Romanization-based Approach to Morphological Analysis in Korean SMS Text Processing

Youngsam Kim
Seoul National University/
Gwanak-1, Gwanak-ro, Gwanak-gu,
Seoul, South Korea
youngsamy@gmail.com

Hyopil Shin
Seoul National University/
Gwanak-1, Gwanak-ro, Gwanak-gu,
Seoul, South Korea
hpshin@snu.ac.kr

Abstract

In this research, we suggest an approach to retrieval-related tasks for Korean SMS text. Most of the previous approaches to such text used morphological analysis as the routine stage of the preprocessing workflow, functionally equivalent to POS tagging. However, such approaches suffer difficulties since Short Message Service language usually contains irregular orthography, atypically spelled words, unspaced segments, etc. Two experiments were conducted to measure how well these problems can be avoided with the transliteration of Korean to Roman letters. In summary, we will argue that such a Romanization-based retrieval method has several advantages since it provides an easier way to preprocess the data with a variety of linguistic rules.

1 Introduction

In this internet era, everyday people express opinions, comments, or sentiments; all of which can be accessed via the web. Particularly with the popularization of mobile computing devices, it has become easier than ever for people to share messages using social media services like Twitter or Facebook. However, such an environment brings new challenges for researchers who aim to analyze or interpret this linguistic data. One of the problems they encounter is that these written texts have a different form than those in published books or articles. They were often called as short message service language, txt-speak, chat-speak, etc. This new data source has received attentions from various fields and researchers working in the field of sentiment analysis and opinion-mining often find that dealing with such texts using traditional approaches is problematic.

For agglutinative languages like Korean, since words are formed by combining lemmas and various affixes, morphological analysis is required to find the functional meaning of each component. Most previous studies used morphological analysis only to preprocess the text, but this approach exhibits several weaknesses when used on the data that is written in SMS-like languages. First of all, texts are often unspaced to save on typing time and sentence length (e.g., Twitter only allows 140 characters per tweet). Secondly, many words are not typed in the same way as their dictionary entries; the letters are changed or reduced to smaller units due to morpho-phonetic variation and abbreviation processes.

This paper will propose a new approach to overcome these shortcomings for morphologically rich languages while making use of Korean case studies. This approach adopts Yale Romanization to transliterate Korean alphabets into Roman letters, which, due to the way it handles Korean characters, allows for a more intuitive and easier way of implementing the relevant rewriting rules and handling morph-phonetic changes.

In Section 2, the problems of morphological analysis will be described and the properties of Korean SMS language will be reviewed. This will be followed up in Section 3 by an introduction to the Romanization-based framework and the method of employing linguistic rules. Section 4 will detail two retrieval experiments which were prepared to show the effectiveness of this approach. The first experiment was designed to observe whether the Romanization method could

International Joint Conference on Natural Language Processing, pages 145–152,
Nagoya, Japan, 14-18 October 2013.

handle unspaced texts. The second experiment explored the possibility of covering phonetic variations of the target words using a small set of linguistic rules.

2 Related Research

Transliteration methods have often been used for the task of keyword matching across different languages (Chen and Ku, 2002; Fujii and Ishikawa, 2001). In contrast, Han (2006) applied the transliteration method to perform part-of-speech tagging for Korean texts using Xerox Finite State Tool. Similarly, this paper proposes using the method not for Korean-English word equivalents but for Korean-to-varied Korean word detection.

2.1 Problems of morphological analysis: lack of lexicon

As the number of the users using social networking services increases rapidly, sentiment analysis or opinion mining capable of automatically extracting the sentiment orientation from online posts has been gaining attention from NLP researchers (Hu and Liu, 2004; Kim and Hovy, 2004; Wiebe, 2000; Pak and Paroubek, 2010). As stated above, Korean is an agglutinative language and the chunks distinguished by space must be further separated into roots and affixes before they can be assigned a part-of-speech tag. This whole procedure is performed by morphological analysis and is critical to determining the meaning of a component. However, it is also known that such analysis can cause errors when not equipped with complete word entries to analyze the text. Such 'lack of lexicon' problems arise because after the morphological analysis categorizes all listed words in the sentence it classifies the remaining words as general nouns (Jang and Shin, 2010). Consider the following.

