opinionated expressions in detail.

3 Sentiment Corpus: KOSAC

3.1 Corpus Selection

Unlike English, Korean is a morphologically rich language, so, rather than words, morphemes should be the units of annotations. However, it is

too time consuming to build a flawless morphologically parsed corpus due to the inaccuracy of part of speech (POS) taggers. For this reason, the Sejong syntactic parsed corpus, which is semi-automatically built, was used as the basis for the sentiment annotation corpus. Syntactic information of sentences is also available, enabling further logical inference on agents or targets of sentimental expressions.

A subset containing a total of 332 articles made up of 7,713 sentences was selected from the Sejong corpus newspaper articles. These articles were taken from the society and life subsections of Hankyoreh and Chosun, the editorial section of Hankook.

3.2 Annotation Process

The size of corpus largely depends on the speed of annotation work. Without an appropriate annotation tool, it is almost impossible to build a large annotated corpus.

Though the MPQA opinion corpus was built with GATE annotation tool, we developed a morpheme based annotation tool for Korean text (Cattle et al., 2013) for three reasons. First, none of current annotation tools, such as GATE or brat, supported switching between word and morpheme views. Second, there are non-continuous sentiment expressions that cannot be annotated by current tools. Third, targets and nested-sources of sentiment expressions need to be annotated in advance of sentiment expressions within those tools, which is not intuitive and in turn makes process of annotation slow. Moreover, to ensure the quality of annotations, three well-trained linguistics students annotated separately, and then double cross-checked the annotations until all annotators agree on the same annotations.

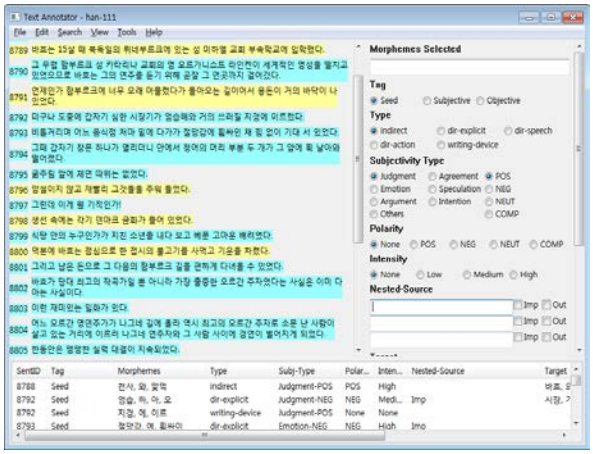

Figure 1: Morpheme Based Annotation Tool

3.3 Annotated Expressions

The accuracy of an annotated corpus is difficult to measure. For KOSAC, twenty frequently occurring sentiment expressions were chosen from six subjectivity types to see how consistently people annotated those expressions. For measurement, the ratio of annotated times to the number of occurring times for each of those expressions is shown in Table 3.

Emotion	ratio	Agreement	ratio	Argument	ratio
두렵- twulyep- ‘fear’	1.00	합의하- hapuyha- ‘agree’	0.86	주장하- cwucangha- ‘insisit’	0.98
분노 pwunno ‘anger’	0.93	인정하- incengha- ‘admit’	0.90	지적하- cicekha- ‘point out’	0.90
사랑하- salangha- ‘love’	0.94	반대하- pantayha- ‘disagree’	1.00	제시하- ceysiha- ‘suggest’	0.82
행복하- hayngpokha- ‘happy’	0.94	거부하- kepwuha- ‘deny’	0.90		

Intention	ratio	Judgement	ratio	Speculation	ratio
-고 싶- -ko siph- ‘want’	0.88	인기 inki ‘popular’	0.87	-는 것 같- -nun kes kath- ‘might’	0.50
-기 위하- -ki wiha- ‘purpose’	0.63	재미 caymi ‘fun’	0.59	-을 것 -ul kes ‘would’	0.20
-도록 -tolok ‘purpose’	0.52	중요하- Cwungyoha- ‘important’	0.90	예상 되- yeysang toy- ‘expected’	1.00
예정 yeyceng ‘plan’	0.61	풍부하- Phwungpwuha- ‘plentiful’	0.91		

