OPINION TOPIC, HOLDER AND POLARITY IN TEXTS: EXPLORATION AND AUTOMATIC IDENTIFICATION FROM CROSS-LINGUAL DATA

BY

KYOUNG-YOUNG KIM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Roxana Girju, Chair
Professor Richard Sproat, Director of Research
Professor Peter Lasersohn
Associate Professor Chengxiang Zhai

ABSTRACT

People express their opinions in various ways in different domains. With the growing interest in what other people think, mining opinions in texts has been the focus of attention for researchers in many different fields. Also, with the rapid development of technology and the internet, more and more multilingual and multicultural information has become available on the web. The objective of the present dissertation is exploring and automatically extracting opinions from multilingual corpora. In pursuing this objective, a bilingual opinion-annotated corpus was constructed focusing on detailed opinion factors with editorial texts. Annotated opinion factors include the holder of an opinion (Holder) and the topic of an opinion with its polarity (Positive Topic, Negative Topic). Factors used to express opinions as well as opinions across languages were investigated with the annotated corpus. The main contribution of this dissertation is the proposal of a multilingual sentiment analysis system for identifying opinion factors using a novel method that explores the linguistic structures used to express opinions. Without using pre-labeled opinion words, this multilingual sentiment analysis system directly identifies opinion factors using syntactic analysis, predicate-argument structure and pragmatic analysis. In the place of pre-labeled opinion words for each language, a clustered lexicon was constructed from bilingual dictionaries. Lexical features crucial for identifying the polarity were learned automatically. In addition to the lexical features, syntactic, morphological and contextual

features were used in the learning algorithm. The syntactic structure of the sentence as well as predicate-argument structures extracted from the Propbank database were investigated and used to assign appropriate features to the target chunk. The experimental results show that the proposed system is significantly more successful than a baseline system. Experiments focusing on each novel method verify that both the clustered lexical dictionary and incorporating more linguistic structures benefit the accuracy of opinion factor extraction. The proposed system was also tested with an existing English monolingual corpus (MPQA corpus) composed of news articles, and yielded consistent results with the annotated corpus. With the experimental set-up of multilingual analysis, the way that opinions are expressed across languages was investigated and utilized to improve the results of the analysis. Experiments with cross-lingual features extracted from parallel sentences show even more improved results, which suggests cross-lingual reinforcement in identifying opinion factors with the proposed system.

To my family, for their love and support.

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to my advisor, Prof. Richard Sproat for his support and guidance throughout all the years of my doctoral study. He always encouraged me and provided insight on the full range of computational linguistics so that I may truly be a researcher. His insightful comments made my research much more enjoyable and let me bravely move forward when I met obstacles.

I also thank the other members of my dissertation committee, Prof. Roxana Girju, Prof. Chengxiang Zhai, and Prof. Peter Lasersohn for their constructive and critical suggestions for the dissertation. I especially would like to express my sincere thanks to Prof. Roxana Girju for serving as the chair of my committee and for her valuable advice on the corpus annotation.

Thanks to Erica Britt, Eun-Kyoung Lee, Heejin Kim, Su-Youn Yoon, Young-Sun Lee, Yoonsook Mo and my colleagues in the Linguistics department, who have been with me as study buddies as well as lifetime friends. I also thank Suna Woo for helping with the annotation which is the basis of my research.

Finally, I give many thanks to my loving family. I would not have completed this dissertation without the patience and support of my soul mate, Byung-il Kwak. I'm really proud that we both made it though the hard time of our long distance relationship between North Carolina and Illinois. My parents and brothers' trust also made me what I am now. Also, thanks to my mother,

father and sisters in-law for their love and encouragement. Last, but not least, I send love to my little one, Benjamin Minje Kwak, who was born and grew up with this dissertation.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1
1.1 Subjectivity vs. opinion 2
1.2 Research Overview 5
1.2.1 Opinion factors in texts 5
1.2.2 Multilingual Sentiment Analysis System 7
1.3 Outline of the Dissertation 9

CHAPTER 2 RELATED WORKS 11
2.1 Opinion 11
2.1.1 Moral opinion: neither true nor false 11
2.1.2 Conveying opinion: a function of language 12
2.1.3 Evaluative opinion 14
2.2 Automatic analysis of opinions from texts 14
2.2.1 Subjectivity and sentiment analysis 15
2.2.2 Identification of opinion factors 17
2.2.3 Multilingual approach 19
2.2.4 Contribution of the dissertation 21

CHAPTER 3 OPINIONS IN TEXT: ANNOTATION 23
3.1 Corpus 25
3.2 Annotation scheme 26
3.2.1 Sentence polarity annotation 27
3.2.2 Topic with polarity annotation 29
3.2.3 Holder annotation 31
3.3 Factors determining opinion topic and polarity 32
3.3.1 Lexis 32
3.3.2 Grammar 34
3.3.3 Pragmatics 35
3.3.4 Context: beyond sentences 37
3.4 Inter-annotator agreement 37

CHAPTER 4 IDENTIFICATION OF OPINION TOPIC, HOLDER AND POLARITY FROM MULTILINGUAL CORPORA 41
4.1 Preprocessing of input sentences: chunking 42
4.2 Feature dictionaries 44
4.2.1 Lexical features 44
4.2.2 Syntactic features 46
4.2.3 Contextual features 48
4.2.4 Morphological features 49
4.3 Feature extraction 51
4.3.1 From the predicate-argument relationship 52
4.3.2 From the syntactic structure 56
4.3.3 Features for beyond chunked units 58
4.4 Machine learning algorithm 58
4.5 Experiment 60
4.5.1 Baseline 60
4.5.2 Result and discussion 62
4.6 Conclusion 73

CHAPTER 5 EXPRESSING OPINIONS ACROSS LANGUAGES . . 75
5.1 Bilingual sentence alignment 77
5.2 Polarity agreement between parallel sentences 80
5.3 Cross-lingual features 83
5.4 Experimental result 85
5.5 Conclusion 88

CHAPTER 6 CONCLUSION 94
6.1 Summary 94
6.2 Future work 96

REFERENCES 99

LIST OF TABLES

3.1 Statistics of the Bilingual-Editorial Corpus 26
3.2 Topics in the Bilingual-Editorial Corpus 26
3.3 Sentence polarity annotation 29
3.4 Topics with polarity annotation 31
3.5 Holder annotation 32
3.6 Inter-annotator agreement (Sentence polarity): Kappa statistics 38
3.7 Inter-annotator agreement (Sentence polarity): *Agr* ratio . . . 39
3.8 Inter-annotator agreement (Opinion Factors): Kappa statistics 39
3.9 Inter-annotator agreement (Opinion Factors): *Agr* ratio . . . 39
3.10 Inter-annotator agreement (All factors) 40

4.1 Chunking units: detailed noun phrase types 43
4.2 Syntactic features: higher path 48
4.3 Contextual features: sentence-based 49
4.4 Contextual features: discourse-based 49
4.5 Morphological features 51
4.6 Annotated information with predicates in the English PropBank 55
4.7 Evaluation: Proposed system vs. Baseline 64
4.8 Evaluation: Holder (H) identification of each step in the proposed system 67
4.9 Evaluation: Topic (T) identification of each step in the proposed system 68
4.10 Evaluation: Negative Topic (NT) identification of each step in the proposed system 68
4.11 Evaluation: Positive Topic (PT) identification of each step in the proposed system 69
4.12 Evaluation: MPQA corpus 72
4.13 Holder identification from the MPQA corpus 73
4.14 Topic identification from the MPQA corpus 73

5.1 Evaluation: bilingual sentence alignment 79
5.2 Evaluation: bilingual sentence alignment (Whole data accuracy) 79
5.3 Sentence polarity agreement between parallel sentences: Kappa statistics 81

5.4 Sentence polarity agreement between parallel sentences: *Agr* ratio 82
5.5 Agreement on opinion factor annotation between parallel sentences: *Agr* ratio 82
5.6 Evaluation: Effect of cross-lingual features 86
5.7 Effect of Cross-lingual features: Holder (H) identification . . . 89
5.8 Effect of Cross-lingual features: Topic (T) identification 89
5.9 Effect of Cross-lingual features: Negative Topic (NT) identification 90
5.10 Effect of Cross-lingual features: Positive Topic (PT) identification 90

LIST OF FIGURES

4.1 Schematic representation of the multilingual sentiment analysis system 42
4.2 Constructing clustered lexical feature dictionary 46
4.3 Examples of lexical feature 47
4.4 Example of Frame files in the PropBank: English 55
4.5 Argument labeling with PropBank database 57
4.6 Results with varying size of training data (F-score(%)): SYSTEM 70

5.1 Algorithm for Bilingual sentence alignment 78
5.2 Extracting cross-lingual features 84
5.3 Results with varying size of training data (F-score(%)): CROSS 87
5.4 Schematic representation of F-score(%) results: Holder identification 91
5.5 Schematic representation of F-score(%) results: Topic identification 91
5.6 Schematic representation of F-score(%) results: Negative Topic identification 92
5.7 Schematic representation of F-score(%) results: Positive Topic identification 92

CHAPTER 1

INTRODUCTION

With the rapid development of technology and the internet, the general public is not only receiving information from the web but also actively including this information in the formation of their private opinions. Mining opinions from web sources such as news articles and blogs has been the focus of many researchers in many different fields. The most popular domain dealing with opinions as primary information is the domain of review-related websites. Members of the general public, as well as the companies providing products, seek various opinions about products on the market. Editorials and public forums on various topics are other domains whose primary information is opinion. Government or political parties might want to track opinions from different holders on specific issues. Moreover, to correctly find answers to questions such as "How does X feel about Y?" in opinion-related question answering, more detailed opinion factors (X: holder, Y: target) should be identified.

On the other hand, when performing information retrieval or question-answering which seeks reliable answers such as "What is the highest mountain in the world?", opinions should be dealt with separately from fact as opinions may have more or less reliability depending upon the holder of the opinion. The performance of the information extraction (IE) system could be improved by filtering out opinion sentences using subjectivity classification (Riloff et al., 2005). That is, opinion is the element which should be

disregarded in this application as it could convey incorrect information. In this dissertation, opinions in texts were investigated with a focus on detailed opinion factors (including holder, topic and polarity). The method for expressing opinion factors in cross-lingual data was explored, with the aim of implementing an authentic multilingual sentiment analysis system to automatically extract opinion factors from text.

1.1 Subjectivity vs. opinion

The term *opinion* is used differently depending on the application in the sentiment analysis field. When we need to filter out opinions and seek reliable information, an opinion is defined as a subjective statement which is the opposite of an objective statement. Lyons (1977) describes the functions of language as descriptive, social and expressive. According to him, descriptive meaning is factual in the sense that "it can be explicitly asserted or denied and objectively verified". (p.50) Factual information which can be objectively verified is conveyed by an objective statement. This aspect of meaning conveying factual information is also described using labels such as referential, cognitive, propositional, ideational and designative. On the other hand, social and expressive meanings cannot be verified objectively. These two types of information are often subsumed under one label such as emotive, interpersonal, and attitudinal. Quirk et al. (1985) present verb types that convey information as factual, suasive, emotive and hypothesis classes. They further divide factual verbs into 'public' and 'private' types. Private types of verbs such as *believe* and *doubt* are not observable, so a statement with these verbs expresses *private state* which is "not open to objective observation or verification". (p.1181) In other words, a statement with a public

type of factual verb can be regarded as an objective statement and conveys factual information. However, verbs are not the only clues that a statement expresses private state; most obviously, adjectives and nouns can express private state. Wiebe et al. (2005) define subjective expressions as the words and phrases used to express private state, and further define private states in terms of their functional components. They denote private states as "states of *experiencers* holding *attitudes*, optionally toward *targets*". They set up the guidelines of opinion annotations including the factors stated above, and created the Multi-Perspective Question Answering Opinion Corpus (MPQA Corpus)[1], which was used by many researchers working on sentiment analysis thereafter.

As described above, subjectivity within a sentence could be determined with the use of predicates. For example, the sentence (1) can be deemed a subjective statement, as it contains private types of the factual verb *believe*. As the sentence is about what the subject in the main clause *I* believes, it is not open to objective observation or verification. The sentence (2) is also clearly identified as a subjective statement as it includes the speech-event of the source *He* expressing a positive opinion toward *this plan*.

(1) I believe you have to use the system to change it.

(2) He said, "This plan needs to be respected".

The definition and description of subjectivity, however, should be interpreted differently depending on the domain.

(3) The price is high.

(4) This restaurant is expensive.

[1] www.cs.pitt.edu/mpqa/databaserelease

Although the sentence (3) and the sentence (4) carry descriptive meanings so that they are assumed to be objective, they could also carry expressive meanings. More specifically, the subjectivity depends heavily on the reliability of the source of the statement. If these sentences are found in a news article, they can be assumed as objective statements that just describe fact without judgment in them. On the other hand, in the domain of user-reviews, the sentences clearly describe the user's opinion on a specific item, most likely negative.

In the domain of user-reviews and editorials, opinions on specific targets are the primary information that the texts deliver. The term *opinion* here can be interchangeable with the term *evaluation* which means "the writer's feeling, judgment or viewpoint about the entities or propositions that he or she is talking about" (Thompson and Hunston, 2000). Biber and Finegan (1989) use the term *stance* and present a list of stance markers defined as "the lexical and grammatical markers for expressing attitudes, feelings, judgments or commitment." (p.93) The sense of opinion is labeled *standpoint* as well to represent a statement which shows either affirmative or negative polarity (Eemeren et al., 1996). Martin and White (2005) introduce the "appraisal system" which emerged from Systemic Functional Linguistics (SFL)(Halliday, 1994) to investigate the language of evaluation. SFL identifies three metafunctions of language operating in parallel: ideational, interpersonal and textual. In the framework of SFL, language is interpreted as a resource for mapping the three metafunctions onto one another in an act of communication. Appraisal theory focuses on the interpersonal meaning and describes how social relationships are negotiated through evaluations of self, others and artifacts. Attitude types in appraisal theory are categorized into three types: *affect* construing emotional responses, *judgement* evaluating

according to a personal or moral code, and *appreciation* evaluating according to aesthetics or social significance. Positive/negative polarity about an opinion topic is encoded in this concept of attitude. In addition to the concept of attitude, the system defines *engagement* distinguishing various types of intersubjective positioning such as attribution and expectation, and *graduation* reflecting the degree of evaluation.

1.2 Research Overview

Two different goals are pursued in this dissertation. One of the aims of this dissertation is to investigate the method for expressing opinions in texts across languages. Language universality in representing opinions is hypothesized within linguistic structure, although the details and surface structures are language-dependent. Another aim is to implement a sentiment analysis system to automatically extract opinion factors (topic, holder and polarity) from multilingual corpora. The system pursued here is an authentic system which explores the linguistic structure of each language in order to induce cross-lingual reinforcement.

1.2.1 Opinion factors in texts

Most previous studies in sentiment analysis other than user-review domains focus on the subjective expression with an optional target as defined in (Wiebe et al., 2005). In the MPQA corpus, *experiences, attitude* and *targets* are annotated as opinion factors for each private state. The types of attitude they mark are categorized into positive and negative sentiment, agreement, arguing, intention, speculation and others.

The opinion factors that are focused on in this dissertation are the *topic,*

holder and *polarity* of an opinion. As a notion of "opinion", the definition and types of "attitude" in the appraisal system by Martin and White (2005) are adopted. That is, "opinion" in this dissertation refers to the positive or negative "evaluation" of a specific target. Therefore, the target of an opinion is a primary factor which is not optional. Stoyanov and Cardie (2008) distinguish the notion of *topic* of an opinion from the term *target* used in the MPQA corpus. They define the *topic* of an opinion as "the real-world object, event or abstract entity that is the subject of the opinion as intended by the opinion holder", and the *topic span* as "the closest, minimal span of text that mentions the topic". On the other hand, *target span* is used to denote "the span of text that covers the syntactic surface form comprising the contents of the opinion".

(5) [John] *adores* [Marseille] and visit it often.

(6) [Al] *thinks* that [the government should tax gas in order to curb CO_2 emissions].

(7) Although he doesn't like government-imposed taxes, he thinks that a fuel tax is the only effective solution.

For example, in the sentence (5), *John* and *Marseille* are the holder and target of the opinion *adores* respectively. It is likely that in the sentence (6), *Al* and *the government should tax gas in order to curb CO_2 emissions* are the holder and target of the opinion *thinks*. The target span *Marseille* in the sentence (5) can be considered as a topic span as well. On the other hand, there are several possible topic spans in the sentence (6) depending on the context:*the government, tax gas*, *CO_2 emissions*. Considering the following sentence (7), the topic of an opinion in the sentence (6) is determined as *tax gas* among the candidates. In this dissertation, I adopt the term "topic"

instead of "target" of an opinion, as the notion I am seeking in this study is "real-world object, event or abstract entity". Instead of determining whether or not the sentence contains opinions, specific opinion topics with polarity (Positive Topic, Negative Topic) are pursued in this dissertation in addition to the opinion holder. In many recent applications such as opinion-related question answering, the focus goes on the detailed opinion factors. In other words, instead of whole sentences containing subjectivity, specific opinion holders and topics should be identified to answer the question satisfactorily. Identification of detailed opinion factors could successfully meet the needs of this kind of application.

To deeply explore the representation of opinion factors across languages, editorial texts were chosen as the primary corpus. As the purpose of an editorial is to express an opinion on a set of issues, various patterns and indirect ways of expressing opinion are by nature present in editorials. This makes editorial texts an ideal but at the same time challenging dataset for sentiment analysis.

1.2.2 Multilingual Sentiment Analysis System

So far, the majority of previous studies on sentiment analysis have worked with monolingual texts, mostly English. With the increasing need to deal with non-English opinion corpora, studies on multilingual sentiment analysis are gaining in interest. Most studies, however, detect sentence subjectivity making use of systems based on English, using machine translation (Banea et al., 2008, 2010; Denecke, 2008; Mihalcea et al., 2007) The present dissertation aims to implement a multilingual sentiment analysis system to identify detailed opinion factors, based on English and Korean, which improves the

performance of the sentiment analysis task by exploiting the different ways that different languages and cultures have for couching opinions. To fulfill this objective, novel methods are adopted as below.

One-step of identifying opinion factors Most work on sentiment analysis starts with identifying opinion words, then, in a separate step, extracting the topic of an opinion anchored to the word. The present study, on the other hand, starts directly with identifying the topic and holder of an opinion without depending on whether or not the sentence contains opinion words. This approach has been motivated by the observation that it is dangerous to determine the subjectivity of the sentence only from the opinion words it contains. First, the definition of an opinion word is not clear. The judgment for opinion words is not always consistent with different persons. Second, there are sentences that express opinions without making use of explicit opinion words. Other linguistic factors such as grammar and pragmatics could induce opinion factors as well. Therefore, in this dissertation, opinion factors in a text are extracted using contextual information without a separate step of identifying opinion words. That is, starting directly with opinion factors within a sentence, clues for opinion factors are inferred through linguistic structure. This approach will make the authentic multilingual sentiment analysis theoretically and technically possible, by investigating and utilizing the underlying linguistic structure in expressing opinion.

Clustered feature dictionaries Possible linguistic features (lexical, syntactic and pragmatic) for opinion factors were designed and bilingual clustered feature dictionaries were constructed. To strengthen the appropriate features across languages while learning, linguistic features assumed to share

the value were clustered into one feature for machine learning. This strategy is also expected to deal with the data sparseness problem not only within monolingual data but also by utilizing bilingual data that shares features.