(1) 너무 진부한 내용
nemu cinpuha-n nayyong
too stale-AD[1] content
'too stale contents'

(2) 너무/a 진부/ncs 하/xpa ㄴ/exm 내용/nc
nemu/a[2]cinpu/ncs ha/xpa n/exm nayyong/nc

(3) 너/npp 무진/nc 부/nc 한/nc 내용/nc
ne/npp mucin/nc pu/nc han/nc nayyong/nc
'you Mujin(place name) wealth resentment contents'

Sentence (3) is a misanalyzed version of sentence (1). The morphological analyzer's dictionary did not include the word entry ('cinbu') so the analyzer had to ignore the previous spacing and take the proper noun ('mucin') as a possible morpheme instead (Jang and Shin, 2010; p. 500).

As can be inferred from examples (1) ~ (3), typical morphological analysis consists of two stages: first, a sentence or clause is decomposed into relevant morphemes and then, second, the distinguished morphemes are assigned part-of-speech tags which denote grammatical function. The reason why the morpheme separation stage precedes POS tagging is to avoid the sparse data problem caused by the multiplicity of morphological variants of the same stem (Han and Palmer, 2005). However, the morpheme-based POS tagger in this process is vulnerable to irregular variations of word stems and, unfortunately, such variants are often found on the web. By the same reason it also produces erroneous results given unspaced texts since the complexity of the decomposing morphemes is very high.

This paper assumes that the morpheme analysis procedure is not feasible to process the SMS texts. In order to alleviate the pain, this research will focus on how one can extract the expected items from the linguistic data with which morpheme analysis does not work.

2.2 Properties of Korean SMS language

Socio-linguistic studies of the Korean SMS language have revealed that the irregular variations within the language are not arbitrarily irregular. The five distinguished properties have been summarized in Table 1 (Park, 2006; Lee, 2010; Kim, 2011).

Some of the properties in Table 1 can be found in English SMS texts as well, hinting that this set of the features may be due to common factors. 'Addition of sounds' is known as epenthesis phenomenon, existing in many languages including English; Crystal (2008) contended that many features of the texting language (logograms, initialisms, pictograms, abbreviations, nonstandard spellings) are not entirely new and have already been in writing systems for centuries.

[1] Abbrebiates: AD(adnominal suffix), NM(nominative particle), IN(instrumental particle), SC(subordinative conjuctive suffix), CP(conjunctive particle), PST(past tense suffix), DC(declarative final suffix), RE(retrospective suffix), CN(conjectural suffix), PR(pronoun), PP(propositive suffix), AC(auxiliary conjunctive suffix), GE (genitive particle), LC(Locative particle)

[2] POS tags: a(adverb), ncs(stative common noun), xpa(adjective-derived suffix), exm(adnominal suffix), nc(common noun), npp(personal pronoun)

Properties	Examples
Ignoring spacing	그녀가학교에갔다. (spaced: '그녀가 학교에 갔다') Ku nyeca-ka hakkyo-ey ka-ss-ta The woman(nyeca)-NM school(hakyo)-LC go-PST-DC 'The woman went to school'
Linking sound or phonetic writing	멋있어 -> 머시써 mes-iss-e 'gorgeous' -> me-si-sse
Reductions or shortenings	메일 -> 멜 meyil 'mail' -> meyl 서울 -> 설 sewul 'Seoul' -> sel
Acronyms or abbreviation	애니메이션 -> 애니 ay-ni-mey-i-syen 'animation' -> ay-ni 비밀번호 -> 비번 pi-mil-pen-ho 'password' -> pi-pen
Addition of sounds	아빠 -> 압빠 a-ppa 'daddy' -> ap-ppa 여보 -> 여봉 ye-po 'honey' -> ye-pong

Table 1. Summarization of properties in Korean SMS text

Ling and Baron (2007) reported that lexical shortening is the one of the most significant characteristics one can see in text messages. However, 'ignoring spacing' is the exception, since Korean suffixes can play as good predictors for the roles or the functions of the preceding stem. As such, removing spaces between phrases does not severely deteriorate the readers' understanding given the content.

This study will focus on only three of the features presented in Table 1: Unspacing, Linking, and lexical reduction. According to linguistic analysis (Park, 2006; Lee, 2010), liaison and vowel reduction were very common among the phonetic variation of the words. Following that observation, this paper will incorporate a set of rules (presented in Park, 2006) in its experiment. Also, it will make use of the Romanization transliteration with the given phonological rules to cope with the lexical variations of the linguistic data.