Table 3: Frequency Cross Table of Expressive and Subjectivity Type

	Agree.	Argu.	Emotion	Intention	Judgment	Speculation	Others
Dir-Action	1	9	71	8	38	0	1
Dir-Explicit	156	277	341	276	2740	157	40
Dir-Speech	8	1149	22	28	86	13	7
Indirect	255	321	720	409	6086	63	22
Writing-Device	5	98	9	306	764	172	2957

Table 4: Frequency Cross Table of Expressive and Subjectivity Type

Among the 7,713 sentences, 2,658 are annotated as subjective and 5,055 sentences as objective. There are 17,615 SEED tags, indicating on average 2.3 expressions tagged as SEED per sentence.

Of the 17,615 SEED annotations, the frequencies of `type` and `subjectivity-type` are given in Table 4. As seen above, the `judgment` subjectivity type is the most predominant type since `judgment` subjectivity type expressions include not just short sentiment words or phrases, but also clauses that show speakers' judgments. Among subtypes of `type`, `indirect` expressions include all sentiment expressions except all main predicates and `writing-device` expressions; accordingly `indirect` type is also the most frequent type of all. A large portion of `writing-device` expressions are categorized `others` subjectivity type because they do not usually belong to any other subjectivity types. To help understand which expressions belong to such types above and how they are annotated, Table 5 shows some examples of some types.

`Direct-explicit` & `Agreement`		
뜻을 모으-	ttusul mou-	'agree'
결의하-	kyeluyha-	'resolve'
반발이 강하-	panpali kangha-	'strongly oppose'
`Direct-action` & `Emotion`		
눈물이 흐르-	nwunmwuli hulu-	'tear drops'
얼싸안-	elssaan-	'hug'
킥킥거리-	khikkhikkeli-	'giggle'
`Writing-device` & `Judgment`		
하지못하면	haci moshamyen	'if do not do (it)'
제아무리	ceyamwuli	'even if'
오히려	ohilye	'rather'

Table 5: Examples of Annotated Expressions

From the examples above, it can be seen that annotated expressions are not restricted to specific syntactic segments, but rather capture segments which reveal one's subjectivity. Also, it is noticeable that intensifiers are not separated from sentiment expressions.

From the fine-grained annotated corpus, characteristics of a subjective or an objective sentence could be described by frequencies of type and subjectivity types.

Type	Objective	Subjective
direct-action	0.015772	0.017097
direct-explicit	0.374925	0.794073
direct-speech	0.225594	0.067629
indirect	0.678179	1.679711
writing-device	0.354761	0.946809
Subjectivity Type	Objective	Subjective
Agreement	0.041925	0.079787
Argument	0.270313	0.18845
Emotion	0.116191	0.216565
Intention	0.118387	0.162234
Judgment	0.830904	2.087006
Speculation	0.030146	0.094225
Others	0.241366	0.677052
Number of SEEDs	1.649231	3.505319

Table 6: Average Frequencies of Types for Objective and Subjective Sentences.

For an objective or a subjective sentence, how many types and subjectivity types it has on average is shown in Table 6. A subjective sentence tends to have more `direct-explicit`, `indirect`, `writing-device` types than an objective sentence. The frequency of the `direct-speech` type is higher for objective sentences due to the reporting predicates. For subjectivity type, a subjective sentence has particularly higher frequency of `judgment`, `speculation`, `emotion`, and `others` than an objective sentence. Also the number of SEED

tags for a subjective sentence is the double of that for an objective.

4 Experiments

4.1 Subjectivity Classification

Firstly, a subjectivity classification test was done by using frequency features from sentence tag attributes. To guarantee the experiment result, a 10-fold cross validation was used; 1/10 is used as a test set and 9/10 as a training set. As a classification model, SVMlight (Joachims, 2002) was chosen using a linear kernel and default options.

Since there could be too many frequency features from attributes, a pair of features was tested to classify sentence subjectivity, and then features were added one by one until the accuracy of SVM began to drop to find the most effective feature set. In detail, we identified the effectiveness of the attributes of SEED tags in terms of classifying polarity of a sentence by adding each attribute feature to the most efficient pairs as per the previous experiment. If an added attribute showed a better result, the combination would be the base pair for the next experiment.