Utilizing linguistic structure As a way of exploring linguistic structure, automatic semantic role labeling has been used in previous studies to extract opinion factors from texts (Bethard et al., 2004; Choi et al., 2006; Kim and Hovy, 2006), and verified to make substantial contributions. A semantic role is defined as "the underlying relationship that a participant has with the main verb in a clause" (Payne, 1997). However, a semantic role anchored to a predicate cannot be an effective solution all the time (Ruppenhofer et al., 2008). In this dissertation, in addition to the semantic-role relationship, other linguistic structures such as syntactic and pragmatic structures are incorporated to identify opinion factors.

Cross-lingual reinforcement By way of utilizing features extracted from parallel sentences, the interaction between different languages is investigated. With the assumption that there exists universality in expressing opinions across languages, a cross-lingual feature dictionary was constructed from aligned parallel sentences in the parallel corpus.

1.3 Outline of the Dissertation

The remainder of this dissertation is structured as follows:

In chapter 2, a review of the literature on opinion definition and analysis is presented. Although the present dissertation is not exactly in line with the previous works on opinion mining, general reviews of sentiment analysis are presented to provide the implications of the current work. Works on subjec-

tivity and sentiment analysis in varying degrees of granularity (document, sentence and phrase level) are presented. Also, recent works on multilingual approaches to sentiment analysis are summarized.

In chapter 3, the annotation scheme and process are explained in detail. In addition to the corpus and annotation scheme, patterns for expressing opinions are presented with examples from the annotated corpus. The inter-annotator agreement for each annotation factor are presented in later in this chapter.

The experiments on the automatic extraction of opinion factors are described in chapter 4 and chapter 5.

In chapter 4, the preparation of the main system — multilingual sentiment analysis — is explained in detail including the feature dictionary and feature extraction. Experimental results from the baseline system and the proposed system are shown to verify the improvements in performance of the proposed system. Results from three more experiments are also shown which verify the effect of the individual factors of the system. Finally, the experimental results of the existing MPQA corpus are presented in addition to the experiment with the annotated corpus from chapter 3.

Chapter 5 focuses on cross-lingual effects in opinion factor extraction. The procedures for extracting cross-lingual features are described in detail, and the results of the experiments with cross-lingual features are presented compared with the results without the cross-lingual features from chapter 4.

Conclusion and directions of future study follow in chapter 6.

CHAPTER 2

RELATED WORKS

As mentioned in chapter 1, the term "opinion" is used differently depending on the application in computational linguistics. The adopted definition of "opinion" in this dissertation is evaluative opinion based on "attitude" as used in the appraisal system by Martin and White (2005). In this chapter, a general review of the literature concerning opinion and evaluation is presented in section 2.1 followed by computational approaches to automatically identify opinions from texts in section 2.2.

2.1 Opinion

2.1.1 Moral opinion: neither true nor false

In the view of ethical subjectivism, a moral opinion is based on the feelings of the holder who expresses it, nothing more (Rachels, 2007). As the first stage of this theory, Simple subjectivism interprets the expression that something is morally good or bad as the holder's approval or disapproval of the target. That is, if a speaker X says "Y is immoral," Simple subjectivism interprets this as a statement of the fact that "X disapproves of Y". Rachels (2007) states that Simple subjectivism faces criticism as it conflicts with the nature of moral evaluation. One of the objections is that Simple subjectivism cannot account for disagreement which surely exists between the utterances of two

people. Another objection is about "fallibility" which Simple subjectivism fails to account for, as it supposes every person expresses his or her feelings sincerely which, of course, is not always true. Emotivism is the improved version of simple subjectivism. Unlike Simple subjectivism which interprets moral judgments as statements *about* the speaker's attitude, Emotivism interprets moral judgments as expressions *of* attitude. Ayer (1952) states that moral judgments cannot be verified as they are mere "pseudo-concepts" irreducible to empirical concepts. He argues that the statement "X is wrong" has no factual meaning which can be either true or false, and it merely expresses moral sentiments. Stevenson (1944) agrees with Ayer (1952)'s concept of moral judgment, while adding an imperative component intended to change the listener's feelings. That is, the statement of "Y is immoral," is interpreted as something like "Y — so wrong", or "Don't do Y". In either case, in the view of Emotivism, a statement conveying moral opinion is neither true nor false.

2.1.2 Conveying opinion: a function of language

As described previously, a moral opinion is characterized as neither true nor false as it is not verifiable according to the Emotivistic view. In this section, how an opinion is represented with the use of language is investigated. In (Lyons, 1977), the functions of language are categorized into descriptive, social and expressive functions. According to Lyons, descriptive meaning is factual in the sense that "it can be explicitly asserted or denied and objectively verified" (p.50). Social and expressive meanings, on the other hand, cannot be verified objectively, which corresponds with the characteristics of moral opinion. These two types of information are often subsumed under one label

such as emotive, interpersonal, and attitudinal. Therefore, the distinction in meaning is redefined as referential and emotive meaning (or cognitive and affective meaning). Further, Lyons describes the "connotation" of a word as "an emotive or affective component additional to its central meaning"(p.176). In other words, an emotive or affective component of meaning that comes from the speaker's subjective idea is carried additionally through connotation. Stevenson (1944) also distinguishes pragmatic meanings into descriptive and emotive meanings. Emotive meaning here is defined as "a meaning in which the response (from the hearer's point of view) or the stimulus (from the speaker's point of view) is a range of emotions". (p.59) Halliday (1994) describes meaning in language with the view of its functional components. Based on the systemic theory, which considers a language as a resource for making meaning, the fundamental components of meaning in language are called "metafunctions" categorized into ideational, interpersonal and textual meanings. Ideational meaning is about construing a model of experience while interpersonal meaning is about enacting social relations between individuals, including the feelings they try to share. Textual meaning is concerned with creating relevance to context. Quirk et al. (1985) suggests the term *private state* which is "not open to objective observation or verification". Based on these definitions, the emotive meaning by Lyons (1977); Stevenson (1944), and the interpersonal meaning by Halliday (1994) can be said to convey "private state", and the statement carrying these meanings is a statement of opinion which is subjective.

2.1.3 Evaluative opinion

Thompson and Hunston (2000) present a distinction between 'opinions about entities' and 'opinions about propositions'. 'Opinions about entities', which are canonically attitudinal, involve positive or negative feelings, while 'opinions about propositions', which are canonically epistemic, involve degrees of certainty. Thompson uses the term *evaluation* to cover these two types of opinions meaning "the writer's feeling, judgment or viewpoint about the entities or propositions that he or she is talking about." Conrad and Biber (2000) presents a similar distinction between 'attitudinal stance' and 'epistemic stance' following the use of the term "stance" in (Biber and Finegan, 1989). In (Halliday, 1994), this distinction is represented as 'modality' and 'attitudinal meaning' within the category of interpersonal meaning. He categorizes modality into modalization, relating to probability and usuality and modulation, relating to obligation and inclination. Martin and White (2005) extend the account of the interpersonal meaning in (Halliday, 1994) focusing 'attitudinal meaning', then proposes the appraisal theory with three types of attitude: *affect, judgment* and *appreciation. Affect* is modeled as a semantic resource construing positive or negative emotional responses, while *judgment* refers to evaluating according to a personal or moral code. Finally, *appreciation* is a resource for positively or negatively evaluating products of behaviors according to a code of "aesthetics" or social significance.

2.2 Automatic analysis of opinions from texts

The earliest work on opinion analysis in computational linguistics was on identifying opinion expressions and has become the basis for further study. Researchers extract subjective words or phrases from dictionaries or cor-

pora, and add a positive or negative semantic orientation to the subjective expressions. Subjectivity/polarity detection using the extracted opinion expressions is pursued at various levels such as document, sentence and phrase. Detailed opinion factors such as holder, topic and polarity started to gain focus relatively recently. In addition to the monolingual sentiment analysis that use mostly English texts, multilingual sentiment analysis has been studied. In this chapter, a general review of the literature on opinion analysis is presented.

2.2.1 Subjectivity and sentiment analysis

Identifying opinion expressions in text has been the starting point for mining opinions by most previous researchers. Opinion expressions (words and phrases) are identified either from corpora (Breck et al., 2007; Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Wiebe, 2000; Wiebe et al., 2001) or dictionaries such as WordNet[1] (Kamps and Marx, 2002; Kim and Hovy, 2005b; Hu and Liu, 2004a; Takamura et al., 2005; Esuli and Sebastiani, 2005, 2006; Andreevskaia and Bergler, 2006).

Hatzivassiloglou and McKeown (1997) infer the semantic orientation of adjectives from conjunctions of conjoined adjectives. They start with the insight that a connective could be a strong clue for the semantic orientations of the adjectives connected with it: for most connectives except for "but", the conjoined adjectives tend to share the semantic orientation. Wiebe (2000) learns subjective adjectives from corpora using higher quality adjective features, such as polarity and gradability, seeded by a small amount of detailed manual annotation. Wiebe et al. (2001) include verbs as candidates of subjec-

[1]G. Miller., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton.edu/ wn

tive language and identify them from corpora. Riloff and Wiebe (2003) learn extraction patterns for subjective expressions using bootstrapping methods. In addition to isolated opinion expressions, Breck et al. (2007) identify the words and phrases used to express opinion in context. More detailed classifications of subjectivity, strong and weak opinion clauses are identified in Wilson et al. (2006).

In Kamps and Marx (2002), the semantic distance from a word from *good* and *bad* in WordNet is used as a classification criterion of the attitude of the word. The semantic orientation of a phrase is also calculated in Turney and Littman (2003) as the mutual information between the given phrase and the word *excellent* minus the mutual information between the given phrase and the word *poor*. Kim and Hovy (2005b) use WordNet to detect opinion-bearing words with the assumption that the synonyms of opinion words share their semantic orientation while antonyms contain an opposite semantic orientation.

Wiebe and Riloff (2005) detect subjectivity at the sentence level by following an unsupervised learning approach that uses unannotated texts for training. They achieve substantially higher recall than previous works by learning extraction patterns associated with subjective expressions (Riloff and Wiebe, 2003). Yu and Hatzivassiloglou (2003) perform both document level and sentence level classification for identifying the polarity of opinion sentences in addition to separating facts from opinions. Pang et al. (2002) also identify document polarity. They suggest that term occurrence is a more effective basis for review polarity than real-valued feature vectors unlike traditional Information Retrieval.

2.2.2 Identification of opinion factors

Beyond the level of identifying opinions in a sentence or document, the identification of opinion factors [topic,holder,sentiment] has been investigated by a number of researchers.

In review mining, the focus of research has been the sentiment toward an item such as product or movie. Hu and Liu (2004b,a) implement feature-based summaries of customer reviews of products sold online. In their studies, features of the products on which customers have expressed their opinions are identified first. For each feature, the polarity of the sentence is identified using seed adjectives tagged with positive and negative labels. Zhuang et al. (2006) try the mining and summarization of movie reviews incorporating WordNet, statistical analysis and movie knowledge. They use feature keywords and opinion keywords from labeled data. Taboada and Grieve (2004) apply the linguistic classification of appraisal by Martin and White (2005) in text classification of reviews. They calculate the semantic orientation of adjectives considering the text structure, then classify the review texts among one of three types of attitude (affect, judgement and appreciation) in the appraisal system. Whitelaw et al. (2005) extract adjectival appraisal groups for the document classification of user-reviews and show significantly improved accuracy. Opinions on commercial products from Weblogs were summarized by Mei et al. (2007). In their work, the mixture of topics and sentiment were captured simultaneously using a probabilistic model. In addition to the major topics with their polarities, the dynamics of each topic and the corresponding sentiments were summarized.

Outside the review mining domain, most research has been done with news media texts. Although most information in newspapers is factual, there ex-

ists some information containing the position of the news agencies. More frequently, opinions from third parties should be dealt with separately from facts. Algorithms to identify opinion holders given a topic (Kim and Hovy, 2004) and given an opinion expression (Kim and Hovy, 2005a) are implemented. Kim and Hovy (2004) start with sentences containing both the topic phrase and holder candidates. Then, the sentence sentiment classifier calculates the polarity of all sentiment-bearing words individually and combine them to produce the holder's sentiment for the whole sentence. Named entities such as person and organization are used as potential holders in their work. In Kim and Hovy (2005a), they add noun phrases to the holder candidates in addition to the named entities, and adopt syntactic parsing to extract features for machine learning. After parsing the sentence, features such as syntactic path information between each holder candidate and given opinion expression are extracted. Choi et al. (2005) also identify the holder of opinions using the conditional random field model and extraction patterns. They use capitalization feature, POS features, opinion lexicon features (binary features that indicate whether or not the words are in the opinion lexicon), dependency tree features and semantic features for semantic tagging via conditional random fields. The accuracy of holder identification has been improved by joint extraction of entities and relations (Choi et al., 2006). Bethard et al. (2004) focus on propositional opinions and extract their holders using the Support Vector Machine paradigm (Joachims, 2006) for semantic parsing.

Unlike source and polarity identification which have been studied by several researchers as presented above, topic identification has not been explored much other than in the domain of product reviews. Kim and Hovy (2006) identify the opinion with its holder and topic in news media texts by exploit-

ing the semantic structure of a sentence anchored to an opinion bearing verb or adjective. The first step of their algorithm is to identify an opinion-bearing word using manually annotated seed data (adjectives and verbs classified into positive, negative and neutral classes). The sentences are parsed using the Charniak parser (Charniak, 1999), and all constituents of the given sentence are collected as the possible topic and holder of the opinion. The semantic role of each selected candidate is determined using FrameNet annotated data (Barker and Sato, 2003). Another attempt to annotate and identify the topic of an opinion is Stoyanov and Cardie (2008)'s work to identify topic using topic coreference resolution. This system adds topic annotation in the existing MPQA Corpus, and extracts clusters of coreferent opinions in order to label the clusters with the name of the topic. Kim et al. (2008) extract opinion targets after determining topic-related opinions in the NTCIR-7 corpus using syntactic path information between opinion clues and an opinion target, and syntactic dependency features. Choi et al. (2010) also extract opinion targets in the NTCIR-8 corpus by considering document-level features and collocation between an opinion target and opinion clue words. Both of these works start with the assumption that opinion targets are related to document topics. Bloom and Argamon (2010) extract appraisal expressions with an unsupervised approach. After extracting attitude groups using a lexicon-based shallow parser, targets associated the attitude groups are identified.

2.2.3 Multilingual approach

Most of the previous studies on opinion mining described above have focused on English, as English is the most popular and has abundant linguistic resources. Studies on sentiment analysis in languages other than English as

well as multilingual sentiment analysis have been performed in two directions: applying English resources and systems to other languages by cross-lingual mapping, and performing the sentiment analysis separately for each language. Mihalcea et al. (2007) automatically generate resources for subjectivity analysis through the cross-lingual projection of available resources and tools for English on parallel corpora. Using an existing English subjectivity lexicon and bilingual dictionary, a subjectivity lexicon for Romanian is derived. With this constructed lexicon, a subjectivity sentence classifier for Romanian is developed. Also, from parallel corpora, a subjectivity-annotated corpus is obtained based on the English results for subjectivity. Denecke (2008) make use of an English resource (SentiWordNet) for multilingual document sentiment analysis of movie reviews. A translation of documents into English using standard machine translation software is performed first, then the sentiment of a document is classified using SentiWordNet and other existing English resources and tools. Bautin et al. (2008) identify the sentiment of entities in international news and blogs, depending on English resources and a state-of-the-art machine translation system. Wan (2009) detects polarity in Chinese product reviews using English training data after translating it into Chinese. He proposes co-traning with both unlabeled Chinese data and translated English data to detect polarity. Banea et al. (2008) perform sentence subjectivity analysis of Romanian using machine translation and English resources for subjectivity classification (OpinionFinder (Wiebe and Riloff, 2005)). They further show in (Banea et al., 2010) that multilingual data automatically translated into English provides benefits in the subjectivity classification of the source language, English, with the use of an additional lexicon drawn from translation.

Boiy and Moens (2009) perform sentiment analysis using a machine learning approach on blog, review and forum texts written in English, Dutch and French. They focus on the sentiment toward an entity (car brands and movie titles) within a sentence. A cascaded architecture of classifiers is used to reduce the computational cost, and active learning is performed to deal with the data sparseness problem. In addition to the features applied to all languages, language-specific features for each language are used, and verified to improve accuracy. Seki et al. (2009) extract opinion and opinion holders based on the differentiation between the author and authority viewpoint in Japanese, English and Chinese. They capture writing style differences such as syntactic constructions and term usages between author- and authority-opinionated sentences, and use features for each language.

2.2.4 Contribution of the dissertation

The aim of the present dissertation is distinguished from any of the previous studies.

First, detailed opinion factors are identified along with the opinion's polarity from general texts. Previously, opinion targets with polarity have been identified only from review-related texts. The targets of opinion in this domain are products or specific attributes of a certain products. Either case is different from the topics of the current work in that product names and attributes tend to be pre-defined in the reviews, so that they could be sufficiently identified through a statistical approach. In (Hu and Liu, 2004b,a), attributes of the products on which customers have expressed their opinions are identified through association rule mining (Agrawal and Srikant, 1994) to find frequent item sets. The opinion holder and topic from general texts

are identified by some previous researchers with news articles as the domain, but polarity is not considered in those works. Starting from the pre-identified opinion expressions, the opinion holders and topics have been identified separately. In the present dissertation, however, all opinion factors including topic, holder and polarity are identified at the same time by exploring linguistic structures.

Furthermore, the present dissertation seeks multilingual sentiment analysis. A multilingual approach in sentiment analysis started to gain focus recently, and several important attempts at a multilingual system are found in some recent studies. However, the previous studies mostly focus on expanding the monolingual system to a multilingual system by making use of cross-lingual projection or machine translation. The present dissertation aims to build an authentic multilingual system which could explore language universal as well as language specific features in order to induce cross-lingual reinforcement. A few recent studies (Boiy and Moens, 2009; Seki et al., 2009) explore the contexts or writing style used to express opinions in each language, but the cross-lingual relation is not explored or the domain is limited to review-related texts.

CHAPTER 3

OPINIONS IN TEXT: ANNOTATION

The primary aim of this dissertation is to investigate and extract evaluative opinions from multilingual corpora — English and Korean corpora are pursued in this dissertation. To fulfill these objectives successfully, one of the main procedures of this study is constructing a bilingual opinion-annotated corpus. Although a limited number of opinion-annotated corpora have been built, they are not exactly suited for the aim of this study in terms of the domain, annotated opinion factors and language.

The most broadly used gold standard annotation for sentiment analysis is the MPQA corpus (Wiebe et al., 2005; Wilson, 2008). This corpus contains news articles that are annotated in detail for subjective expression factors. The corpus adopts the notion of *private state*, "a state that is not open to objective observation or verification" suggested by Quirk et al. (1985), to define subjectivity, and focuses on three main types of private states: explicit mentions of private states, speech events expressing private states, and expressive subjective elements.

(1) "The U.S. **fears** a spill-over," said Xirao-Nima.

(2) "The repost is **full of absurdities**," Xirao-Nima **said**.

In the sentence (1), a private state of the source *U.S* about the target *a spill-over* is explicitly mentioned by the word *fears*. On the other hand, in the sentence (2), a private state of the source *Xirao-Nima* about the target

report is expressed by using the expressive subjective element *full of absurdities* in addition to explicitly mentioning it using *said*. Both sentences contain speech events that are either objective (1) or subjective (2). The primary factors annotated in this corpus are opinion expressions (*fears*, *full of absurdities* and *said*). Details such as opinion holder (source) and intensity are annotated anchored to the opinion expression. The opinions the authors define are subjective expressions, so the opinions in this corpus could contain neutral polarity in addition to positive or negative polarity, unlike evaluative opinions. The newest version (2.0) additionally includes attitude and target annotation when they are present based on the annotation scheme explained in (Wilson, 2008). Attitude types are categorized into sentiment, agreement, arguing, intention, speculation and others.