3 Romanization-based morpheme retrieval process

This section will provide the detailed contents of the lexical variation generation process. Basically, the generation process consists of the three main sub-modules: word-ending addition, vowel-change rules, and vowel omission. Each of these modules contains a set of linguistic rules. As a result, each target word in the list obtains its variants. These variants can then be used to check the input sentence for derived forms of the target word.

3.1 Yale Romanization

Yale Romanization is the transliteration systems developed at Yale University for Romanizing Mandarin, Cantonese, Korean, and Japanese. The Yale system of Korean[3] is generally used in linguistics and is adopted as the application of the transliteration process in this work. There are two other Romanization systems, Revised Romanization of Korean and McCune-Reischauer system, but since the emphasis of the systems is on how to transliterate entire Korean words to a string of elements of a pronounceable alphabet, only Yale Romanization has a one-to-one correspondence between Korean letters and English letters. Therefore, the other two systems are not considered in this study.

3.2 Korean syllable

The Korean alphabet, called Hangul, consists of blocks of multiple letters with each block representing a single syllable. For example, the first word of the Korean word, 한글 (hangul), can be decomposed into three letters ('ㅎ′/'h', 'ㅏ′/'a', and 'ㄴ′/'n′) though it is represented as a single character (or block) in Korean orthography. One advantage of using Yale Romanization is the ability to linearize the Korean syllables into a sequence of the phonemes and thus allowing the linking of alphabets with their sound properties. The examples in Table 1 show this phenomenon

[3] http://search.cpan.org/dist/Encode-Korean/lib/Encode/Korean/Yale.pm

clearly. Although it seems '멋있어'(mes-iss-e) and '머시써'(me-si-sse) have quite different word forms, their romanized forms are identical; implicating that the latter is the phonetic writing version of the former.[4] Morphological analysis has difficulty when analyzing such phonetically written words since it makes distinctions based on Hangul syllables instead of the string of the letters. That is, 'mes-iss-e' and 'me-si-sse' are discriminated because the hyphens are taken as the boundary of the syllables even though this is not the case during pronunciation.

3.3 Implementation of linguistic rules

3.3.1 Conjugation of verbs and adjectives

In Korean grammar, verbs or adjectives do not come as independent morphemes, but always present along with an appropriate conjugation. This paper considers 17 word endings for the romanzied target words, following the standard grammar of Korean (~다 '~ta', ~은 '~un', ~는 '~nun', ~고 '~ko', ~기 '~ki', ~냐 '~nya', ~었다 '~essta', ~았다 '~assta', ~든지 '~tunci', ~던지 '~tenci', ~지 '~ci', ~게 '~key', ~음 '~um', ~ㅁ '~m', ~습니 '~supni', ~읍니 '~upni', ~구 '~kwu'). When the target lexical entry is given with its part-of-speech information, and if it belongs to the categories of noun or adjective, the 17 endings are added to the base word, generating 17 different word forms to be included in the lexicon paradigm set.

3.3.2 Vowel contraction or change

This paper accepted the five vowel variation rules from Park (2006) as follows:

(4) 'o' + 'a' -> 'wa'. e.g., pho-hang ('Pho-hang') -> phwang[5]
(5) 'wu' + 'e' -> 'ye'. e.g., swu-ep ('a class') -> syep
(6) 'wu' + 'i' -> 'wi'. e.g., pwu-in ('wife') -> pwin
(7) 'i' + 'a' -> 'ya'. e.g., ki-an ('draft') -> ky-an
(8) 'i' + 'e' -> 'ye'. e.g., ki-ek ('memory') -> kyek

The rules in (4) ~ (8) are supplied to the 'vowel-change' function that takes the Romanized target word as input and returns its changed form as the output.

3.3.3 Vowel reduction

The vowel reduction rules used in this paper aim to catch two types of shortening; the first type is concerned with the middle syllable of the whole word while the second works on the last syllable. As described in section 3.2, one Hangul syllable consists of several letters and, if the syllable is the target area of the reduction process, the contained vowel may be removed. Therefore, considering the first word of the Korean word, 한글 (hangul), Romanized as 'han', if one omits the vowel ('a') then the result would be 'hngul'.

Previous studies showed that Korean SMS language has frequent vowel reductions (Park, 2006; Lee, 2010; Kim, 2011) with the middle and final syllables being the most common targets for reduction. The example sentence (9) presents the omission of the vowel in the middle syllable and (10) provides an example of reduction in the final syllable.