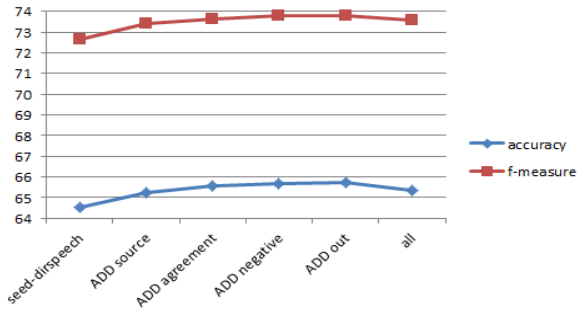

Figure 2: The result of polarity classification tests

Figure 2 shows experimental results of subjectivity classification. The best pair of features was the number of SEED tags and the `direct-speech` frequency, so another feature was added to the pair until the accuracy dropped. In the end, it was found that the best result was a feature set of the number of SEEDs, `direct-speech`, `nested-source`, `agreement`, `out` (`nested-source`), and `negative` value of `polarity`. The best performance of the SVM classifier was accuracy 65.72%, precision 59.76%, recall 96.41%, F-measure 73.78%.

However, the best classification result by SVM is not satisfactory, even though this test was done within a gold standard data. The reason was that sentence subjectivity surprisingly does not depend on the frequency of attributes. Rather, it is decided how a sentence ends. It is intuitively noticeable that a subjective sentence has features that make it subjective, and an objective sentence does not. We found almost all subjective sentences end with expressions that have a `direct-explicit` tag or include a `writing-device` seed. Among subjective sentences, 84.9% included a `direct-explicit` or `writing-device` seed. Table 7 shows how much sentence subjectivity depends on `direct-explicit` and `writing-device` expressions. Furthermore, the position of `writing-device` expression is important for the subjectivity of a sentence; a subjective sentence tends to have it within a main clause or close to main predicate.

Type	Subjective Sent.(1)	Objective Sent	(1) / Total Subj Sent
D-E	2102	935	2102/2658 (79.08%)
W-D	1543	1197	1543/2658 (58.05%)

Table 7: Ratio of `direct-explicit` and `writing-device` for Sentence Subjectivity

4.2 Polarity Classification

Secondly, sentence polarity classification experiments were conducted. The experimental method was the same as the sentence subjectivity classification experiments. The following Figure 3 shows the best results and the experimental result of using all attributes.

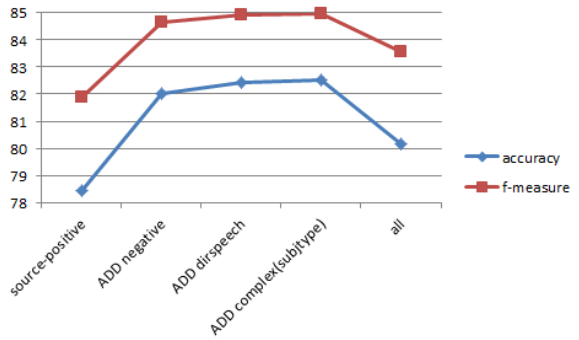

Figure 3: The result of polarity classification tests

Attributes leading the best results (Accuracy 82.52%, Precision 77.64%, Recall 93.93%, F-measure 84.96%) in the sentence polarity classification experiments were the number of `nested-source`, `positive` (`polarity`), `negative` (`polarity`), `direct-speech` (`type`), and `complex` (directional cue of `subjectivity-type`).

Among the contributory features in the experiment, the directional cue `complex`, which combines only with `emotion`, `judgment`, and `argument` subtype of `subjectivity-type`, is worthy of notice. These subtypes express private states in a relatively direct way and so the intensity of expressions is usually higher than other subtypes. In such aspects, polarity of expressions classified as these subtypes would be easier.