With book reviews as the corpus, Read et al. (2007) annotate expressions of appraisal in English; appraisal-bearing terms with detailed appraisal types are annotated. 38 documents containing a total of 1,245 sentences from the websites of four British newspapers (The Guardian, The Independent, The Telegraph, and The Times) on two different dates make up this corpus. It is shown that inter-annotator agreement varies depending on the level of abstraction in the appraisal theory.

In this dissertation, I focus an "evaluative opinion" which corresponds to the "attitude" in the appraisal system (Martin and White, 2005). As opinions defined here evaluate some targets, the polarity and the target of an opinion are the primary factors. In addition, I investigate various patterns of expressing opinions including subtle and indirect patterns. Therefore, I choose bilingual editorial texts as a primary corpus for annotation and use in the sentiment analysis system. Moreover, many more sentences in the editorial texts are opinionated than typically occur in regular news media

texts, as the purpose of editorials is to express opinions. In the rest of this chapter, the corpus, annotation scheme, and patterns of expressing opinions are described in detail. Finally, the results of inter-annotator agreement are shown for sentence polarity as well as for each of the opinion factors.

3.1 Corpus

English and Korean editorial texts were collected and annotated to build a bilingual opinion-annotated corpus. Editorials are classified as a representative of public argumentative text (Werlich, 1976). As the purpose of an editorial is to express an opinion on a set of issues, various patterns and indirect ways of expressing opinion are present in editorials by nature. This makes editorial texts an ideal but at the same time challenging dataset for sentiment analysis, requiring the use of techniques that are less dependent on opinion expressions. Editorials are also known to show difference in style depending on the culture (Tirkkonen-Condit, 1994). Therefore, it is expected that the cross-lingual differences in expressing opinions in editorials are caused not only from the linguistic structures but also cultural differences.

Data were collected from three different news agencies through online resources [1] dated from March 2007 to November 2007. Corpus statistics and topics in each language corpus are illustrated in Table 3.1 and Table 3.2 respectively. As shown in Table 3.1, the number of words in the Korean corpus is much smaller than those in the English corpus although the number of sentences in the Korean corpus is bigger. One of the possible reasons is that Korean is an agglutinative language whose grammatical markers are attached to the content words. Therefore, morphological analysis is necessary

[1] http://english.donga.com/editorial/, http://www.hani.co.kr, http://www.joins.com

Table 3.1: Statistics of the Bilingual-Editorial Corpus

	English	Korean
Documents	113	121
Sentences	2553	2824
Words	52643	32178

Table 3.2: Topics in the Bilingual-Editorial Corpus

Topics	No. of documents	
	English	Korean
Culture	10	10
Economics	17	19
Education	11	14
International	11	10
North Korea	21	23
Politics	30	30
Science	2	1
Society	11	11

for Korean to be processed in the sentiment analysis system. The editorials are about various topics including politics, economics and education.

Most (105) of 121 Korean and 113 English documents are parallel texts which are translations of Korean texts into English. However, the parallel texts in this corpus are not always direct translations of each other. Some English texts only contain the summary of Korean texts. More frequently, sentences are not matched one-by-one between parallel texts.

3.2 Annotation scheme

Most existing opinion-annotated corpora annotate opinion expressions as primary factors. In the MPQA corpus, the occurrences of subjective expressions used to express private state and their functional components (experiencers,

attitudes and optional targets) are annotated. Targets of opinions are added in the MPQA corpus anchored to the pre-annotated opinion expressions (Wilson, 2008), and Stoyanov and Cardie (2008) add topic annotation to the MPQA corpus. Unlike the previous resources, annotation in this dissertation is not dependent upon the opinion expressions. Instead, the topic and holder of an opinion are directly annotated without considering whether or not the sentence contains subjective expressions. Specifically, the holder of an opinion (H), the opinion topic with positive polarity (PT), and the opinion topic with negative polarity (NT) are annotated. The motivation for this approach is that there are sentences that express opinion without making use of explicit opinion expressions. Thompson and Hunston (2000) suggest how people recognize evaluation both conceptually and linguistically. Conceptually, comparative, subjective and value-laden characteristics make the information evaluative. This conceptual characteristic of evaluation is realized linguistically, with lexis, grammar and text. "Stance markers" described in (Biber and Finegan, 1989) include both lexical and grammatical expressions. Opinion lexis has been acknowledged by previous researchers as a key clue to identify opinions. However, this is not the only clue for determining opinion factors within the sentence. In addition to opinion lexis, grammar, pragmatics, context and even culture play a key role in identifying opinion topics from the text. This is especially frequent in the editorial domain: many more patterns and indirect ways of expressing opinion are present.

3.2.1 Sentence polarity annotation

As a first step of annotation, annotators were directed to judge if the sentence contains opinion or not. Then, for the sentences containing opinion,

sentence polarity annotation was performed. Although sentence polarity is not one of main types of opinion factor highlighted in this dissertation, it is closely related to the opinion topic annotation. In this dissertation, only the most prominent opinion within a sentence was annotated, unlike the MPQA corpus which identifies all the subjective expressions. The motivation of this approach is the assumption that the prominence of opinion within a sentence could be determined from the linguistic structure and context when there is more than one opinion present. In other words, people might annotate the prominent topic of an opinion more easily and consistently considering the context. For example, if the sentence (4) is presented alone, there should be a possibility to annotate *the resolution* as a positive topic (in the writer's opinion) in addition to the annotation of *We*—coreferent of *South Korea* in the previous sentence— as a negative topic. Another possible option is to annotate *the resolution* as a negative topic of the holder *We*. In the context that considers the previous sentence (3), however, it becomes pretty clear that the polarity is a negative, and *We* is a negative topic for the writer.

(3) ***South Korea*** shunned its responsibility as a liberal democracy and as a viable member of international society.

(4) ***We*** should have voted for *the resolution*.

The types of sentence polarity annotation include positive (P), negative (N) and positive & negative (PN). When positive and negative opinions are equally prominent in one sentence, the sentence polarity is annotated as PN. The sentence (5) is considered to carry both positive and negative sentiment as the positive topic *It* and the negative topic *the regulations over mortgage loans* within the sentence are equally prominent. Both sentiments are also captured in the sentence (6) which is the Korean counterpart of the sen-

Table 3.3: Sentence polarity annotation

	English	Korean
Sentences	2553	2824
Negative (N)	919 (36%)	1101 (39%)
Positive (P)	471 (18%)	507 (18%)
Postive and Negative (PN)	52 (6%)	54 (2%)

tence (5): positive topic 주택공급 규제의 해제 (housing-supply regulation release) and negative topic 주택담보대출 규제 (morgage-loans regulation). Table 3.3 shows the statistics of the annotated sentence polarity.

(5) The OECD advised, "***It*** is far more important to deregulate the housing supply, including reconstruction in Gangnam-gu, Seoul. And the government should also ease ***the regulations over mortgage loans***."

(6) OECD는 서울 강남의 재건축 규제를 비롯한
OECD-nun seul kangnam-uy caykenchwuk kyucey-lul pilos-han
OECD-TOP Seoul Gangnamgu-GEN reconstruction regulate-ACC including

주택공급 규제의 해제가 훨씬 중요하며
cwuthak-kongkup kyucey-uy haycey-ka hwelssin cwungyohamye
housing-supply regulate-GEN release-NOM far-more important

주택담보대출 규제도 완화해야 한다고 조언한다.
cwutayk-tampotaychwul kyucey-to wanhwa-hayya han-ta-ko coenhan-ta.
morgage-loans regulation-too ease-should do-MOD-COMP advise-COMP

3.2.2 Topic with polarity annotation

Stoyanov and Cardie (2008) describes the difficulty of opinion topic annotation in the fine-grained subjectivity analysis if no context beyond sentence level is provided. They provide a different notion of opinion "topic" com-

pared to the "target" of the opinion annotated in the MPQA corpus. In their work, the *topic* of an opinion is defined as "the real-world object, event or abstract entity that is the subject of the opinion as intended by the opinion holder", and the *topic span* is described as "the closest, minimal span of text that mentions the topic. Here, in turn, they use the term *opinion* to cover all types of private states. On the other hand, *target span* is used to denote "the span of text that covers the syntactic surface form comprising the contents of the opinion." To annotate and identify opinion topics, they use topic-coreference resolution and try to find clusters of coreferent opinions.

In this dissertation, however, the notion of the opinion is limited to the "opinion with polarity which represents appraisal". More specifically, only the most ***prominent*** opinion within a sentence is focused when more than one opinion is present. As described in chapter 1, opinion topics in this study include types of attitude in the appraisal system (judgement, affect and appreciation). In the annotation process, however, types of attitude were not distinguished and the opinion expression used to express the attitude was not annotated. Annotators were directed to determine the most ***prominent*** opinion topic within a sentence considering the context. The smallest noun phrase as well as the head noun of the opinion topic were annotated. There should be linguistic clues to determine the opinion topics such as opinion lexis and grammar. These clues as well, however, are not distinguished in the annotation. What was focused on in this study is the opinion topic and the polarity it is carrying, without specifying the detailed types of attitude.

The statistics of the annotated topics are shown in Table 3.4. The ratio among NT, PT and PNT is almost similar to the ratio among N,P and NP in Table 3.3 although the numbers are not exactly the same. Negative polarity is much more frequent than positive polarity in both sentence polarity and

Table 3.4: Topics with polarity annotation

	English	Korean
Negative Topic (NT)	846	813
Positive Topic (PT)	455	460
PT and NT (PNT)	49	48

opinion topics. Le (2009) reports the same tendency of negative prominence in *Le Monde's* editorials.

3.2.3 Holder annotation

The holder of an opinion is annotated when the holder is present within the sentence containing an opinion topic with polarity. [Some commentators in the United States] in the sentence (7) holds a negative opinion on the topic [this alliance].

(7) [Some commentators in the United States]H are arguing that [this alliance]NT should be re-assessed when the new administrations of both countries take office.

(8) [President Roh Moo-hyun]H said that [the Northern Limit Line] is not a border.

On the other hand, in the sentence (8), the holder [President Roh Moo-hyun] talks about the topic [the Northern Limit Line], but no obvious polarity is shown. The *source* of an objective speech event like in this sentence was annotated as an opinion holder as well.

As illustrated in Table 3.5, in many cases, the opinion holder is not present within the sentence where a topic of the opinion is identified (in parentheses).

Table 3.5: Holder annotation

	English	Korean
Holder of NT	109 (846)	72 (813)
Holder of PT	81 (455)	55 (460)
Holder of PNT	18 (49)	13 (48)
Holder without polarity	92	67

An inherent holder, in this case, is assumed to be the writer who represents the news agency.

3.3 Factors determining opinion topic and polarity

As previously described, linguistic factors other than opinion lexis play a role in expressing opinions. In this section, patterns of expressing opinions by different level of linguistic structure are demonstrated from the annotated corpus.

3.3.1 Lexis

As mentioned, a lexicon containing subjectivity is the key factor in determining the opinion topic and polarity in most cases across languages.

L1. **opinion noun or adjective: [Subject] + be verb + *adjective/noun***

(9) [This]NT is an *insult* to the people who support him.

L2. **opinion verb: [Subject] + *verb***

(10) But [the government]H said, "[The report]NT *failed* to accurately reflect the real situation of the Korean economy," adding that it would have the OECD revise the draft.

L3. **(Subject: holder) +*verb* + [Object]**

(11) As for the peace regime, [the U.S.]H prefers [a roadmap consisting of "completion of denuclearization first, a peace treaty second, and then, finally, U.S.-North Korea diplomatic ties]PT."

Opinion topics induced from the opinion lexis are mostly a subject or object within the sentence depending on the context. In the sentence (9), [this] is directly pointed out as the writer's negative topic with the use of the following noun *insult*. The subject within the speech-event in the sentence (10), on the other hand, is marked as a negative topic of the holder [the government] with the verb *failed* as the most obvious clue. In the sentence (11), the positive topic [a roadmap ~] of the holder [the U.S.] is identified with the verb *prefers* in between.

L4. **preposition+[object]**

(12) But critical thinking and insight *are fostered by* [reading]PT.

(13) [He]H said that from [books]PT, he not only *gets* information but also *enhances* concentration.

Opinion topics within the prepositional phrase could also be expressed with opinion lexis; this pattern occurs relatively infrequently in our corpus, as we focus only on the most prominent opinion within a sentence. The positive topic [reading] in the sentence (12) is expressed with the preposition *by* followed by the passive voice of the verb *foster*. Likely, in the sentence (13), the positive topic [books] of the holder [He] within the prepositional phrase is cued by the words *gets* and *enhances*.

While individual words contain subjective meaning in the above examples, there are cases where idiomatic expressions are used to express opinions. Mi-

halcea et al. (2007) suggest that a significant portion of the subjective lexicon in English is composed of multi-word expressions, which cannot be effectively translated to build bilingual opinion resources with the use of existing bilingual dictionaries. In the sentence (14) and sentence (15), negative topics are captured with idiomatic expressions.

(14) President Roh's biggest regret must be that [he]NT *had issues with* dignity as a president.

(15) [We]H have *heard enough of* [slogans]NT.

3.3.2 Grammar

As a grammatical factor, *were*-subjunctive mood (or past subjunctive) which is hypothetical and unreal in meaning (Quirk et al., 1985) is observed to be used mostly to express negative polarity. In the sentence (16), the writer's negative opinion on [they] is expressed with the use of the adverbial representing subjunctive mood *like*.

G1. **[topic] + be verb + *like* — as if [topic]**

(16) With socialist President Roh's strong support behind them, [they]NT are acting *like* they are above the law.

G2. **[topic] should/ought to + perfective/progressive**

(17) If he was displeased with Lee Myung-bak and the party, [he]NT *should have run* in the party primary so his political positions could be evaluated.

(18) [Law enforcement agencies]NT should have taken action sooner.

Opinion topics are also expressed with the use of modals like *should* or *ought to* when they are used in the perfective or progressive tense to induce the meaning of obligation. This combination of modals and prerfective/progressive tense tends to imply "non fulfillment of obligation (Quirk et al., 1985)". With this usage, the polarity of the opinion topic [he] in the sentence (17) is negative, as [he] did not fulfill the obligation from the writer's point of view. Likely, the writer shows negative opinion on the topic [Law enforcement agencies] in the sentence (18) as they didn't take action sooner although they should have.

G3. **should/must/ought to + *verb* + [Object]**

(19) In short, the message from [the OECD]H is that, "The government *should attempt* [a U-turn in its economic policy according to market principles]PT."

When modals containing the meaning of obligation are used in present tense, an object in the sentence could be identified as an opinion topic depending on the verb following the modal. [a U-turn in its economic policy ∼] in the sentence (19) is expressed as a positive topic of the holder [the OECD] with the use of the combination of modal *should* and verb attempt.

3.3.3 Pragmatics

In addition to the opinion lexis and grammar within sentences, pragmatic meanings determine the opinion and polarity in several ways. If a sentence contains a subordinate clause led by a conjunction, the polarity of an opinion within the subordinate clause should be determined beyond the clause level. Opinion is expressed with the combination of the meaning of each clause (main and subordinate) as well as the type of conjunction.

P1. **Conditional clause**

- **If + [topic] +polarity, +polarity ⇒ -polarity**
- **unless + [topic] +polarity, +polarity ⇒ +polarity**

(20) If [this principle]PT *recedes*, there will be *no real peace* for the South.

(21) There is *no hope* of reelection for the pan-ruling party circle and the Roh administration *unless* they *disavow* [this mindset]NT.

In the sentence (20), the negative polarity induced by the verb *recedes* in the conditional clause headed by *if* is shifted with the combination of the negative polarity contained in the main clause. As a result, [this principle] is expressed as a positive topic which "should not recede". On the other hand, in the sentence (21), the negative polarity of the topic [this mindset] in the conditional clause headed by *unless* remains with the negative polarity in the main clause.

P2. **Conditional prepositional phrase: without + [topic], +polarity ⇒ -polarity**

(22) *Without* [security]PT, economic cooperation is *in vain*.

Similarly, the polarity of the opinion topic in the prepositional phrase headed by *without* is determined by the content of the main clause: polarity shifts from the polarity of the main clause.

P3. **polarity in Rhetorical question ⇒ -polarity**

(23) The economic association also asked, "Can Korea become the hub of Northeast Asia with [its regulation on the Seoul Metropolitan area]NT in place?"

Another pragmatic factor for inducing opinion is the rhetorical question. The rhetorical question is syntactically interrogative, but it does not expect an answer; instead, it is a statement with a strong assertion (Quirk et al., 1985). In the sentence (23), the negative polarity of the opinion topic [its regulation on the Seoul Metropolitan area] is induced from the rhetorical question within the sentence.

3.3.4 Context: beyond sentences

Opinion topics in a sentence sometimes should be identified beyond the sentence level. That is, it is not possible to detect the polarity of the opinion topic without considering the broader context. The negative polarity of the opinion topic [His comment in Washington] in the sentence (26) is derived from the two previous sentences (24) (25), as we should recognize what "the same attitude" is.

(24) [Abe]NT is someone who has denied the comfort women issue is an issue at all since before his inauguration as prime minister last September.

(25) Recently [he]NT went further, saying he would see to it that there is an inquiry that questions the matter.

(26) [His comments in Washington]NT are part of *the same attitude*.

3.4 Inter-annotator agreement

To validate the annotation process and scheme, additional annotation was performed by another annotator. The second annotator is female in her mid-thirties. She has both a computer science and linguistics background from her

Table 3.6: Inter-annotator agreement (Sentence polarity): Kappa statistics

κ: 0.82	P	N	PN	None	Total
P	872	5	4	115	996
N	5	1860	1	164	2030
PN	10	7	87	4	108
None	111	186	0	1946	2243
Total	998	2058	92	2229	5377

undergraduate and graduate study respectively. The first and second annotators both speak Korean as native language and English with high proficiency. The annotator was trained in the annotation scheme using five extra articles for each language pre-annotated by the author. In the opinion factor annotation, the sets of opinion factors annotated by each annotator should be different. That is, a specific word phrase could be annotated as opinion topic by one annotator but not chosen by the other annotator. Therefore, in addition to the traditional Cohen's Kappa κ which is more appropriate tasks involving the same set of objects, the *agr* metric proposed in Wiebe et al. (2005) was adopted to calculate the inter-annotator agreement. The *agr* metric measures the recall of the annotated set A by annotator a, with respect to the set B by the other annotator b based on the following equation:

$$recall(a||b) = \frac{|\text{A matching B}|}{|\text{A}|} \tag{3.1}$$

As for the sentence polarity, κ is 0.82 which shows almost perfect agreement as illustrated in Table 3.6. Mean ratios of *recall (a||b)* and *recall (b||a)* from the *agr* metric in Table 3.7 also show very high agreement which is above 0.87 for all types of sentence polarity.

Inter-annotator agreement for opinion factors was calculated based on head-nouns: see if the head noun of the opinion factor from each annota-

Table 3.7: Inter-annotator agreement (Sentence polarity): *Agr* ratio

agr	recall (a‖b)	recall (b‖a)	mean
N	0.92	0.90	0.91
P	0.88	0.87	0.88
PN	0.80	0.95	0.88

Table 3.8: Inter-annotator agreement (Opinion Factors): Kappa statistics

κ: 0.89	H	PT	NT	None	Total
H	587	0	1	40	628
PT	2	1025	0	165	1192
NT	12	2	1737	227	1978
None	72	133	120	80690	81015
Total	673	1160	1858	81122	84813

tor matches. Agreement for all types of opinion factors shows a mean of more than 0.87 *agr* with a 0.89 κ value.