(9) sa-mwu-sil ('office) -> sam-sil
(10) key-im ('game') -> keym

4 Experiment

Sentiment analysis or opinion mining techniques that utilize retrieval tasks to obtain the training sets or corpus data have to extract subjective chunks or morphemes from the real-world data. In fact, if one chooses to use an annotated subjective word list for the study, one must still go through the process of confirming whether the items in the given list are in the raw input data. For that reason, an effective retrieval operation is required for research which needs to manage unorganized message texts. This section documents two experiments. The first is on the effectiveness of the proposed approach for unspaced tweet texts, while the second focuses on lexical variation.

4.1 Data

A large tweet dataset was obtained from another study (Lee et al., 2011). This dataset contained 5,913,888 tweets from 11,379 users up

[4] It is worth to noting that it becomes easier to apply rewriting rules to the romanized Hangul text because of its' linearity.
[5] Note that the rule of 'H-weak' is manipulated here and the rule functionally works by omitting any 'h' between of sonorants. This rule helps to capture the typical linking sound phenomenon in Korean.

Method \ Condition	Spaced			Unspaced		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Romanization-based method	0.67	**0.79**	0.73	0.67	**0.79**	**0.73**
Morpheme analysis method	**0.94**	0.72	**0.82**	**0.95**	0.29	0.44

Table 2. Results of retrieval test for spacing factor

until the date of 14th Mar 2011. All the Twitter-specific components were filtered beforehand such as Twitter ID, Retweet marker, URL, and hash-tags. To form the list of the target sentiment words, 2823 sentiment word-morphemes, all annotated with their POS tags, were exploited from the previous study of the sentiment analysis on Korean movie reviews (Ko and Shin, 2010).

Since it is needed to construct the test dataset for the first experiment, 100 tweets were randomly selected from the tweet corpus and were manually annotated using the target sets found in the sentiment word list (as a result, 128 items were found in the 100 tweets).

For the second experiment, because no annotated corpus of Korean SMS texts was available, 80 tweets from the corpus were manually collected, each containing at least one irregular word (92 types in total). The varied word in the tweet was marked as the target and its corresponding original entry was restored and recorded in the target lexicon list.

4.2 Experiment 1: Spaced vs. Unspaced

This experiment involved conducting a simple retrieval test for the selected 100 tweets using the sentiment word list as described above. To make a comparison with the proposed approach, the performance of the morphological analysis method also needed to be evaluated. As such, the data was tested using a Korean morphology analyzer. [6]

For the experimental conditions, one factor (spacing) was manipulated, providing two types of test dataset for the different approaches. Since removing all the spaces from the sentences would have left the morphological analyzer inoperable, only the spaces around the target were deleted to create the unspaced condition.

Table 2 shows the results of the retrieval experiment: how well each method found the target items and how many they picked incorrectly. The morpheme analysis-based approach barely chose any wrong targets, but it missed too many right answers (the precision was 27% higher than the precision of Romanization-based method, while marking 7% lower recall rate). Although the morpheme analysis-based approach showed higher performance on the spaced text (0.82 versus 0.73 on F-Measure), the method proved ineffective against unspaced texts (the recall, compared to the Romanization method, was severely decreased from 0.72 to 0.29).

Following expectations, the Romanization-based method was very robust against unspaced texts. This phenomenon is easily explained by considering that the method searched for the target strings without any regard for morpheme boundaries. In contrast, the morpheme analysis-based method took the incoming chunks and separated them into morphemes, but when text is unspaced the morpheme analyzer has to perform word-segmentation as well as morpheme-analysis. Thus one would anticipate an increase in errors when the input text is not properly spaced, because it would increase the complexity of the analysis process.

However, unlike the predictions, the Romanization-based method recorded a lower precision than the morphological analysis-based approach. This result might be due to the set of short-length words in the target list. For example, words consisting of one or two letters such as ‘ak’ (both ‘evil’ or ‘music’ in English) may be erroneously identified in other words such as in ‘ak-ki’ (‘musical instrument’) since such short strings are likely to occur if only by chance. Thus, the Romanization-based method has a higher risk of errors if the system is supplied with such short terms. In the experiment above, the employed sentiment words were morphemes (not phrases or clauses), which is unfavorable for the Romanization approach. However, it is worthwhile to acknowledge that this is mitigated by employing the conjugation module, implying that well-defined rules can enhance performance.