We suppose that the characteristics of news articles are the reason why `nested-source` and `direct-speech` (`type`) are the main features in the experimental results. In general, writers of news articles try to maintain objective distance. When citing other people's comments or statements, however, they have to convey the exact words of the speaker. Therefore, cited sentences could include more direct opinionated expressions showing obvious polarity. A number of `nested-source` and `direct-speech` (`type`) are important factors to distinguish whether an expression is a writer's own thinking or a citation of another's utterance.

In another manner, we can classify polarity of a sentence simply by checking for the inclusion of specific attributes. Checking attributes can be different according to the corpus. In the experiment using KOSAC corpus, we only used three attributes of SEED tags: `type` (only `direct` subtypes), `polarity`, and `intensity`. Table 8 describes the algorithm to classify polarity of a sentence by checking these attributes.

Through this checking algorithm, we obtained an 82.15% accuracy on sentence polarity classification. This result is slightly lower than the best experimental result using the SVMlight. However, considering that many sentences could slip through the net of checking at any phase of the algorithm since the algorithm is too simple, such accuracy can be rated high. In addition, this method does not need any other classifier, and we can get good results by using attributes which are understood intuitively as important factors in classification of polarity.

For all sentences in the KOSAC corpus,

1. if a sentences have SEED tags of `direct` subtypes,
   for only corresponding SEED tags,
   A. if a number of `positive` polarity tags and a number of `negative` polarity tags are different, classify the sentence as the bigger polarity.
   B. else,
      i. if `intensity` values of the polarity tags are different, classify the sentence as the polarity having the highest intensity value.
      ii. else, classify the sentence as the polarity having `dir-explicit` type value.
2. else,
   for every SEED tags,
   do the same process of phase 1.

Table 8: Checking algorithm for polarity classification

Therefore, we confirm that the theoretical background forming the KSML annotation scheme is highly effective at describing subjectivity and polarity of opinionated expressions.

5 Conclusion and Future Work

This paper described a fine-grained annotation scheme KSML and the manually-annotated Korean Sentiment Analysis Corpus, KOSAC. This scheme pulls together into one linguistic annotation scheme both the concept of private states and nested source based on the MPQA. However, the frame and some attributes were modified in order to reflect the characteristics of Korean language. The scheme was applied comprehensively to a large 7,713-sentence corpus. Several examples illustrating the scheme and basic observations of the corpus were described in section 3. The results of sentence subjectivity and polarity classification experiments using the corpus were also presented in section 4. Such experimental results show wide possibilities of application of the KSML annotation scheme and the KOSAC corpus.

The main goal behind the KSML and KOSAC is to support the development and evaluation of NLP systems that exploit opinions and sentiments in applications. Our hope is that including rich information of opinionated expressions in our corpus annotations will contribute to a new understanding of how

sentiments are expressed linguistically in Korean language. We hope this work will be useful to others working in corpus-based explorations of subjective language and that it will encourage NLP researchers to experiment with subjective language in their applications.

Acknowledgments

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-327-A00322).

References

Cattle, Andrew, Munhyong Kim, and Hyopil Shin. 2013. Morpheme-based Annotation Tool for Korean Text. In Proceedings of the American Association for Corpus Linguistics.

Joachims, Thorsten. 2002. Learning to Classify Text Using Support Vector Machines. Ph.D Dissertation, Cornell University.

Martin, J.R. 2002. Appraisal: An overview. http://www.grammatics.com/appraisal/AppraisalGuide/UnFramed/Appraisal-Overview.htm

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 79-86.

Shin, Hyopil, Munhyong Kim, Yu-Mi Jo, Hayeon Jang, and Andrew Cattle. 2012. Annotation Scheme for Constructing Sentiment Corpus in Korean. In proceedings of the 26th Pacific Asia Conference on Language, Information and Compuation, pages 181-190.

White, P.R. 2002. Appraisal-the language of evaluation and stance. In Jef Verschueren, Jan-Ola Ostman, Jan Blommaert, and Chris Bulcaen, editors, Handbook of Pragmatics, pages 1-27.

Wiebe, Janyce. 2002. Instructions for Annotating Opinions in Newspaper Articles. Departiment of Computer Science Technical Report TR-02-101, University of Pittsburgh.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, 39(2/3):164-210.

Wilson, Theresa Ann. 2008. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. Ph.D Dissertation, Brandeis University.