In addition to the inter-annotator agreement for sentence polarity and each opinion factor, the agreement of the annotation as a whole within a sentence was calculated. Sentence pairs which were not annotated in exactly the same way between annotators were treated as disagreed sentences, even though part of the annotation is in agreement. As shown in Table 3.10, 60% and 52% of all sentence pairs in English and Korean respectively contain at least one annotation among opinion factors (Tagged). Among the "Tagged" sentences, 16% and 18% of sentences in the English and Korean pairs were disagreed

Table 3.9: Inter-annotator agreement (Opinion Factors): *Agr* ratio

agr	recall (a‖b)	recall (b‖a)	mean
H	0.93	0.87	0.90
NT	0.88	0.93	0.91
PT	0.86	0.88	0.87

Table 3.10: Inter-annotator agreement (All factors)

English			Korean		
All	Tagged	Dis	All	Tagged	Dis
2553	1549	254	2824	1478	272

·All: No. of all sentences in the corpus
·Tagged: No. of sentences where any of the opinion factor is annotated
·Dis: No. of sentences not matched between annotators

upon considering all the opinion factors. The disagreed upon sentence pairs account for about 10% of all sentence pairs in both English and Korean. In the sentiment analysis experiment presented later in this dissertation, the disagreed upon sentences were omitted during learning in order to make the annotated corpus as confident as possible.

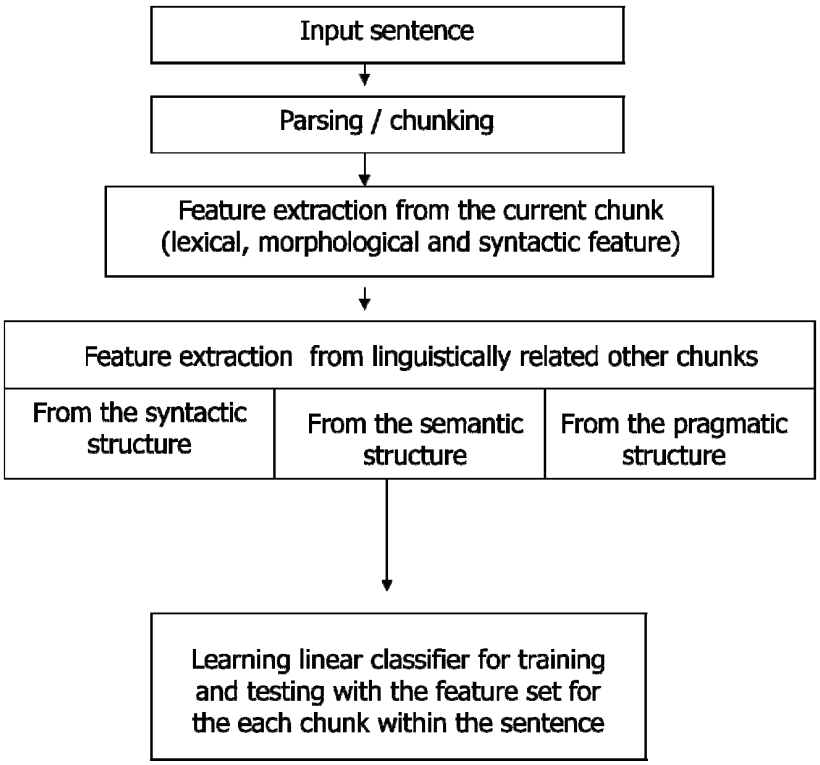
CHAPTER 4

IDENTIFICATION OF OPINION TOPIC, HOLDER AND POLARITY FROM MULTILINGUAL CORPORA

To automatically extract the opinion holder and topic along with the opinion's polarity, a multilingual sentiment analysis system was designed and implemented based on a supervised machine learning algorithm with an annotated corpus. The implemented sentiment analysis system is a bilingual system designed using English and Korean, which could be expanded into a multilingual by adding additional languages with the same procedure. The schematic representation of the implemented system is illustrated in Figure 4.1.

As illustrated, when an input sentence is fed to the system, the first procedure is parsing and chunking. Instead of individual words, chunked units act as the basic unit for the feature extraction and machine learning. Opinion feature dictionaries constructed in various linguistic aspects are used to extract features for the current chunk and other linguistically related chunks within a sentence. Feature extraction from the current chunk is performed first, and then feature extraction from other linguistically related chunks follows. With all the features set for each chunk, the linear classifier is learned to identify opinion factors from new data.

Figure 4.1: Schematic representation of the multilingual sentiment analysis system



4.1 Preprocessing of input sentences: chunking

The sentiment analysis system implemented in this dissertation aims to utilize linguistic structure across languages to identify opinion factors. Rather than an individual word, a word group composing a syntactic phrase is more effective as a basic unit for the feature extraction and the machine learning in this regard. Therefore, as a part of the preprocessing of the input for the system, input sentences were chunked after being fully parsed. This parsing and chunking was performed instead of performing shallow parsing, as the internal structure and the role in the sentence of a chunk are to be used in the system. Chunking fitted to the objective of this system was performed

Table 4.1: Chunking units: detailed noun phrase types

Korean	NP_SBJ	NP subject with nominative case marker
	NP_OBJ	NP object with accusative case marker
	NP_CMP	NP complements
	NP_MOD	modifying NP
	NP_AJT	adjectival NP
	NP_CNJ	conjunctive NP
	NP_PRN	pronoun
English	NP_SBJ	NP under S node
	NP_OBJ	NP under VP node
	NP_POBJ	NP under PP node

independently in English and Korean with different criteria depending on the characteristics of each language: i.e. head-initial English and head-final Korean. English sentences were parsed using the Charniak parser (Charniak, 1999), and Korean sentences were parsed using the Probabilistic chart parser implemented at Postech in Korea (Eun et al., 2006). The phrase types which were chunked were noun phrases (NP), verb phrases (VP), adverbial phrases (ADVP) and adjectival phrases (ADJP). While the parsed outputs of the English data only identify the general category of phrases such as NP and VP, the parsed outputs of the Korean data contain information about the grammatical functions of the phrase, which are derived more clearly from the grammatical markers within a phrase. The detailed information from the parsing engine for Korean was retained in chunking. In addition, English NPs are categorized into three groups (NP_SBJ, NP_OBJ, NP_POBJ) considering the governing category (S,VP and PP respectively), in order to make the chunked phrases as parallel as possible between languages. Detailed noun phrase types among chunked outputs are shown in Table 4.1. Other than noun phrases, phrasal types from the parsed outputs were used directly without modification both in English and Korean.

4.2 Feature dictionaries

As a major preparation for the sentiment analysis experiment, opinion feature dictionaries were designed and constructed. In addition to the opinion lexis, grammar and text act as clues for the representation of opinions within a sentence (Thompson and Hunston, 2000). Therefore, lexical, syntactic and pragmatic feature dictionaries were constructed to capture the possible linguistic features for opinion factors. As the aim of this dissertation is multilingual sentiment analysis exploring each language at the same time in order to reinforce the performance, the feature dictionaries were applicable for both English and Korean except for the morphological dictionary which is mostly applicable to Korean only. Feature dictionaries were constructed in advance from existing dictionaries or linguistic knowledge, and utilized in the machine learning process.

4.2.1 Lexical features

Opinion lexis is known to be the most important clue for expressing opinions within a sentence, and has been pursued by most previous researchers. An English subjective lexicon extracted in Riloff and Wiebe (2003) was publicly shared and used by previous researchers as lexical clues for subjectivity. To identify polarity of opinions from texts, words are categorized as positive, negative or neutral mostly utilizing manually collected seed words. The lexical feature dictionary in this dissertation, however, is distinct from the previous resources in that prior polarity is not pre-labeled. Although the prior polarity of the word itself is very important information for identifying opinion factors, actual subjectivity or polarity should be determined within context (Choi and Cardie, 2008; Ding and Liu, 2007; Ikeda et al., 2008; Kennedy and

Inkpen, 2006; Wilson et al., 2005, 2009). Therefore, the system is designed to determine the clues for opinion factors automatically in the learning process combining all possible features in addition to the lexical features. Instead of collecting opinion words based on their prior polarities, words having the same meaning across English and Korean were clustered into one feature set to make the machine learning process more effective. With this approach, lexical features important for identifying opinion factors could be strengthened in the learning process if they are present in both languages. Also, the data sparseness problem could be solved to some extent with the clustering strategy.

As a first step of clustering, words, nouns, adjectives, verbs and adverbs sets from English-Korean bilingual dictionaries[1] were collected. Unlike most previous studies utilizing adjectives and verbs to detect opinion factors (Bethard et al., 2004; Kim and Hovy, 2006), nouns and adverbs were also collected as candidate lexical features for opinion factors: for the machine learning process, as many lexical features as possible were collected as candidates. To effectively reduce the computational cost, however, nouns were filtered using the English subjective lexicon by Riloff and Wiebe (2003) containing 2172 nouns, while all entries of adjectives, verbs and adverbs from the bilingual dictionaries were used. The procedure for constructing the lexical feature dictionary is illustrated in Figure 4.2.

As illustrated, starting word sets are composed of English key words and Korean word sets such as ENG_i;kor_{i1};kor_{i2};...;kor_{ij}. Then, English key words were expanded using the synsets from WordNet. Korean words were clustered along with the English Key words. Finally, overlapping feature sets were excluded. As a result, clustered features of 1857 nouns, 6321 adjectives, 4602

[1]Dong-A's Prime English-Korean Dictionary (4th edition)

Figure 4.2: Constructing clustered lexical feature dictionary

---

▪ **Sources**

- **English-Korean bilingual dictionary (Nouns, adjectives, verbs, adverbs)**
- **English subjective lexicon by Riloff and Wiebe (2003) : 2172 nouns**

1. starting with English Key word- Korean word sets from bilingual dictionary:

   ENG_0;kor_{01};kor_{02};…kor_{0j} ENG_1;kor_{11},kor_{12};… kor_{1j} …… ENG_i;kor_{i1};kor_{i2};… kor_{ij}

2. if ENG_0 and ENG_1 are WordNet Synset Sy1(ENG_0 ENG_1 ENG_2 ENG_3):
   feature set is clustered by expanding English word set

   ENG_0; ENG_1;ENG_2,ENG_3:kor_{01};…kor_{0j};kor_{11};…kor_{1j}

3. Making the dictionary cleaner by excluding overlapped feature sets

---

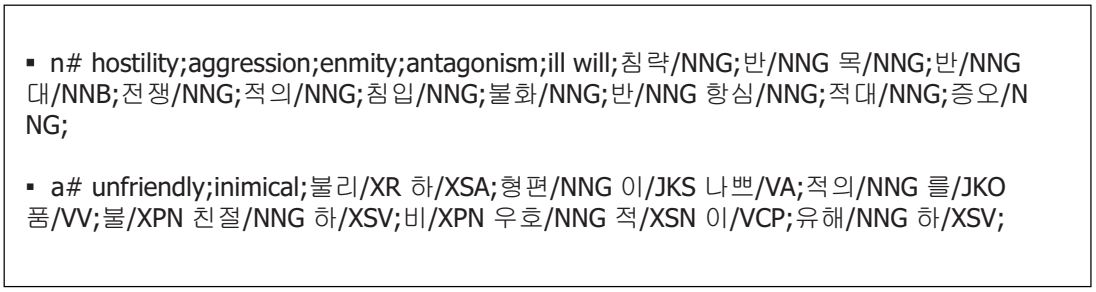
verbs and 1580 adverbs are in the dictionary. After clustering, Korean words were morphologically analyzed and the functional segments were filtered out. Examples of constructed feature sets are shown in Figure 4.3.

4.2.2 Syntactic features

To reflect the status of a candidate chunk for opinion factors within a sentence, a syntactic feature dictionary was designed referring to Gildea and Jurafsky (2002)'s work using various syntactic features such as phrase type, governing category and voice for automatic semantic role labeling.

As a first type of syntactic feature, the phrasal type of a chunked unit was used. The targets of the sentiment analysis system are opinion factors,

Figure 4.3: Examples of lexical feature



holder and topic. The opinion topics pursued here is defined as "real-world object, event or abstract entity", which are expressed as noun phrases in most cases. Therefore, in the case of noun phrases, more detailed types that mark grammatical functions within a sentence should be more effective. Noun phrases in the English data were categorized into three types considering the governing category as mentioned in section 4.1. All phrasal types containing the grammatical functions of chunked units are used as a syntactic feature in Korean. Next, the higher path in the parse tree — the path between the root and the word — is considered as a syntactic feature. This reflects the syntactic relation of a constituent to the rest of the sentence, whether or not the higher path contains node sequences representing the syntactic status

Table 4.2: Syntactic features: higher path

S/NP, S/VP
S/S/NP, S/S/VP
VP/S/NP, VP/S/VP
NP/S/NP, NP/S/VP
SBAR/S/NP, SBAR/S/VP
SBAR/S/S/NP, SBAR/S/S/VP
S/VP/VP
S/VP/VP/PP
S/VP/PP

within a sentence including embedded sentence (S), verb phrase (VP) and prepositional phrase (PP). The complete list of the higher path features is illustrated in Table 4.2.

4.2.3 Contextual features

Polanyi and Zaenen (2004) suggest there are contextual valence shifters based on the sentence or discourse. The first type of contextual valence shifters they describe are *negatives* and *intensifiers*. *Negatives* such as "not" or "never" shift the valence between positive and negative. *Intensifiers*, on the other hand, just influence the strength of the valence instead of flipping the valence, so they are not strongly relevant to the current system. Therefore, the negatives presented in Table 4.3 were used as sentence-based contextual features. Another important type of sentence-based valance shifter is the class of *modals*, which plays an important role in English. Modals representing obligation, ability and intention were chosen to be used as contextual features, and corresponding Korean fragments were added to the feature lists if applicable (Table 4.3). Beyond sentences, *connectors* were collected as discourse-based valence shifters, following Polanyi and Zaenen (2004). In addition, the ques-

Table 4.3: Contextual features: sentence-based

Negation	never, none, nobody, nowhere, nothing, neither 못하(mosha)/VX, 않(anh)/VX, 안(an)/MAG, 없(eps)/VA 아니(ani)/VCN, 비(pi)/XPN, 잃(ilh)/VV
Modals	should, have to, much, ought to, 이어야할(ieyahal), 해야할(hayyahal) will, would, 하기로(hakilo) can, be able to

Table 4.4: Contextual features: discourse-based

as long as, 는한
although,though,however, 불구하고(pulkuhako)
after, if, 면(myen), ㄴ다면(n-tamyen)
no matter how
why
unless
when
whether
whenever
as if, 인양(injang)
question mark, 는가(nunka), ㄴ가(n-ka), 가(ka)

tion mark is treated as a valence shifter affecting the whole sentence. The discourse-based contextual features should be dealt with separately when extracting features, as they influence the whole or part of the sentence beyond the chunked unit. The list of discourse-based features are shown in Table 4.4.

4.2.4 Morphological features

As Korean is an agglutinative language, morphological features are very critical for identifying the role of a word in a sentence. Therefore, morphological analysis is performed as well as syntactic parsing, then grammatical markers as well as suffixes conveying special meanings are used as morphological features for Korean. The morphological features used are presented in Table 4.5.

The subjective case marker 이(i), 가(ka) and the objective case marker 을(ul), 를(lul) could represent semantic roles in active voice sentences. One of the most particular aspects of Korean markers which is distinct from English is the use of the topic markers 은(un) and 는(nun). A topic marker, as the name represents, encodes the topic of a given sentence. Therefore, it is highly likely to be related to the opinion factors pursued in the present study if the sentence contains an opinion. On the grammatical account, there are several interpretations of the topic marker in a sentence (Park et al., 1994). One way to interpret the topic marker is to consider the topic marker as ambiguously being a case marker such as a subject and object marker. The other interpretation is that the topic marker is optionally adjoined onto the beginning of the sentence irrespective of the argument structure of the verb in a sentence. In this case, the argument in a sentence is considered to be empty if there is no other candidate noun phrase. The Penn Korean Treebank (Han et al., 2002) whose data are used by the Propbank database adopts the second interpretation of empty argument. On the other hand, the probabilistic chart parser (Eun et al., 2006) used for parsing in the current study assigns a case to the noun phrase containing the topic marker. Grammatical markers and suffixes with special meanings are also listed as morphological features, as they act in the same way as the topic marker in that they can be replaced with any case markers depending on the usage. The presence of a topic marker in Korean is expected to play an essential role in identifying opinion factors. Also, the other features within the opinion factors identified with the help of the topic marker are expected to be a benefit to identifying opinion factors in English data with the same sentence structure.

Table 4.5: Morphological features

이(i)/JKS, 이가(ika)/NNG, 가(ka)/JKS
을(ul)/JKO, 를(lul)/JKO
은(un)/JX, 는(nun)/JX (topic markers)
나마(nama)/JX, 나(na)/JX
으로(ulo)/JKB, 으로써(ulosse)/JKB
라도(lato)/JX, 도(to)/JX, 든지(tunci)/JX
야말로(yamallo)/JX, 만(man)/JX, ㄹ 랑(l-lang)/JX, 나마(nama)/JX, 도(to)/JX
마저(mace)/JX, 부텀(puthem)/JX, 조차(cocha)/JX, 부터(puthe)/JX 까지(kkaci)/JX, 오히려(ohilye)/MAJ, even
뿐(ppun)/JX, 밖에(pakkey)/JX, 만(man)/JX, merely, only, just nothing less than, nothing less of

JKS: subjective case, JKO: objective case, JKB: adverbial
JX: auxiliary, MAJ: adverb

4.3 Feature extraction

In the present system, the basic units for feature extraction and machine learning are chunked units described in Section 4.1. When input sentences are fed into the system, the sentences are chunked and feature extraction is performed for each chunked unit. As a first step, features within the chunked unit are extracted. Then, features from linguistically related other chunks are extracted.

(1) We *are sick of* [the *deception* and *audacity* of the current government]NT which named the suppression of the freedom of speech as the so-called "Advanced Media Support System".

(2) [The ruling]PT is a *welcome move* and signals the *greater importance* of education rights than teachers rights.

The negative topic in the sentence (1) [the *deception* and *audacity* of the current government] contains lexical clues for negativity within the phrase as well as outside the phrase. On the other hand, lexical clues for the positive topic in the sentence (2) [the ruling] are only present outside the phrase. To identify the topic chunks in both sentences effectively, it is optimal to extract relevant features not only from the current chunk but also from outside the chunk. As syntactic and morphological features are only relevant for the current chunk, the features extracted from outside the chunk are lexical features and sentence-based contextual features. Contextual features based on discourse should be extracted from outside the chunk as well, but the extraction of those features goes through a different procedure in that discourse-based features affect a whole sentence or part of the sentence instead of a specific chunk. To extract lexical features from other linguistically related chunks, predicate-argument relationship is explored and utilized as a semantic structure. Syntactic structure is also explored with the use of the parse tree of a sentence.

4.3.1 From the predicate-argument relationship

Semantic-role labeling has been made use of as a way of utilizing semantic structure to extract opinion factors in several previous studies. Bethard et al. (2004) use both the PropBank (Palmer et al., 2005) and the FrameNet database (Barker and Sato, 2003) to identify propositional opinions and their holders. Choi et al. (2006) use the PropBank argument role labeling to investigate the joint extraction of opinions and sources. To identify topics and hold-

ers of opinion expressions, Kim and Hovy (2006) label the semantic roles of each opinion word using the FrameNet database. In the Propbank database, a layer of predicate-argument relationship is added to the syntactic tree of the Penn Treebank corpus. The FrameNet database describes semantic frames consisting of lexical units and frame elements. For example, the lexical units *hope, wish, interested* and *desire* are members of "Desiring frame" with frame elements such as *event, experiencer* and *location_of_event*. While the PropBank database only contains the semantic role relationship anchored to a verb, lexical units in the FrameNet include adjectives and nouns as well.