[6] We used the Korean morpheme analyzer distributed from the 21st century Sejong Project (http://www.sejong.or.kr/dist_frame.php).

Model	Precision	Recall	F-Measure
Vowel-reduction+, H weak+, Vowel-change+	0.80	**0.55**	**0.65**
Vowel-reduction+, H weak+, Vowel-change−	0.79	0.52	0.63
Vowel-reduction+, H weak−, Vowel-change+	0.79	0.53	0.63
Vowel-reduction−, H weak+, Vowel-change+	0.96	0.25	0.40
Vowel-reduction−, H weak−, Vowel-change+	0.96	0.23	0.37
Vowel-reduction−, H weak+, Vowel-change−	0.89	0.22	0.36
Vowel-reduction+, H weak−, Vowel-change−	0.78	0.5	0.61
Vowel-reduction−, H weak−, Vowel-change−	**1.0**	0.24	0.38
Morphological analysis-based method	**1.0**	0.067	0.13

Table 3. Results of retrieval tests for phonetically changed words

4.3 Experiment 2: Covering phonetic changes in the lexicon

Experiment 1 dealt with the cases where morpheme's grammatical category information was given, allowing the use of conjugation rule functions. Experiment 2 considers the situation in which specific words or expressions are given without POS tags and with phonetic variations of the targets which must be resolved before its original can be retrieved from the tweet data.

A retrieval experiment was conducted given the test data as described in section 4.1. Unlike Experiment 1, this experiment utilized the sub-modules of the lexical shortening (as stated in section 3.3). The result is displayed in Table 3.

The numbers in bold of Table 3 refer to the highest values for the column (tied values are treated as the same). The conjugation function is not carried out here because of a lack of grammatical category information, thus only three kinds of functions were manipulated as above. While vowel-change rules only care about the replacement of vowels, vowel-reduction rules cope with the circumstances in which the vowels in the word are omitted, resulting in a shortened form. H-weak rule is the only component that relates to any consonant change phenomena in this system; removing the phoneme 'h' between word syllables under specific conditions (e.g., The Korean word, 'coh-a' meaning 'good' is reduced to 'co-a'). The notation [+/-] indicates whether the mentioned function was employed in the construction of the target paradigm set.

As can be seen in Table 3, the full model (including all the three sub-modules) outperforms the other models, proving the research assumption that implementation of linguistic rules would cover a subset of the lexical variations in the SMS language. With capturing the case alone, even the weakest model (with neither vowel-reduction/change nor H-weak functions) showed better results than those of morphological analysis. This is because it could find type-equivalence between tokens such as 'cwuk-um' (죽음, 'death') and 'cwu-kum' (주금, 'death'), obtaining the higher F-score (0.38 vs. 0.13).

Obviously, the strongest module affecting the results is the vowel-reduction function. Remember that this function has two omission rules for the middle and the last syllables of the target items.

The model (with vowel-reduction off and the other two functions on) clearly reveals the effect of this sub-module by exhibiting a rapid drop in F-score from 0.65 for the full-model to 0.40 for the current model.

This effect is due to the high frequency of the vowel-reduction variations. Table 4 summarizes the types of variation in the test data, providing an explanation for the results in Table 3. The proportion of phoneme reduction instances can be seen to be about a third of the total occurrences (36 out of 104, or approximately 35 percent), and it accounts for the steep decrease in F-score when the vowel-reduction function is not adopted. It is also worth noting that vowel-reduction in the first-syllable is quite rare; consistent with the linguistic analysis of empirical research (Park, 2006; p. 466). The creation of vowel-reduced forms clearly had a large effect, lowering the accuracy from 0.96 to 0.80. This is because the shortened targets can also be found as sub-string of bigger words. However, this shortcoming does not weaken the efficiency of the whole approach. The morphological analysis-based retrieval method found only a few items in the data, which was expected considering that this analysis is dependent on a syllable-based word lexicon.

In short, though a small set of the linguistic rules were employed, and even using them is still far from achieving complete coverage, the results of the experiment implicate that such a rule-based system can capture at least part of the vast, complicated range of linguistic variations.