Unlike the previous studies that explore the semantic structure from the pre-identified opinion word to the possible opinion factors, the present system starts from the possible opinion factors to collect any clues within a sentence about opinion factors. Although the FrameNet has the advantage over the PropBank in that it covers adjectives and nouns additionally, the PropBank database was used to extract predicate-argument relationships in this dissertation. The first reason is the bilingual applicability of the PropBank database: A Korean PropBank as well as an English PropBank exists. Second, the PropBank database adds the predicate-argument relations to the data of the Penn Treebank corpus, which means the annotated predicates and arguments are connected with the parse tree information in the Penn Treebank corpus. As a step of extracting features for opinion factors, possible predicate-argument relationships within a sentence were extracted from the PropBank database instead of performing semantic-role labeling as a separate step. The English PropBank used in this study contains 112,917 total propositions, which covers the entire Wall Street Journal section of the Penn Treebank corpus excluding auxiliaries and the verb "be". The total number of framesets in this data is 4,659. In the Korean PropBank, 9,588 and 23,707

predicates are annotated from the Virginia corpus and Newswire corpus of Penn Korean Treebank data respectively. 2,800 framesets are contained in the Korean data.

Semantic roles are generally represented as three levels of generalities. The first level is verb-specific semantic roles such as *runner, killer* and *hearer*. Thematic-relations such as *agent, instrument* and *experiencer* are the next level of generality. The most general level is representing semantic roles with only two types: agent-like and patient-like. They are also called "proto-roles" or "macroroles" (Dowty, 1991; Robert D. Van Valin, 2005). In the PropBank, arguments are labeled with numbered arguments of which detailed semantic roles are verb-specific. The description of the detailed semantic roles is presented in the frame files as shown in Figure 4.4. The numbered arguments span from Arg0 to Arg6. Although the details of the arguments are verb specific, the criteria for labeling Arg0 and Arg1 are quite consistent: the Arg0 label is usually assigned to the argument of agent, causer or experiencer, while the Arg1 is assigned to the patient argument. In addition to the numbered arguments, ArgA (causative agents) and ArgM (adjuncts of various sorts) are annotated if present. In the current system, possible Arg0, Arg1, Arg2 and ArgA of a specific predicate are used as features. In addition to the arguments of a specific predicate, the inter-argument relationship is also used as feature.

(1) [The conservatives]PT chose their candidate by going through fair procedures in the primary.

(2) [The government]H banned [the demonstration]NT, fearing the traffic night,are and public inconvenience.

The predicates in the English Propbank include detailed information about

Figure 4.4: Example of Frame files in the PropBank: English

Frame File for the verb 'expect':

Roles:
Arg0: expecter
Arg1: thing expected

Example: Transitive, active:
Portfolio managers expect further declines in interest rates.

Arg0:	Portfolio managers
REL:	expect
Arg1:	further declines in interest rates

Table 4.6: Annotated information with predicates in the English PropBank

form	i=infinitive g=gerund p=participle v=finite
tense	f=future p=past n=present
aspect	p=perfect o=progressive b=both perfect and progressive
person	3=3rd person
voice	a=active p=passive

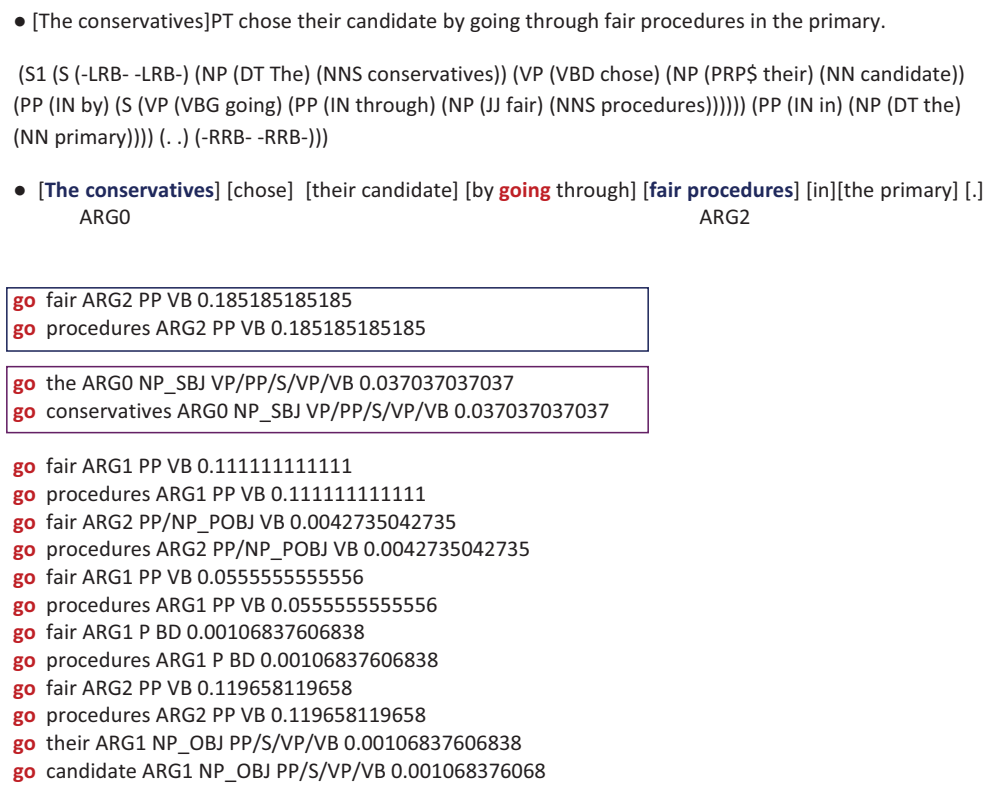
form, tense, aspect, person and voice as described in Table 4.6. Among this information, the form and voice information was selected to be used in the present system. Possible arguments anchored to a specific predicate were extracted using statistics based on the syntactic paths of the predicate and the candidate arguments. The procedure for labeling arguments anchored to a predicate in the sentence (1) is illustrated in Figure 4.5. In the example, the input sentence contains a positive topic [the conservatives] to be identified. As a first step, the input sentence is parsed and chunked. Features from the

current chunk [the conservatives] were extracted first such as the phrasal type of NP_SBJ. However, the most obvious clue to identify this chunk as a positive topic is the lexical feature *fair* in the chunk placed later in the sentence with a long distance. To identify the relation between the two chunks (*target chunk* and *chunk with lexical clue*), predicate-argument relations anchored to the predicates in the sentence (*chose, going*) are extracted. For each predicate, the possible combination of predicate, argument and their conditional probability was extracted based on the syntactic path under the node shared between them. Based on the conditional probability, the most probable arguments of the predicate *go* were extracted: [the conservative] as Arg0 and [fair procedure] as Arg1. Therefore, the lexical feature *fair* is assigned to the current chunk [the conservatives] as the inter-argument feature of the predicate *go*. In the sentence (2), on the other hand, the holder [the government] and the negative topic [the demonstration] are linked through the predicate *ban*: Arg0 and Arg1 of *ban* respectively.

4.3.2 From the syntactic structure

Although the predicate-argument relation is effective for identifying clues for opinion factors in most cases, there still exists the limitation of coverage. The first limitation is from the coverage of the PropBank database itself, especially for Korean. Second, the opinion clues cannot be identified if they are not related to the opinion factors through the predicate. Syntactic structure is additionally explored in this regard. In the case where the previous step of pragmatic-argument relation misses the clues, the syntactic structure acts as a complement. Where the opinion clues are already extracted from the pragmatic-argument relations, on the contrary, the syntactic structure could

Figure 4.5: Argument labeling with PropBank database



strengthen the effective features for opinion factors.

Lexical features for identifying opinion factors were extracted through the syntactic structure as (1) sister node features and (2) relative clause features. As the first type, the sister node features of the current NP chunk were extracted from the adjacent chunk, which has the same higher path as the current NP chunk. Features from the corresponding relative clause for each NP were added as well, since a relative clause frequently contains opinion clues for an adjacent NP.

4.3.3 Features for beyond chunked units

As described above, various kinds of linguistic features for the current chunk were extracted from linguistically related other chunks as well as their own chunk to identify opinion factors. Discourse-based contextual features (DCfea) described in section 4.2, on the other hand, should be assigned beyond the level of chunked units. As shown in Table 4.4, most DCfeas are conjunctions in both English and Korean except for question marks. In English, conjunctions could shift the polarity of the following noun phrases while preceding noun phrases are affected in Korean. Therefore, in English, DCfea was assigned to the all chunks present after conjunctions within a sentence unless a "Comma" breaks the effect of conjunction. Likewise, DCfea were assigned to the chunks before conjunctions in Korean. When a question mark is found within a sentence, all chunks between the "Comma" (if applicable) and the question mark are assigned the DCfea.

4.4 Machine learning algorithm

The problem of identifying the holder and topic of an opinion is dealt with as a multi-class classification problem. The goal is to map the current chunk to the correct class among four classes (holder, negative topic, positive topic, and None) based on the feature sets derived from the extraction process described above. With the annotated corpus, the classification task was performed through supervised learning, which infers a function by generalizing the training data. The number of features used in this proposed system is very large, even though not all of them are active in the actual learning process. Each feature in the feature set is assumed to be independent of the others. Considering the nature of the feature set in this system, the SNoW

(Sparse Network of Winnows) learning architecture (Carlson et al., 1999) was used for the classification process. The SNoW learning architecture is a sparse network of linear units which are called *target nodes*. The linear unit is *active*(y=1) or not(y=0) based on the following equation:

$$
\begin{aligned}
&A_t = \{i_1, ..., i_m\} \\
&y = 1 \ if \sum_{i \in A_t} w_i^t > \theta_t \\
&y = 0 \ otherwise \\
&w_i^t\text{: the weight on the edge connecting the } i\text{th feature to the target node } t \\
&\theta_t\text{: its threshold}
\end{aligned}
\tag{4.1}
$$

Winnow, Perceptron, and Naive Bayes update rules could be taken by users in learning. Winnow and Perceptron update rules are similar in that they are mistake-driven: update the weight vector only when a misclassified instance is encountered. While the Perceptron update rule takes only two parameters, *threshold* and *learning rate*, the Winnow update rule takes two more parameters: *promotion* and *demotion* parameters. Winnow is called an *attribute-efficient learner*, which means it is a very efficient learning algorithm if the dataset has many features but a relatively small number of them are relevant(Witten and Frank, 2005). Both Winnow and Perceptron update rules are taken for learning in the present experiment. In addition, it is possible to assign the strength of the feature in the SNoW architecture. Using this strategy, different strengths for the features from syntactic, semantic and pragmatic structures were assigned in the experiment. Features from syntactic and pragmatic structures are assigned the same weight, while features from semantic structure are assigned more weight. This is motivated by the observation that the predicate-argument structure plays a more important

role in expressing opinions. Moreover, features shared by parallel sentences are strengthened to test the cross-lingual reinforcement in identifying opinion factors across languages.

4.5 Experiment

Opinion factors in texts were identified with the experimental set-up previously described. In addition to the bilingual editorial texts which have been annotated in this dissertation, an experiment with the MPQA corpus was also performed to verify whether the current system also works for existing resources although there are some differences in annotation scheme. The effect of the novel approaches in the present system was tested in several different experiments.

4.5.1 Baseline

As a baseline, a system which lacks the most important novel approaches in the proposed system was designed to evaluate contribution of the proposed system. First of all, the baseline system does not make use of a clustered lexical feature dictionary. A lexical feature dictionary composed of only the subjective lexicon by Riloff and Wiebe (2003) without clustering was constructed for the baseline system. Moreover, the baseline system only incorporates predicate-argument relationships to extract features from other chunks.

Simple lexical feature dictionary The lexical feature dictionary used in the present system is unique in that:
(1) It does not label the prior polarity of a word. Instead, the dictionary collects as many lexical clues for opinion factors as possible.

(2) Words with the same meaning across languages are clustered, and fed into the system as the same feature.

For the baseline system, a simple bilingual lexical feature dictionary was constructed starting with the English subjective lexicon by Riloff and Wiebe (2003). 2160 nouns, 3235 adjectives, 332 adverbs, 1322 verbs and 1136 words tagged as "anypos" are contained in the lexicon. The Korean counterpart was collected from the English-Korean bilingual dictionary used in Section 4.2: only the first word among glosses were collected. Each word in the lexicon with its part of speech is fed to the system as a separate feature irrespective of their similarity.

Limited features and feature extraction from linguistic structures To extract the proper clues that the current chunk is an opinion factor, various features from other chunks within a sentence were extracted through linguistic structures. In several previous studies of identifying opinion factors anchored to the opinion expression, semantic-role labeling was used to explore the structure of the sentence (Bethard et al., 2004; Choi et al., 2006; Kim and Hovy, 2006) although the details and strategy is quite different from the proposed system. Therefore, as a baseline system, only the predicate-argument structure was used to extract possible features from other chunks. The additional approach using syntactic structure in the proposed system is not used in the baseline system. For the current chunk, the lexical and morphological features were extracted without considering higher path features and contextual features.

4.5.2 Result and discussion

The corpus was divided into 10 groups, so that 10-fold cross-validation could be performed to evaluate the results. The classification of chunked units was evaluated: the chunked unit is considered to be target-annotated when the head noun of the opinion factor is present within a chunk. In addition to the detection of topic with polarity, topic detection irrespective of its polarity was evaluated. This additional evaluation was performed as the subjective topic itself is important information in many applications even without its polarity. Moreover, in the MPQA corpus, topics without polarity, neutral targets of subjective expressions, are annotated as well in addition to the topics with polarity. As head nouns of opinion factors are not annotated in the MPQA corpus, overlap match was performed for evaluation instead. Overlap match considers the chunked unit to be target-annotated if it overlaps with any part of the opinion factors.

Precision, recall and F-score were calculated as evaluation standards based on the following equations.

$$Precision(\%) = \frac{No.\ of\ chunks\ correctly\ calssified}{No.\ of\ classifised\ chunks\ as\ each\ class} \tag{4.2}$$

$$Recall(\%) = \frac{No.\ of\ chunks\ correctly\ classified}{No.\ of\ chunks\ annotated\ as\ each\ class} \tag{4.3}$$

$$F\text{-}score(\%) = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.4}$$

Overall result

Evaluation results for the baseline and the proposed system are shown in

Table 4.7. As illustrated, the overall result of the proposed system is much improved over the baseline results in terms of F-score for both English and Korean. One notable result is that precision of the baseline system is not really lower than the proposed system. The difference in F-score relies more on the difference in recall. As previously explained, a non-clustered lexical feature dictionary composed of only the subjective lexicon was used in the baseline system. On the other hand, the lexical feature dictionary in the proposed system clusters words with the same meaning across languages. This clustering strategy could be beneficial with sparse data such as in the current experiment, as collecting more opinion factor-annotated data costs a great deal of effort. However, this strategy could also have drawbacks in precision, as it is liable to treat different senses of words in terms of polarity as the same. The precision of positive topic (PT) in English and holder (H) in Korean actually show more accurate results in the baseline system. In English, identification of the opinion holder shows better performance than topic identification, while the opposite tendency is shown in Korean. According to Byon (2006), Korean people manipulate politeness along with indirectness and use of honorifics. Directly expressing negative feelings is considered to be impolite in Korean culture, so the holder of an opinion in Korean texts is not represented as directly as in the English texts. Therefore, the expression of opinion holders in Korean is generally more subtle and within much deeper linguistic structure. Considering that the annotated corpus is editorial texts, the difference in the style of expressing opinion in each language could be a plausible reason for this result. In the experiment of the Proposed system-E, training and testing was performed within the same language. That is, only English data are used in training to test English. On the other hand, all language data were used for training in the experiment of the Proposed

Table 4.7: Evaluation: Proposed system vs. Baseline

	English								
	Baseline			Proposed system-E			Proposed system-A		
	P	R	F	P	R	F	P	R	F
Holder	39.3	7.8	13.1	53.6	47.7	50.5	33.3	34.2	33.7
Topic	31.9	12.7	18.2	37.5	21.1	27.0	29.5	29.0	29.3
NT	28.2	11.2	16.0	33.2	14.7	20.3	27.3	26.2	26.7
PT	25.9	10.4	14.9	22.1	13.9	17.1	18.5	13.5	15.6
	Korean								
	Baseline			Proposed system-E			Proposed system-A		
	P	R	F	P	R	F	P	R	F
Holder	33.3	2.7	4.9	22.1	8.0	11.7	27.3	9.6	14.2
Topic	28.6	9.5	14.2	33.6	23.7	27.8	34.2	23.7	28.0
NT	21.8	6.7	10.2	29.4	19.8	23.0	27.2	19.3	22.6
PT	19.0	7.1	10.4	24.6	14.0	17.9	29.0	16.3	20.8

·P: precision (%), R: recall (%), F: F-score (%)
·Proposed system-E: learning with each language data only
·Proposed system-A: training with both language data in learning

system-A. The Korean annotated corpus was additionally used as training data to test English, and vice versa. This experiment was performed to verify that the proposed system is designed for multilingual data, which means that an annotated corpus of other language can be beneficial for extracting opinion factors. In the case of English, the improvements in results come from improvements in recall, while precision drops in all opinion factors. In the Korean result, however, both precision and recall are improved with the use of English data as additional training data.

Effect of each approach in the system

Other than the baseline and the current system, three more experiments were performed to evaluate the contribution of each approach in the system. The experiments *woCF* and *woCB* were performed to see how the clustering strategy affects the performance. In the experiment *woCF*, each word or

expression in the feature dictionaries is fed into the system as a separate feature, unlike the proposed system where features are clustered based on similarity. The lexical feature dictionary used in the experiment *woCB* is the one from the baseline system, which is composed of only a subjective lexicon. In the experiment *onlyS*, only predicate-argument structure is incorporated to extract features from other chunk as in the baseline system. Experimental results for each of the opinion factors are shown in the tables below. To verify that the differences in performance from each experiment are not by chance, the statistical significance among the results was tested. Sproat and Emerson (2003) propose a way to decide whether different precision and recall measures are significantly different for the results of Chinese word segmentation. By assuming that a binomial distribution is appropriate for the experiments, the confidence interval from each experiment is decided given the Central Limit Theorem for Bernoulli trials. Based on their method, the 95% confidence intervals for each experiment for identifying opinion factors were calculated with the same assumption of binomial distribution.

$$C_p = \pm 2\sqrt{p(1-p)/N}$$

p= precision rate : the probability that retrievd as an opinion factor is really an opinion factor

N : No. of each opinion factor annotated in the corpus

(4.5)

$C_r = \pm 2\sqrt{p(1-p)/N}$

p= recall rate : the probability that opinion factors are successfully retrieved

N : No. of each opinion factor annotated in the corpus

(4.6)

After calculating c_p and c_r based on the equation 4.5 and equation 4.6 respectively, it is investigated whether the 95% confidence level of any of two experiments overlap. Two systems are determined to be significantly different if at least one of either c_p and c_r are different. Although the confidence level of either precision-based (c_p) or recall-based (c_r) overlap between parts of the experiment pairs, no two experiments show the overlapped confidence level of both precision-based (c_p) and recall-based (c_r). Therefore, the improvements in the system's performance are verified to be meaningful.

Table 4.8 shows the results of identifying opinion holders from each experiment. In the case of English, *SYSTEM-E* yields the best results in precision, recall and F-score. As the performance of holder identification in Korean is much lower, adding other language data in training is not beneficial in this case, while it helps to enhance the performance for Korean. Unlike the English results, the systems making use of whole features and the feature extraction process do not yield improved results in Korean. Precision is the best in the experiment *woCB*, while the experiment *onlyS* shows the best recall and F-score. That is, it is shown that using possible lexical clues other than the subjective lexicon does not improve the result. However, the clustering strategy and adding more candidates in addition to the subjective lexicon are verified to benefit accuracy in most cases. The difference in the baseline system and the experiment *onlyS* is that the experiment makes use of the clustered lexical feature dictionary of the proposed system. The overall result

Table 4.8: Evaluation: Holder (H) identification of each step in the proposed system

English						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	281	0.393	±0.0583	0.078	±0.032	0.131
woCB	281	0.434	±0.0591	0.082	±0.0327	0.138
woCF	281	0.494	±0.0597	0.416	±0.0588	0.452
onlyS	281	0.477	±0.0596	0.445	±0.0593	0.46
SYSTEM-E	281	**0.536**	±0.0595	**0.477**	±0.0596	**0.505**
SYSTEM-A	281	0.333	±0.0562	0.342	±0.0566	0.337
Korean						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	188	0.333	±0.0687	0.027	±0.0236	0.049
woCB	188	**0.476**	±0.0728	0.053	±0.0327	0.096
woCF	188	0.347	±0.0694	0.09	±0.0417	0.143
onlyS	188	0.333	±0.0687	**0.122**	±0.0477	**0.179**
SYSTEM-E	188	0.221	±0.0605	0.08	±0.0396	0.117
SYSTEM-A	188	0.273	±0.0650	0.096	±0.0430	0.142

of the experiment *onlyS* is better than the baseline system and worse than the proposed system, which implies that both the clustered feature dictionary and incorporating syntactic structure improve the result.

Experiments with varying size of training data To more deeply investigate the performance of the system, experiments with varying size of training data were performed. As the size of the annotated data collected for the current system is limited, the opinion identification results presented in this section could be improved with more training data. Figure 4.6 shows the schematic representation of experiments with three different sets of training data. The annotated data were divided into three groups depending on the date of the newspaper across all three news agencies: up to September 2007 (Eng1, Kor1), Oct 2007 (Eng2, Kor2) and November 2007 (Eng3, Kor3). The first sets of training data (Eng1 and Kor1) contain 30 English and 38 Ko-

Table 4.9: Evaluation: Topic (T) identification of each step in the proposed system

English						
Holder	N	Precision	c_p	Recall	c_r	F-score
Baseline	1289	0.319	±0.0260	0.127	±0.0185	0.182
woCB	1289	0.237	±0.0237	0.180	±0.0214	0.205
woCF	1289	0.369	±0.0269	0.230	±0.0234	0.283
onlyS	1289	0.286	±0.0252	0.171	±0.021	0.214
SYSTEM-E	1289	**0.375**	±0.027	0.211	±0.0227	0.27
SYSTEM-A	1289	0.295	±0.0254	**0.29**	±0.0253	**0.293**
Korean						
Holder	N	Precision	c_p	Recall	c_r	F-score
Baseline	1227	0.286	±0.0258	0.095	±0.0167	0.142
woCB	1227	0.242	±0.0245	0.32	±0.0266	0.276
woCF	1227	0.330	±0.0268	0.22	±0.0237	0.264
onlyS	1227	0.213	±0.0234	0.188	±0.0223	0.2
SYSTEM-E	1227	0.336	±0.027	**0.237**	±0.0243	0.278
SYSTEM-A	1227	**0.342**	±0.0271	**0.237**	±0.0243	**0.28**

Table 4.10: Evaluation: Negative Topic (NT) identification of each step in the proposed system

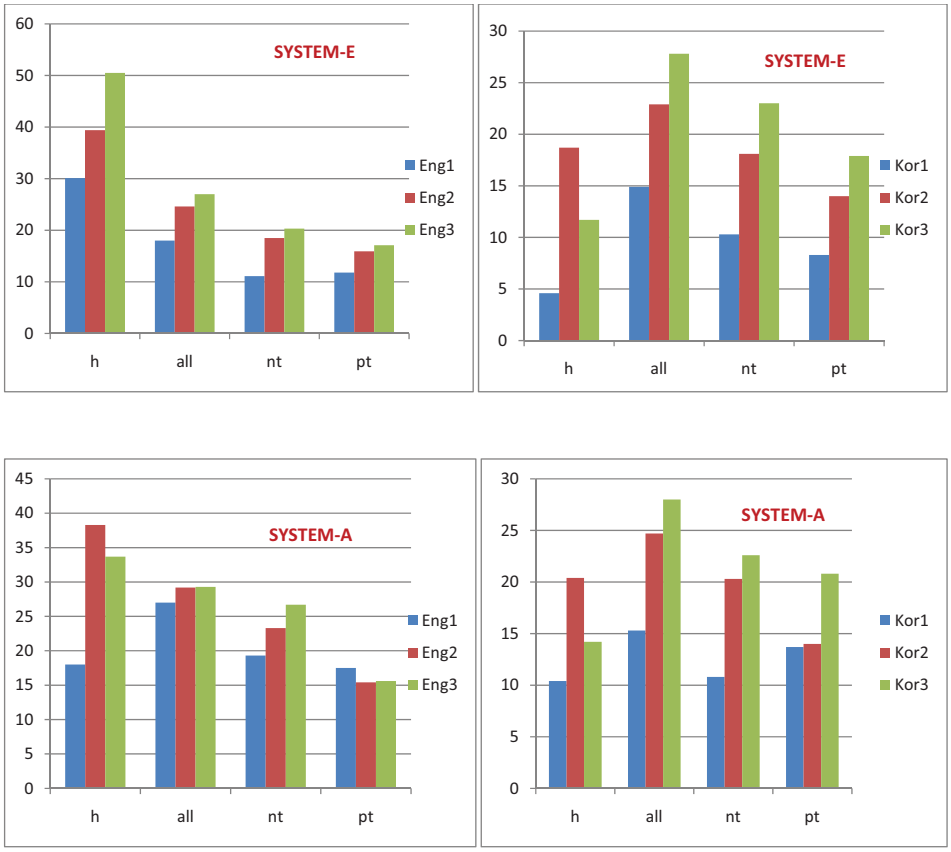
English						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	832	0.282	±0.0312	0.112	±0.0219	0.16
woCB	832	**0.369**	±0.0335	0.107	±0.0214	0.166
woCF	832	0.303	±0.0319	0.12	±0.0225	0.172
onlyS	832	0.269	±0.0307	0.133	±0.0235	0.178
SYSTEM-E	832	0.332	±0.0327	0.147	±0.0246	0.203
SYSTEM-A	832	0.273	±0.0309	**0.262**	±0.0305	**0.267**
Korean						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	778	0.218	±0.0296	0.067	±0.0179	0.102
woCB	778	0.216	±0.0295	**0.198**	±0.0286	0.207
woCF	778	**0.278**	±0.0321	0.184	±0.0278	0.221
onlyS	778	0.234	±0.0304	0.162	±0.0264	0.191
SYSTEM-E	778	0.274	±0.032	**0.198**	±0.0286	**0.23**
SYSTEM-A	778	0.272	±0.0319	0.193	±0.0283	0.226

Table 4.11: Evaluation: Positive Topic (PT) identification of each step in the proposed system

English						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	460	**0.259**	±0.0409	0.104	±0.0285	0.149
woCB	460	0.109	±0.0291	**0.176**	±0.0355	0.135
woCF	460	0.254	±0.0406	0.133	±0.0317	**0.175**
onlyS	460	0.16	±0.0342	0.124	±0.0307	0.14
SYSTEM-E	460	0.221	±0.0387	0.139	±0.0323	0.171
SYSTEM-A	460	0.185	±0.0362	0.135	±0.0319	0.156
Korean						
	N	Precision	c_p	Recall	c_r	F-score
Baseline	449	0.19	±0.037	0.071	±0.0242	0.104
woCB	449	0.104	±0.0288	**0.205**	±0.0381	0.138
woCF	449	0.265	±0.0417	0.12	±0.0307	0.165
onlyS	449	0.107	±0.0292	0.131	±0.0318	0.118
SYSTEM-E	449	0.246	±0.0407	0.14	±0.0328	0.179
SYSTEM-A	449	**0.29**	±0.0428	0.163	±0.0349	**0.208**

rean files among the whole of 113 and 121 files respectively. The second sets (Eng2 and Kor2) contain 71 English and 81 Korean files. As shown in the figure, the performance of the SYSTEM-E was improved with more training data in most cases. The only exception is the holder identification in Korean, which suggests the system needs improvement with deeper investigation of expressing patterns. Other than holder identification, performance improvement with bigger size of training data is more remarkable in Korean than in English. The performance of SYSTEM-A shows less consistent increase from more training data, possibly because the effect of using other language data as training is different depending on the datasets. For example, the dataset Eng1 shows much more improvement in topic identification (all) by adding another language data in training (SYSTEM-E to SYSTEM-A) than Eng2 and Eng3, so the performance with the dataset Eng1 is even better than the performance with Eng2 or Eng3.

Figure 4.6: Results with varying size of training data (F-score(%)): SYSTEM



Evaluation with the MPQA corpus Additional experiments with the MPQA corpus were performed to verify whether the proposed system works for the existing corpus with consistent accuracy. The MPQA corpus contains

692 English documents where the detailed factors of subjective expressions are annotated. Opinion target and attitude annotation have been added in the most recent version (2.0) (Wilson, 2008). The opinion target annotation is along with the annotation for the attitude types, intensity and the polarity of the opinion. Opinion targets with neutral polarity are annotated as well as the targets with positive or negative polarity. Attitude types annotated in the corpus include sentiment, arguing, agreement, speculation and others. Among the opinion targets with the attitude types, targets with arguing attitude are omitted in the experiment as the concept of positive and negative polarity in arguing is not consistent with those in the present system. As targets without polarity are also annotated in the MPQA corpus, two different approaches are used in testing with the MPQA corpus. In the previous section, the task of identifying opinion factors from the annotated corpus was treated as a multi-class classification problem. In the case of the MPQA corpus, bi-class classification for each opinion factor was also tried: None-Holder;None-Topic;None-Positive Topic;None-Negative Topic (System-I in Table 4.12.)

As shown in Table 4.12, the result of the experiment using the MPQA corpus in training and testing shows improved results compared with the baseline result. Notably, the results of *System-I* which classify each opinion factor independently show better results: the recall rate of topic (T) was most significantly improved. In the case of the annotated corpus, however, the two different approaches do not yield noticeable differences. Considering that the experiment was performed in monolingual data, both the precision and recall of the experiment shows better performance than those of the experiment with the annotated corpus except for the case of negative topic (NT). As the MPQA corpus is mostly news articles and the annotated data comes from editorials, there should exist differences in annotation. First of all,

Table 4.12: Evaluation: MPQA corpus

	Baseline			Proposed system			System-I		
	P	R	F	P	R	F	P	R	F
Holder	53.6	16.8	25.6	73.7	48.2	58.3	71.6	50.7	59.4
Topic	35.6	13.7	19.8	40.6	15.4	22.3	40.3	29.0	33.7
NT	25.6	8.3	12.5	31.2	8.7	13.7	31.3	10.8	16.1
PT	20.1	7.3	10.7	27.2	11.3	16.0	28.0	15.1	19.6

·System-I: Performing machine learning independently for each opinion factor

the seeming opinions in editorials could be treated as facts in news articles, as it is generally assumed that news articles deal with objective facts if no opinion holder is present. In other words, only opinion targets with more obvious clues are annotated in news articles compared with editorials. In the same way, opinion holders in news articles are represented with more clear clues for opinions. The most regular pattern, the source of the speech event, is much more frequent in news articles than editorials. For these reasons, the annotation of opinion factors in editorials is more difficult as suggested with the survey from annotators in (Wilson, 2008), in that the deep linguistic structure and nuance should be considered.

As the MPQA corpus is the existing resource utilized by many previous researchers, the results of identification opinion holder and topic identification are compared with the previous studies in Table 4.13 and Table 4.14 respectively. As stated in chapter 2, opinion holder identification has been attempted in several pervious studies while topic identification has not been explored much. Choi et al. (2006) jointly extract opinion expressions and holders of opinion using integer linear programming, and yield the best results among the published works on the MPQA corpus. As shown in Table 4.13, the precision of the proposed system is comparable with (Choi et al., 2006)'s result while recall and F-score are worse. Interestingly, the SYSTEM (A+M),

Table 4.13: Holder identification from the MPQA corpus

Holder	P (%)	R (%)	F (%)
Choi et al. (2006)	75.7	80.6	78.1
SYSTEM-I	71.6	50.7	59.4
SYSTEM (A+M)	90.1	23.7	37.6

·System (A+M): training with annotated corpus as well as MPQA corpus

Table 4.14: Topic identification from the MPQA corpus

Topic	P (%)	R (%)	F (%)
Bloom and Argamon (2010)	11	37	17
SYSTEM-I	40.3	29	33.7

making use of both the annotated corpus and the MPQA corpus in training, yields much improved results in precision (90.1%) with decreasing results in recall. As topic and attitude annotation in the MPQA corpus has been recently added, Bloom and Argamon (2010) is the only published work that could be compared with the result from the current system for topic identification. They extract opinion expressions first, then identify the opinion topics of the expressions using linkage specifications. As shown in Table 4.14, both the precision and F-score of the proposed system are greater than the results from (Bloom and Argamon, 2010).

4.6 Conclusion

In this chapter, a sentiment analysis system for identifying opinion topic, holder and polarity was designed and evaluated with an annotated corpus. The most notable aim of the proposed system is to work with more than one language effectively. I pursue the system which explores the linguistic structure and for the way that opinions are expressed for each language at

the same time, not just making use of a system designed for one designated language. The experimental results verify that words other than those in the subjective lexicon play a role in expressing opinions. Also, clustering words with the same meaning across languages improves the overall performance by enhancing the recall rate, although there is some loss of precision. Various types of linguistic features and making use of linguistic structures other than predicate-argument relations are verified to improve the performance as well. With all these meaningful achievements, however, I should admit there is room for improvement in performance. As the present system aims for multilingual data, all the features and feature extraction steps are designed to be parallel between languages, which could induce less accuracy in each step for each language. For example, in extracting features from other chunks, possible predicate-arguments pairs extracted from PropBank database are used instead of performing separate semantic-role labeling. Using more accurate semantic-role labeling systems for each language when extracting features could be one possible way of improving baseline results.

CHAPTER 5

EXPRESSING OPINIONS ACROSS LANGUAGES

One of the main objectives of this dissertation is exploring opinions across languages. Most previous studies on sentiment analysis have focused on English, so resources for sentiment analysis including subjectivity lexicons and sentence sentiment classifiers are limited to English only. As more and more multilingual and multicultural information becomes available on the web, there is an increasing need to mine opinions from multilingual corpora. Studies on sentiment analysis from multilingual corpora have been attempted by either applying English resources and systems to other languages by cross-lingual mapping or performing the sentiment analysis separately for each language. The method of utilizing English resources and cross-lingual mapping by machine translation has an advantage in that it could make use of existing resources, and it is relatively easy to expand to other languages. However, the performance of the projected system in other languages is likely to be less accurate than that of the English system due to the mapping errors in the process. Banea et al. (2010) demonstrate that multilingual data translated into English are beneficial for English sentence subjectivity classification, but the results for other languages are worse than English. This drawback should be worse for the identification of detailed opinion factors, as this requires much deeper linguistic analysis. Moreover, the system cannot capture the linguistic clues used to identify opinion factors present only in the language, as the system is based on English. For example, morphological

features in Korean play a crucial role in opinion factor identification, which might not be considered in the English system. Therefore, as described in chapter 4, I pursue a multilingual analysis directly working on each language with a unified system. In addition to exploring the separate ways that opinions are expressed for each language, the expression of opinion factors across languages was investigated using parallel documents in the annotated corpora.. To get the maximum performance for each language, it seems ideal to build a carefully designed monolingual system suited for each language. However, a series of separate monolingual systems for multilingual analysis should not be the appropriate solution. Not only does it require much more effort, it is also not possible to make use of the possible cross-lingual reinforcement. The simplest cross-lingual reinforcement could be one benefit of using an annotated corpus from additional language data as shown in the result of the experiment SYSTEM-A in chapter 4. Identification of opinion factors for both Korean and English shows improved results by adding more cross-lingual training data. More importantly, when the same document is presented in more than one language, or parallel data, it is expected that more direct benefits from the cross-lingual reinforcement could be obtained. With this aim, cross-lingual features from parallel corpora were designed and added to the proposed sentiment analysis system. Preprocessing to extract parallel sentences by bilingual sentence alignment is described later in this chapter in section 5.1, and agreement in polarity between parallel sentences is investigated with the extracted pairs in section 5.2. Finally, extracting cross-lingual features and experimental results are described in detail in section 5.3 and section 5.4 respectively.

5.1 Bilingual sentence alignment

In the annotated corpus, although most of the texts are bilingual data (pairs of the same date and same topic), they are not always direct translations of each other. Some English texts only contain a summary of the Korean texts. More frequently, sentences do not match one-by-one between parallel texts. Therefore, bilingual sentence alignment was performed as a preprocessing step to extract cross-lingual features. Brown et al. (1991) implement a bilingual sentence alignment algorithm using only sentence length which is calculated from the number of words in each sentence, while Chen (1993) use lexical information by way of word-for-word translation. In the present study, lexical information as well as sentence length were used as features for sentence alignment. Two types of bilingual dictionaries were used to capture the parallel lexical features within aligned sentences: an existing machine-readable bilingual dictionary and a named-entity dictionary. As most of named-entities in editorial texts are not present in the bilingual dictionary, so a bilingual named-entity dictionary from the collected data was constructed. As a first step, named-entities from each language corpus were extracted. English named-entities were extracted using the named-entity recognizer described in (Li et al., 2004), based on the SNoW machine learning toolkit (Carlson et al., 1999). A similar system for extracting Korean named-entities was implemented and used. As a result, 577 English named-entities and 376 Korean named-entities were extracted. A phonetic transliteration model (Yoon et al., 2007) was used for all 216,952 (577×376) English-Korean word pairs, and the top 5 ranked transliteration pairs were extracted. Finally, a 376 item English-Korean bilingual named-entity dictionary was constructed after manually pruning for finding the answer by the author.

Figure 5.1: Algorithm for Bilingual sentence alignment

Algorithm: Bilingual sentence alignment

Input: Korean/English sentence (S_i, S_j) $0<i<n$, $0<j<m$
Input: Bilingual dictionary (bdic), Named-entity dictionary (nedic)
Output: aligned sentence pairs

```
1: for i=1 to n
2:         for j=1 to m:
3:                 Alignscore=0
3:                 if abs(i-j)<5 do
4.                         Alignscore=dlen+nsco+dsco+wsco
                           dlen=abs(log(len(Si)-len(Sj))/max(len(Si), len(Sj)))
                           nsco+= No. of word pairs in nedic ×3
                           dsco+= No. of word pairs in bdic ×1.5
                           wsco+= No . of same script word pairs ×2
5.                 retrieve Si-Sj pair of max(Alignscore),
                   if not Sj in the previous retrieved pairs
```

With the prepared dictionaries, a sentence alignment was performed based on the algorithm illustrated in Figure 5.1. As illustrated, the sentence alignment was performed with a five-line window: sentence pairs which are farther than five lines in distance were not considered as candidates for alignment. The alignment score was calculated as a sum of the sentence length score (dlen), the named-entity match score (nsco), the word match score (dsco) and the words of the same scripts match score (wsco).