Type	Specified type		Count
Linking Sound			8
Phoneme Reduction	Vowel reduction	Head-syllable vowel reduction	1
		Middle-syllable vowel reduction	17
		Final-syllable vowel reduction	14
		Others	4
	Consonant reduction	H-weak	9
		Others	5
Phoneme Change	Vowel change		22
	Consonant change		11
Abbreviation			5
Addition	Vowel addition		6
	Consonant addition		2
Total			104

Table 4. Types and counts of instances in test dataset of Exp. 2

5 Discussion and Conclusion

This paper confirmed that employing language-specific rules to handle SMS language text can enhance the results of the retrieval process. Although it is known that morphological analysis hardly produces erroneous results in formally written texts such as newspaper articles, the analysis results were made much worse for the SMS data in our experiments, which presented the motivation to pursue an additional approach. The procedure of sentiment analysis or opinion mining generally involves searching for items which are defined as subjectively meaningful, but typical morphological analysis cannot deal with the irregular changes of the web texts.

The reason why the morphological analysis does not work on such data is clear. The built-in stemmer or normalization process of the analyzer is not designed to cope with that kind of the text. However, in this paper, we tried to point out that judging the text as not well-formed enough to be processed is too quick. Instead, a set of generative rules to handle such texts were proposed and implemented in our experiments. Although those rules could be imported to a future morphological analyzer giving it broader coverage, suffice it to state that the text on the internet is not as simple as newspaper articles to the analyzers currently available.

For such a case, this proposed method could be an alternative way to preprocess Korean SMS texts and it should be noted that there could be similar approaches for other morphologically rich languages like Japanese or Turkish. Normalizing text is a very complicated task for the type of the languages and well-organized module would be needed if it has to manipulate SMS texts for any morpheme-level retrieval process.

A Romanization transliteration scheme is used in this study because it naturally represents the phonetic properties of Korean syllables while providing a more intuitive way to apply a set of defined rules to the sequence. Since phonemic variation is quite common in SMS texts, as mentioned, this approach seems useful and practical regarding the results of the experiments. Although the size of the dataset which was used for the test is small, the sample set contained cases which were well known in previous literature and their linguistic patterns were consistent with reports (Park, 2006; Lee, 2010; Kim, 2011). However, to make the approach practical enough to be used by field engineers, a large scale corpus would be required to find the optimal set of the transformation rules, which is left for future study due to the lack of such annotated data at the time of writing.

Acknowledgments

We would like to thank Lee, W., Cha, M. and Yang, H. for their kind approval to use the Tweet corpus and the three anonymous reviewers for their helpful comments.

References

Chen, H.-H., and Ku, L.-W. (2002). An NLP & IR approach to topic detection Topic detection and tracking (pp. 243-264): Kluwer Academic Publishers.

Crystal, D. (2008). Txtng: The Gr8 Db8, Oxford University Press.

Fujii, A., and Ishikawa, T. (2001). Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4), 389-420.

Han, N. R. (2006). Klex: A finite-state transducer lexicon of Korean. In *Finite-State Methods and Natural Language Processing* (pp. 67-77). Springer Berlin Heidelberg.

Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA.

Jang, H., and Shin, H. (2010). Language-specific sentiment analysis in morphologically rich languages. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China.

Kim, S. (2011). Phonological and Morphological Characters of Junmal in Korean Net Lingo. *Linguistics. 61*, 115-129. In Korean.

Kim, S.-M., and Hovy, E. (2004). Determining the sentiment of opinions. Paper presented at the Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland.

Ko, M., and Shin, H. (2010). Grading System of Movie Review through the Use of An Appraisal Dictionary and Computation of Semantic Segments. *Korean Journal of Cognitive Science. 21(4),* 669-696. In Korean.

Lee, J. (2010). A Study of Phonological Features and Orthography in Computer Mediated Language. *Linguistic Research, 27(1)*, 1-18. In Korean.

Lee, W., Cha, M., Yang, H. (2011). Network Properties of Social Media Influentials : Focusing on the Korean Twitter Community. *Journal of Communication Research. 48(2)*, 44-79. In Korean.

Ling, R., and Baron, N.S. (2007). Text Messaging and IM: Linguistic Comparasion of American College Data. *Journal of Language and Social Psychology, 26(3)*, 291-298, doi:10.1177/0261927X06303480

Pak, A., and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Paper presented at the Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Park, C. (2006). A Phonological Study of PC Communication Language Noun. *Korean Education, 119*, 457-486. In Korean.

Wiebe, J. M. (2000, July 30–August 3). Learning subjective adjectives from corpora. Paper presented at the In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, TX.