For evaluation of the bilingual sentence alignment, five files were randomly selected from each news agency which were the sources of the annotated corpus, and sentence-alignments were manually performed for reference. As the nature of the parallel corpora could be different depending on the news agencies, evaluation of sentence-alignment was performed separately for each agency. As shown in Table 5.1, the accuracy of the sentence alignment of

Table 5.1: Evaluation: bilingual sentence alignment

Dong-A			Hani			Joins		
P (%)	R (%)	T-P(%)	P (%)	R (%)	T-P(%)	P (%)	R (%)	T-P(%)
92.9	60.6	61.5	89.7	57.8	70	51.6	28.7	23.5
No_si	No_pa	No_tr	No_si	No_pa	No_tr	No_si	No_pa	No_tr
2	81	13	1	99	10	43	84	17

·P: Precision, R: Recall; T-P: partial precision for (more than two)-(more than one) sentences alignment
·No_si: No. of sentences not having translated pairs
·No_pa: No. of aligned sentence pairs in reference
·No_tr: (more than two)- (more than one) sentences alignment pairs

Table 5.2: Evaluation: bilingual sentence alignment (Whole data accuracy)

Dong-A (79.3 %)			Hani (84.3 %)			Joins (48.5 %)		
No_mat	No_tr	No_dis	No_mat	No_tr	No_dis	No_mat	No_tr	No_dis
374	16	102	396	19	77	189	9	210

·No_mat: No. of correctly aligned sentence pairs
·No_tr: No. of partially correct aligned pairs of (more than two)- (more than one) sentences
·No_dis: No. of incorrectly aligned sentence pairs

"Joins" is much worse than the other two news agencies, as there are many sentences not having aligned pairs (No_si) in the corpus. It turns out that parallel files from "Joins" are summaries of the other language articles instead of line by line translations. Therefore, recall as well as precision is very low in the sentence alignment of "Joins".

Table 5.2 shows the accuracy of all of the aligned sentences without considering recall, which was evaluated manually by the author. The accuracy of the sentence alignment for each news agency is little less than the evaluation result of randomly selected articles as shown in Table 5.1. It is observed that the accuracy of the sentence alignment varies greatly depending on the article. A total of 1003 sentences which were correctly aligned (including partially correct) were used to extract cross-lingual features in the next section.

5.2 Polarity agreement between parallel sentences

Although parallel sentences are supposed to share the polarity of sentence and opinion factors as they are translations of each other, there still are discrepancies in some cases. The first type of discrepancy in polarity is caused by the difference in subjective meaning between the word pairs.

(3) [The government's intention to put reporters in the closed briefing rooms] is *clear*.

(4) 정부가 사실상 밀폐된 공간이나 다름없는
cengpu-ka sasilsang milphyeytoyn kongkanina talumepsnun
Government-NOM actually closed place like

브리핑룸으로 기자들을 몰아넣으려는 의도는 *뻔하다*.
puliphinglwum-ulo kica-tul-ul molanehulyenun uyto-nun ppenha-ta
briefing-room-to reporter-s-ACC put intention-TOP clear-COMP

(5) [The results of the opinion polls]NT *do not accurately reflect reality*.

(6) [여론조사 결과]가 현실을 정확하게 반영하는 **건** 아니다.
yeloncosa keylkwa-ka hyensil-ul cenghwakhakey panyeng-hanun **ke-n** ani-ta
opinion-polls result-NOM reality-ACC accurately reflect-ATTR **thing-TOP** not-COMP

For example, the word *clear* in the sentence (3) is used as the correspondent of the word *뻔하다 (ppenha-ta)* in the sentence (4). Although these two words share some meaning (obvious), the Korean word *뻔하다 (ppenha-ta)* contains a subjective meaning with strong negative polarity while the English word *clear* does not. Therefore, the English sentence (3) doesn't seem to obviously contain negative polarity compared with the Korean sentence (4). On the other hand, the negative polarity captured in the sentence (5) is not apparent in the Korean sentence (6) because of the difference in expressions. Expression *반영하는 **건** 아니다 (panyeng-hanun **ke-n** ani-ta)* in the sentence (6) does

Table 5.3: Sentence polarity agreement between parallel sentences: Kappa statistics

κ: 0.77	P	N	PN	None	Total
P	144	3	1	19	167
N	7	244	2	41	294
PN	0	7	11	1	19
None	19	23	1	303	346
Total	170	277	15	364	826

contain the dubious position of the writer unlike the obvious negative polarity in the sentence (5), as a result of the addition of a topic marker. Other types of discrepancies in polarity between parallel sentences occurred in the cases where one-to-one mapping is not possible between words. In Korean, a missing subject is possible and fairly frequent unlike English, which leads to sentences with an omitted topic. Moreover, sometimes a different topic is created by the different modes of expression.

With the extracted parallel sentence pairs, agreement in sentence polarity across languages was calculated in both Kappa statistics and the *agr* metric. Among 1003 aligned sentence pairs extracted in section 5.1, 826 sentence pairs where annotation for all opinion factors agree between annotators (section 3.4) were used for calculating agreement between parallel sentences. According to Mihalcea et al. (2007)'s investigation with English and Romanian, sentence-level subjectivity is preserved in most cases. In this study as well, positive and negative sentence polarity in one language tends to be retained as the same in another language in parallel sentences (more than 0.86 of mean *agr*), as shown in Table 5.4. Less simpler sentence polarity (PN), on the other hand, shows a lesser degree of agreement.

Agreement on opinion factor annotation between parallel sentences cannot be perfectly calculated without manual work, unless word alignment within

Table 5.4: Sentence polarity agreement between parallel sentences: *Agr* ratio

sentence	recall (Eng‖Kor)	recall (Kor‖Eng)	mean
N	0.83	0.88	0.86
P	0.86	0.85	0.86
PN	0.58	0.73	0.66

Table 5.5: Agreement on opinion factor annotation between parallel sentences: *Agr* ratio

opinion factor	recall (Eng‖Kor)	recall (Kor‖Eng)	mean
H	0.95	0.77	0.86
NT	0.86	0.81	0.84
PT	0.82	0.79	0.81

parallel sentences is performed first. As the second best option, the *agr* rates of each opinion factor within parallel sentence pairs were calculated: see if the same type of opinion factor is present between parallel sentences. Although this method cannot perfectly show the agreement ratio because of the possibility that different words are annotated as opinion factors between parallel sentences, the general tendency of annotation agreement could be captured. As illustrated in Table 5.5, the annotation of opinion factors were generally retained between parallel sentences (more than 0.81 of *agr* ratio). Noticeably, the recall rate of (Eng‖Kor) is higher than (Kor‖Eng) in all three opinion factors. The difference between the recall rate in the case of holder (H) annotation is the greatest (0.95:0.77). One of the possible reasons for this phenomenon is the frequently missing subject in Korean. Also, considering the Korean culture of indirectly expressing opinions as mentioned in section 4.5, the missing subject phenomenon is expected to occur more frequently when the subject is the holder of an opinion.

5.3 Cross-lingual features

A way to investigate cross-lingual effects in extracting opinion factors is proposed, with the parallel sentences extracted from the annotated corpus. The hypothesis is that the linguistic features of the same types of opinion factors in parallel English and Korean chunks (noun phrases) could be stronger clues for identifying opinion factors. Two types of cross-lingual features are pursued here: shared features and features from other languages. If the same linguistic feature is found within the same type of opinion factors from two languages, the confidence level about clues for opinion factors should increase. In addition to the shared features, co-occurring feature pairs from each language and in the same type of opinion factors could also provide hidden clues for identifying opinion factors

To extract effective cross-lingual features, *mutual information* was calculated between the feature pairs. *Mutual Information* compares the joint probability of x and y with the probability of x and y independently (Fano, 1961) based on the following equation:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{5.1}$$

Church and Hanks (1990) suggest the way to apply the concept of *mutual information* to measure word association norms. Following this, *mutual Information*, especially Pointwise Mutual Information (PMI) has been used to calculate the semantic association between words in previous studies (Turney, 2001; Turney and Littman, 2003). In this study, however, *mutual information* was adopted to calculate the association between features used to express opinion factors. Here, the joint probability of the x (feature from language A which is tested: *wfea*) and the y (feature from language B: *cfea*) suggests the

Figure 5.2: Extracting cross-lingual features

English sentence:

[North Korea and the U.S.] H wrapped up the first round of talks on normalizing relations yesterday and showed satisfaction by and large at **[the result]PT**.

Korean counterpart:

[북한과 미국]H 은	어제	뉴욕에서	끝난	**[관계정상화 1 차 회의 결과]PT 에**
[pwukhan-kwa mikwuk]H-un	ecey	nyuyokeyse	kkuthna-n	**[kwankyeycengsanghwa 1chahoyuy kyelkwa]PT-ey**
[Norh Korean-and U.S.A]-TOP	yesterday	Newyork-LOC	end	**[talks-on –normalizing-relations result]PT**

대해	대체로	만족감을	나타냈다.
tayhay	taycheylo	mancokkam-ul	nathanay-ss-ta
about	by-and-large	satisfaction-ACC	show-Past-Decl

[North Korea and the U.S.] H == [북한과 미국]H 은

$wfea_i$ 0<i<a $cfea_j$ 0<j<b

[the result]PT == [관계정상화 1 차 회의 결과]PT 에

$wfea_i$ 0<i<m $cfea_j$ 0<j<n

probability that x and y occur in parallel noun phrase pairs with the same annotated opinion-factor.

Examples of parallel sentences used to extract cross-lingual features are illustrated in Figure 5.2.

In the parallel sentences in Figure 5.2, holder (H) and positive topic (PT) are annotated in both sentences. Say, a noun phrase annotated as a positive topic (PT) in an English sentence, *the result*, has m number of features, while parallel counterpart PT-Korean noun phrase , *kwankyeycengsanghwa 1cha-hoyuy kyelkwa-ey*, has n number of features. If opinion factor identification is performed with English as the testing data, *mutual information* between $wfea_i$ and $cfea_j$ in the PT-tagged phrase is calculated (0<i<m,0<j<n) based on the Equation 5.1 where x=$wfea_i$, y=$cfea_j$. The probabilities of each fea-

ture P(x) and P(y) are estimated by counting the number of observations of each feature in the extracted feature set from all data, f(x) and f(y), and normalizing by N, the size of feature set. The joint probability P(x,y) is estimated by counting the number of times normalizing by N, that $wfea_i$ occurs in PT-tagged noun phrase and $cfea_j$ occurs in the noun phrase of the parallel counterpart tagged also as PT. If the result of *mutual information* between $wfea_i$-$cfea_j$ pairs is above the threshold and the pair exclusively occurs in PT-tagged chunks, a cross-lingual feature $cfea_j$ is extracted. In the present experiment, f(x)>4, f(y)>4 and I(x,y)>8 are used as the threshold. The statistics of the features were drawn from the training data for each run, then applied to the testing data: if $wfea_i$ occurs in noun phrase chunk in testing data, $cfea_j$ was added as a cross-lingual feature to the feature set of that chunk with designated weight. When $wfea_i$=$cfea_j$, namely a shared feature, the weight of $cfea_j$ for learning is more than the original weight of $wfea_i$, as it is highly likely to be a confident clue for the specific opinion factor. Otherwise, a little less weight is assigned to the cross-lingual feature $cfea_j$ than the original weight of $wfea_i$. In the current study, the weight is assigned as 1:3 for a shared feature and 0.8:1 for other cross-lingual features compared to the original weight of $wfea_i$.

5.4 Experimental result

Based on the experimental set-up described in chapter 4, experiments were performed with additional cross-lingual features. As in the previous experiments, each language corpus was divided into 10 groups to perform 10-fold cross-validation for evaluation. Table 5.6 shows the evaluation results of the experiments incorporating cross-lingual features (CROSS-A) compared with

Table 5.6: Evaluation: Effect of cross-lingual features

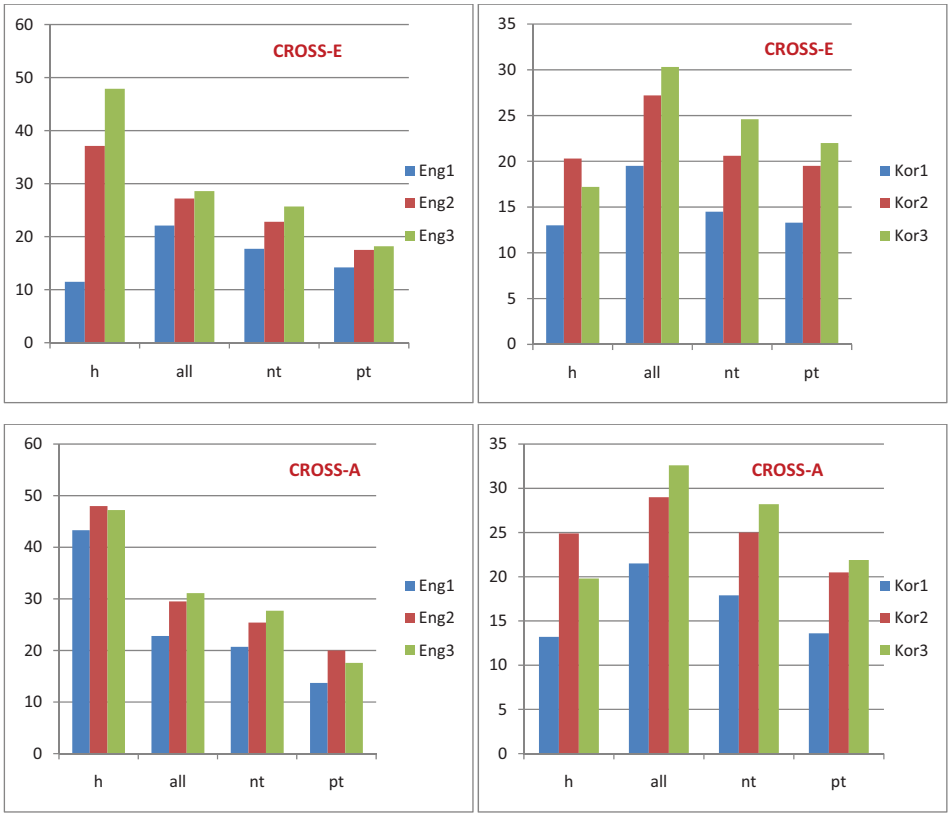
	English					
	SYSTEM-A			CROSS-A		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Holder	33.3	34.2	**33.7**	44.6	50.2	**47.2**
Topic	29.5	29.0	**29.3**	30.0	32.3	**31.1**
NT	27.3	26.2	**26.7**	25.9	29.7	**27.7**
PT	18.5	13.5	**15.6**	18.0	17.2	**17.6**
	Korean					
	SYSTEM-A			CROSS-A		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Holder	27.3	9.6	**14.2**	26.1	16.0	**19.8**
Topic	34.2	23.7	**28.0**	34.6	30.9	**32.6**
NT	27.2	19.3	**22.6**	28.4	28.0	**28.2**
PT	29.0	16.3	**20.8**	25.6	19.2	**21.9**

the system that doesn't use cross-lingual features (SYSTEM-A). In both experiments, data from other language are used in training to test each language. As illustrated, cross-lingual features improve the F-score results in all opinion factors for both English and Korean.

Figure 5.3 demonstrates there is more room for performance improvement if more annotated data are available, by illustrating the improved results with the use of bigger size of training data. The dataset used in these experiments are the same as in Figure 4.6 in section 4.5, which are about 1/3, 2/3 and the whole of the annotated data size.

Comparisons of four different experiments (with/ without cross-lingual features, whether or not training with both language data) are shown in the following tables for each opinion factor. Confidence intervals (c_p and c_r) described in section 4.5 are presented with the precision, the recall and the F-score of each experiment. The improvements of the system with cross-lingual features are verified to be meaningful with the results of the confi-

Figure 5.3: Results with varying size of training data (F-score(%)): CROSS



dence intervals: at least one of either precision-based (c_p) or recall-based (c_r) is independent of each other between any two pairs from all four experiments. The schematic representation of the F-score(%) results of four experiments

for each opinion factor is illustrated in Figure 5.4 - Figure 5.7.

As illustrated, the system using data from both languages in training with cross-lingual features (CROSS-A) shows the best performance in terms of F-score, topic (T) and negative topic (NT) identification. On the other hand, Holder (H) identification in English shows the best performance in SYSTEM-E. As shown in the previous chapter, the Korean data is not beneficial in holder identification in English. However, when comparing SYSTEM-A and CROSS-A, cross-lingual features are verified to improve results even in the holder identification for English by 13.5%. In the identification of positive topic (PT), CROSS-E shows the best performance for both English and Korean. Generally, the Korean results are shown benefit greater from the use of cross-lingual features in that both the precision and recall results are consistently improved for all opinion factors. The results of topic (T) identification in English, on the other hand, are improved based on improvements in recall with a little loss in the precision rates. It is shown in the figures that the performance of the system is enhanced with the use of the other language data (E→A) and cross-lingual features (SYSTEM→CROSS).

5.5 Conclusion

In this chapter, the expression of detailed opinion factors from cross-lingual data were investigated. With the parallel sentences extracted from the annotated corpus, cross-lingual agreement in expressing opinion factors as well as sentence polarity was presented. Expressed opinion factors are mostly matched between parallel sentences with a few exceptions. Although the surface structure of each language is different, the clues for the expression of opinion factors from the other language data could be extracted with the

Table 5.7: Effect of Cross-lingual features: Holder (H) identification

English						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	281	**0.536**	±0.0595	0.477	±0.0596	**0.505**
CROSS-E	281	0.482	±0.0596	0.477	±0.0596	0.468
SYSTEM-A	281	0.333	±0.0562	0.342	±0.0566	0.337
CROSS-A	281	0.446	±0.0593	**0.502**	±0.0597	0.472
Korean						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	188	0.221	±0.0605	0.08	±0.0396	0.117
CROSS-E	188	0.203	±0.0587	0.149	±0.0519	0.172
SYSTEM-A	188	**0.273**	±0.0650	0.096	±0.0430	0.142
CROSS-A	188	0.261	±0.0641	**0.16**	±0.0535	**0.198**

·-E: learning with each language data only
·-A: training with both language data in learning

Table 5.8: Effect of Cross-lingual features: Topic (T) identification

English						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	1289	**0.375**	±0.027	0.211	±0.0227	0.27
CROSS-E	1289	0.367	±0.0269	0.234	±0.0236	0.286
SYSTEM-A	1289	0.295	±0.0254	0.29	±0.0253	0.293
CROSS-A	1289	0.3	±0.0255	**0.323**	±0.0261	**0.311**
Korean						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	1227	0.336	±0.027	0.237	±0.0243	0.278
CROSS-E	1227	0.291	±0.0259	**0.316**	±0.0265	0.303
SYSTEM-A	1227	0.342	±0.0271	0.237	±0.0243	0.28
CROSS-A	1227	**0.346**	±0.0272	0.309	±0.0264	**0.326**

Table 5.9: Effect of Cross-lingual features: Negative Topic (NT) identification

English						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	832	0.332	±0.0327	0.147	±0.0246	0.203
CROSS-E	832	**0.337**	±0.0328	0.208	±0.0281	0.257
SYSTEM-A	832	0.273	±0.0309	0.262	±0.0305	0.267
CROSS-A	832	0.259	±0.0304	**0.297**	±0.0317	**0.277**
Korean						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	778	0.274	±0.032	0.198	±0.0286	0.24
CROSS-E	778	0.222	±0.0298	0.275	±0.032	0.246
SYSTEM-A	778	0.272	±0.0319	0.193	±0.0283	0.226
CROSS-A	778	**0.284**	±0.0323	**0.28**	±0.0322	**0.282**

Table 5.10: Effect of Cross-lingual features: Positive Topic (PT) identification

English						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	460	0.221	±0.0387	0.139	±0.0323	0.171
CROSS-E	460	**0.227**	±0.0391	0.152	±0.0335	**0.182**
SYSTEM-A	460	0.185	±0.0362	0.135	±0.0319	0.156
CROSS-A	460	0.18	±0.0358	**0.172**	±0.0352	0.176
Korean						
	N	Precision	c_p	Recall	c_r	F-score
SYSTEM-E	449	0.246	±0.0407	0.14	±0.0328	0.179
CROSS-E	449	0.248	±0.0408	**0.198**	±0.0376	**0.22**
SYSTEM-A	449	0.29	±0.0428	0.163	±0.0349	0.208
CROSS-A	449	**0.256**	±0.0412	0.192	±0.0372	0.219

Figure 5.4: Schematic representation of F-score(%) results: Holder identification

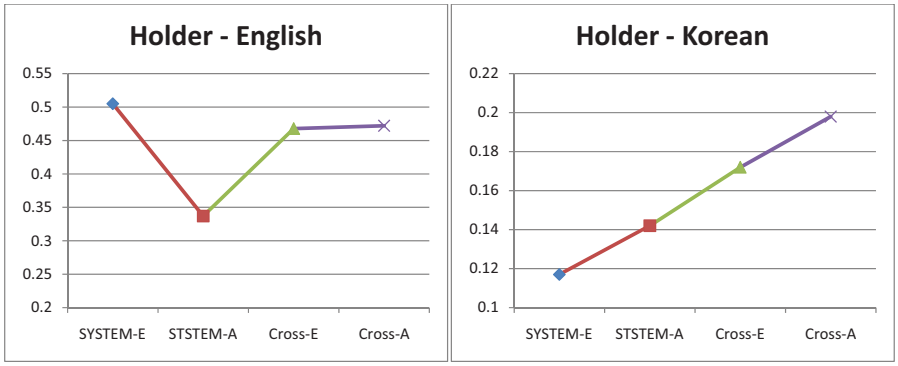


Figure 5.5: Schematic representation of F-score(%) results: Topic identification

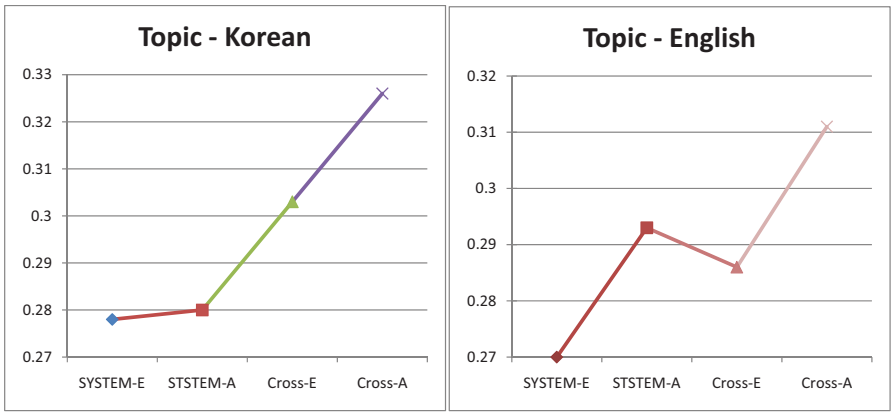


Figure 5.6: Schematic representation of F-score(%) results: Negative Topic identification

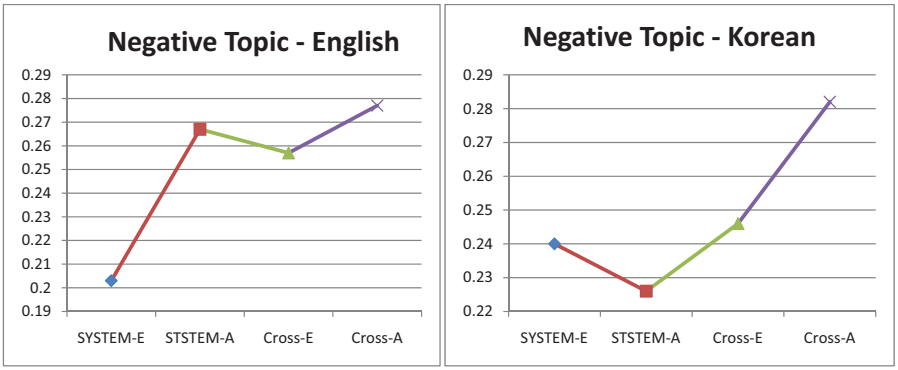
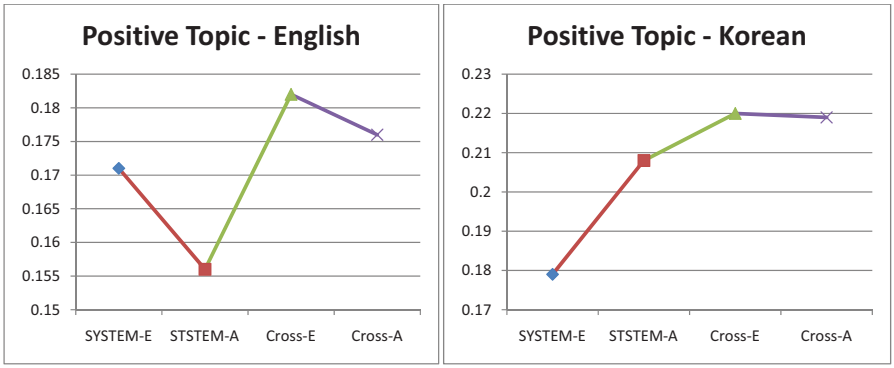


Figure 5.7: Schematic representation of F-score(%) results: Positive Topic identification



use of the unified system for exploring linguistic structures. Cross-lingual features from extracted parallel sentences were designed and applied based on the proposed multilingual sentiment analysis system which directly works with multilingual data. *Mutual information* was used as a statistical method to extract effective cross-lingual features. *Mutual information* between a feature from the chunk of the language tested and a feature from the parallel chunk was calculated, and highly associated feature pairs were extracted as cross-lingual features. It is shown that incorporating additional cross-lingual features improves the performance of the proposed system for both languages, which suggests cross-lingual reinforcement in identifying opinion factors.

CHAPTER 6

CONCLUSION

The present dissertation explores opinions in texts focusing on two main objectives. The first aim is to investigate detailed opinion factors including the opinion holder (H) and topic with polarity (NT, PT) beyond the level of sentence level polarity. These detailed opinion factors could be utilized directly in many applications after extraction. Second, an authentic multilingual system is proposed instead of a system totally depending on one language or the combination of separate monolingual systems. By exploring the linguistic structures used to express opinions in a unified way for all languages, the system benefited from the cross-lingual features which were extracted from parallel sentences. In this chapter, a summary of the present dissertation and the directions for future work are presented.

6.1 Summary

As one of the main procedures in the present dissertation, opinion annotation was performed using bilingual editorials as the corpus. "Opinion" is defined as the evaluative opinion on a specific target correspondent to the definition and types of "attitude" in the appraisal system by Martin and White (2005). Unlike previous resources that annotated opinion expressions as a starting point, opinion factors (holder and topic with polarity) in this study were annotated directly without specifying the clues that signal opinion factors.

Patterns used to express opinions were presented with the examples from the annotated corpus: in addition to the opinion lexis, grammar, pragmatics and context play a role in expressing opinions. Inter-annotator agreement was calculated with using Kappa statistics as well as *agr* metrics: annotation of all opinion factors and sentence polarity show substantial agreement between annotators.

A multilingual sentiment analysis system that automatically identifies the opinion holder and topic is proposed in this study. The proposed system was performed with one step of opinion factor identification using a machine learning algorithm without a separate step of extracting opinion expressions. The input sentences were parsed and chunked and served as a basic unit for feature extraction and machine learning. Opinion factors were extracted as noun phrase chunks from the system. Clustered bilingual feature dictionaries were constructed considering various linguistic factors: lexical, syntactic, morphological and contextual. Then, features for the current chunk were extracted from another set of linguistically related chunks in addition to the features from their own chunk. The proposed system explores the syntactic structure as well as the predicate-argument structure used to extract appropriate features for identifying opinion factors. Experimental results verify that, in general, elements of the lexicon other than the subjective lexicon also play an important role in identifying opinion factors. The clustering strategy also turns out to be beneficial for improving the performance of the system although there exist some drawbacks in precision. The experimental result performed with the MPQA corpus verifies that the proposed system yields consistent accuracy with the existing resources although the domain and annotation scheme are not exactly matched.

The expression of opinions across languages was investigated based on

the multilingual sentiment analysis system that explores the linguistic structures used to identify opinion factors. By making use of parallel sentences extracted from the annotated corpus, agreements in annotation for each of the opinion factors as well as the sentence polarity were calculated. Sentence polarity mostly tends to be retained in the parallel counterpart (more than 0.86 of mean *agr*). Opinion factor annotation also shows high agreement between parallel sentences, with different tendencies for each language. Recall rates of (Eng‖Kor) are higher than (Kor‖Eng), which could be induced from the Korean culture of not directly expressing opinions. Cross-lingual features from parallel sentences were extracted by calculating the *mutual information* of feature sets between parallel chunks. Cross-lingual reinforcement in identifying opinion factors is verified by the improved result of the system that incorporates cross-lingual features.

6.2 Future work

"How does X feel about Y?" is the question that many opinion-related question answering systems seek answers for. To meet the need of retrieving the exact answers about what other people think, identifying detailed opinion factors such as the holder and topic of an opinion is the essential prerequisite. The present dissertation made contributions to the study of detailed opinion factors, an area which has recently gained much interest among researchers. Also, this dissertation proposed a multilingual system for identifying opinion factors which could be reinforced by the data from other languages in the system.

As a future line of research, several directions for improving the performance of the opinion factor identification system are proposed. The first

direction is to enhance the baseline performance for each language by incorporating a more accurate means of exploring linguistic structure. For example, making use of a semantic-role labeling system instead of using the possible predicate-argument relation could improve the baseline performance. However, this was not the direction that I pursued in this proposed system, as building a separate language-dependent system was not the scope of the present dissertation. Considering the realm beyond sentences is another direction for improving the performance of the proposed system. In the review-related domain, the document topic is closely related to the opinion topics, so works on sentiment analysis with topic-modeling show successful results (Lin and He, 2009; Lu and Zhai, 2008; Mei et al., 2007; Titov and McDonald, 2008). Although there are some previous works that detect opinion targets in general texts (news articles) using document topics (Choi et al., 2010; Kim et al., 2008), the performance of co-reference resolution is first required in order to successfully make use of the document topic used to identify opinion topics in the proposed system. This step as well requires language-dependent knowledge and systems. More desirably, another direction for future research would be to design more sophisticated cross-lingual features that capture the relation between parallel sentences in identifying opinion factors.

Another interesting direction for future work with the proposed system is expanding the domains used. The primary corpus used in the present dissertation is editorials, and the proposed system was tested with English news articles as well. Both types of corpora are from well-structured domains, so that linguistic analysis is appropriately performed using standard methods. Putting aside the review-related domain whose characteristics are much different from what the proposed system aims for, personal blogs and public forums for debating political issues are examples of unstructured domains.

Identifying opinion factors from these domains should be more challenging, as we should be faced with difficulty in processing the texts. By expanding the proposed system into these unstructured texts, the system could be more broadly utilized in real applications.

Furthermore, the proposed system, which is currently cross-lingual, could be expanded into a multilingual system working with more than three languages. The required resources are a bilingual dictionary connected to either English or Korean, a parsing engine and a resource similar to PropBank. Chinese could be a readily available third language as there exists a English-Chinese parallel PropBank (Xue and Palmer, 2009) as well as other NLP tools. By incorporating more languages, the effect of the cross-lingual reinforcement is expected to be strengthened.

REFERENCES

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithm for mining association rules. *VLDB*, pages 487–499, 1994.

Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings of the 11rd Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Alfred J. Ayer. *Language, Truth and Logic*. Dover Publications, inc., New York, 1952.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010.

Collin F. Barker and Hiroaki Sato. The Frame-Net data and software. In *Poster and Demonstration at Association for Computational Linguistics*, 2003.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In James G. Shanahan, Janyce Wiebe, and Yan Qu, editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US, 2004.

Douglas Biber and Edward Finegan. Styles of stance in english:lexical and grammatical marking of evidentiality and affect. *Text*, 9:93–124, 1989.

Kenneth Bloom and Shlomo Argamon. Unsupervised extraction of appraisal expressions. In *Lecture Notes in Computer Science*, volume 6085, pages 290–294. Springer Verlag, 2010.

Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Inf. Retr.*, 12(5):526–558, 2009. ISSN 1386-4564.

Eric Breck, Yejin Choi, and Claire Cardie. Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

Peter F. Brown, Jennifer C. Lai, and Nd Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 169–176, 1991.

Andrew Sangpil Byon. The role of linguistic indirectness and honorifics in achieving linguistic politeness in Korean requests. *Journal of Politeness Research. Language, Behaviour, Culture*, 2(2):247–276, 2006.

Andrew J. Carlson, Chad M. Cumby, Jeff L. Rosen, and Dan Roth. The SNoW learning architecture. *Technical Report UIUCDCS-R-99-2101*, UIUC, 1999.

Eugene Charniak. A maximum-entropy-inspired parser. *Technical Report CS99-12*, Brown University, 1999.

Stanley F. Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 1993.

Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language (HLT-EMNLP)*, 2005.

Yejin Choi, Eric Breck, and Claire Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

Yoonjung Choi, Seongchan Kim, and Sung-Hyon Myaeng. Detecting opinions and their opinion targets in ntcir-8. In *Proceedings of NTCIR-8 Workshop*, 2010.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990. ISSN 0891-2017.

Susan Conrad and Douglas Biber. Adverbial marking of stance in speech and writing. In Susan Hunston and Geoff Thompson, editors, *Evaluatin in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, 2000.

Kerstin Denecke. Using SentiWordNet for multilingual sentiment analysis. In *Data Engineering Workshop (ICDEW)*, 2008.

Xiaowen Ding and Bing Liu. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 811–812, 2007.

David R. Dowty. Thematic protp-roles and argument selection. *Language*, 67(3):547–619, 1991.

F. Eemeren, R. Grootendorst, and F. Henkenmans. *Fundamentals of argumentation theory: A Handbook of historical backgrounds and comtemporary developments*. Lawrence Erlbaum Associates, 1996.

Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624, Bremen, Germany, 2005.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06*, pages 417–422, 2006.

Jihyun Eun, Minwoo Chung, and Gary Keunbae Lee. Korean dependency structure analyzer based on probabilistic chart parsing. In *Proceedings of 17th Annual Conference on Human and Cognitive Language Technology*, 2006.

R Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, 1961.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

M.A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1994.

Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. Penn Korean Treebank: Development and evaluation. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation.*, 2002.

Vasileios Hatzivassiloglou and Kathy McKeown. Predicting the semantic orientation of adjectives. In *acl97*, pages 174–181, Madrid, Spain, 1997.

Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI)*, 2004a.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004b.

Daisuke Ikeda, Hiroya Takamura, Lev-Arie Ratinov, and Manabu Okumura. Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.

Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceeding of the European Conference on Machine Learning*, 2006.

Jaap Kamps and Maarten Marx. Words with attitude. In *Proceedings of the first International Conference on Global WordNet*, 2002.

Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

Soo-Min Kim and Eduard Hovy. Determining the sentiments of opinion. In *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

Soo-Min Kim and Eduard Hovy. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domain*, 2005a.

Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing(IJCNLP-05)*, 2005b.

Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*, 2006.

Youngho Kim, Seongchan Kim, and Sung-Hyon Myaeng. Extracting topic-related opinions and their targets in ntcir-7. In *Proceedings of NTCIR-7 Workshop Meeting*, 2008.

Elisabeth Le. Editorials' genre and media roles: Le monde's editorials from 1999 to 2001. *Journal of Pragmatics*, 41(9):1727–1748, 2009.

Xin Li, Paul Morie, and Dan Roth. Robust reading:identification and tracing of ambiguous names. In *Proceedings of NAACL*, 2004.

Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 121–130, New York, NY, USA, 2008. ACM.

John Lyons. *Semantics*. Cambridge University Press, Cambridge, 1977.

James R. Martin and Peter R.R. White. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, London, 2005.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the World Wide Web Conference*, pages 171–180, 2007.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1): 71–105, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumps up? Sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

Hyun Seek Park, Dania Egedi, and Martha Palmer. Recovering empty arguments in Korean. In *Joint Conference of 8th ACLIC and 2nd PacFoCol*, 1994.

Thomas E Payne. *Describing morphosyntax: A guide for field linguists*. Cambridge University Press, Cambridge; New York, 1997.

Livia Polanyi and Annie Zaenen. Contextual valence shifters. *Computing attitude and affect in Text: Theory and Application*, pages 1–10, 2004.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A comprehensive grammar of the English language*. Longman, 1985.

James Rachels. *The Elments of Moral Philosophy*. McGrow-Hill companies, 2007.

Jonathon Read, David Hope, and John Carroll. Annotating expressions of appraisal in english. In *LAW '07: Proceedings of the Linguistic Annotation Workshop*, pages 93–100, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 105–112, 2003.

Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 1106–1111, 2005.

Jr. Robert D. Van Valin. *Exploring the Syntax-Semanctics interface*. Cambridge University Press, 2005.

J. Ruppenhofer, S. Somasundaran, and J. Wiebe. Finding the sources and targets of subjective expressions. In *LREC*, Marrakech, Morocco, 2008.

Yohei Seki, Noriko Kando, and Masaki Aono. Multilingual opinion holder identification using author and authority viewpoints. *Inf. Process. Manage.*, 45(2):189–199, 2009. ISSN 0306-4573.

Richard Sproat and Thomas Emerson. The first international chinese word segmentation bakeoff. In *SIGHAN '03: Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 133–143, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Charles L. Stevenson. *Ethics and Language*. Yale University Press, New Haven, Conn., 1944.

Veselin Stoyanov and Claire Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 817–824, Manchester, UK, August 2008.

Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161, 2004.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, US, 2005. Association for Computational Linguistics.

Geoff Thompson and Susan Hunston. Evaulation: An introduction. In Susan Hunston and Geoff Thompson, editors, *Evaluatin in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, 2000.

S. Tirkkonen-Condit. Explicitness vs. implicitness of argumentation: an intercultural comparison. *Multilingua*, 15(3):257–273, 1994.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA, 2008. ACM.

Peter Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag. ISBN 3-540-42536-5.

Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL-09, 47rd Annual Meeting of the Association for omputational Linguistics*, 2009.

Egon Werlich. *A text grammar of English.* Heidelberg: Quelle and Meyer, 1976.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6.

Janyce Wiebe. Learning subjective adjectives from corpora. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI)*, pages 735–740, 2000.

Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*, pages 475–486, 2005.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational Linguistics*, 30(3): 277–308, 2001.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in languages. *Language Resources and Evaluation*, 39:164–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99, 2006.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Linguist.*, 35(3):399–433, 2009.

Theresa Ann Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh, 2008.

Ian H Witten and Eibe Frank. *Data mining:Practical machine larning tools and techniques*. Elsevier, 2005.

Nianwen Xue and Martha Palmer. Adding semantic roles to the chinese treebank. *Natural Language Engineering*, 15(1):143–172, 2009.

Su-Youn Yoon, Kyoung-Young Kim, and Richard Sproat. Multilingual transliteration using feature based phonetic method. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

Hong Yu and Vasileios Hatzivassiloglou. Toward answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP)*, pages 129–136, 2003.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50, 2006.