International Journal of Database Theory and Application
Vol.10, No.7 (2017), pp.13-26
http://dx.doi.org/10.14257/ijdta.2017.10.7.02

A Cross-Domain Analysis using Morphological Sentence Pattern Approach for Extracting Aspect-based Lexicon

Youngsub Han[1], Yanggon Kim[2] and Jin-Hee Song[*]

[1,2]Department of Computer and Information Sciences, Towson University, 7800 York Road, Towson, MD, USA
[*]School of IT Convergence Engineering, Shinhan University, South Korea
*[1]yhan3@students.towson.edu, [2]ykim@towson.edu, *jhsong@shinhan.ac.kr*

Abstract

Social media data contains people's emotions, opinions and experiences. Sentiment analysis aims to analyze the data to observe meaningful information. However, building lexicon for a lexicon based sentiment analysis is one of the biggest challenges without human-coding efforts. In this study, we proposed a cross-domain approach for building sentimental lexicon using the morphological sentence patterns for analyzing social media data. Our approach shows relatively higher F-score (79.64) than existing approaches. In addition, this approach can be used for multi-source data such as online reviews and social media data without human-coded knowledge bases.

Keywords*: Data mining; Aspect-based Lexicon Building; Social Media; cross-domain analysis*

1. Introduction

In recent years, social media is one of the most popular online media such as Twitter, Facebook, Instagram and YouTube for sharing and communicating news, promotions, advertisements and emotions [1]. The information contains people's opinions about brand equity and value with unpleasant or dissatisfied experiences. Various industries have spared no efforts to build advantages using social media because it allows to reach target audiences efficiently. The data contains sufficient information to understand trends, issues, individuals, human behavior, and identifying influential people [2]. Sentiment analysis aims to analyze the textual data. It helps to observe and summarize people's opinions or emotional states. Despite the demands of sentiment analysis methods for analyzing social media data, fundamental challenges still remain, because user-generated online textual data is unstructured, unlabeled, and noisy to be analyzed accurately. Especially, building lexicon usually needs human-coding efforts because the lexicon affects a quality of analysis in the lexicon based sentiment analysis approach [3-5].

Accordingly, we proposed a morphological sentence patterns model in our previous research [6]. In the research, we suggested some manners for extracting aspects and expressions considering efficiency and accuracy. This paper showed relatively higher F-score (82.81) than existing researches. In addition, this model considers not only individual word but also phrases. This function is important to build lexicon because a phrase represents better the feature term of the document than an individual word [7-8]. However, the research focused only movie reviews despite demands of social media analysis. Therefore, we applied the model for extracting aspects and expressions to be used for aspect-based analysis considering characteristics of social media. Also, we examined how the morphological sentence patterns work across multi-domain which are movie reviews, YouTube and Twitter. We expect that this system can help to minimize human-coding efforts to building sentiment lexicon from social media data.

*Corresponding Author

2. Related Works

2.1. Lexicon Building on Aspect-based Sentiment Analysis

The purpose of sentiment analysis is extracting opinions or emotional states regarding certain topics such as events, products, entertainers, politicians and movies from the textual data to find people's interests and thoughts [3-5]. Especially, aspect-based sentiment analysis is one of advanced approach of lexicon based sentiment analysis. This approach is broadly used and it can be in-depth analysis based on sentimental lexicon. In this approach, the results are categorized into each aspect with one or more expressions. For example, when an object is a "mobile phone", aspects are "display", "size", "price", "camera", or "battery". In this case, aspects seem attributes of the object to describe more detail. Thus, expected result are pairs of aspects and expressions such as "display-clean", "price-good", or "camera-awesome" [9, 10]. In this approach, the building lexicon is a fundamental challenge because the lexicon is used as measurement. A lot of researchers proposed unsupervised or semi-supervised approaches for building lexicon [11-13]. Especially, J. Bross, and H. Ehrig proposed a method for automatically adapting and extending lexicons to a specific product domain, but they simply used morphological patterns to extract aspect-based lexicon [14]. Therefore, we proposed a method for extracting aspect-based lexicon builder using morphological sentence patterns to analyze movie reviews in our previous research [6].

2.2. Morphological Sentence Pattern Model

Table 1. Examples of Extracted Aspects and Expressions

Rank	Aspect	Expression	Count
1	MOVIE	GOOD	149
2	PARK	OPEN	100
3	MOVIE	GREAT	47
4	ACTING	GOOD	47
5	MOVIE	PREDICTABLE	43
6	CGI	GOOD	37
7	CHARACTER	LIKABLE	35
8	DINOSAUR	GREAT	34
9	FILM	GOOD	30
10	MOVIE	AWESOME	28

The morphological sentence pattern model (MSP Model) was designed to extract aspects and expressions from online movie reviews [6]. In this model, the pattern recognizer generates morphological sentence patterns based on part-of-speech tags using a natural language processing tool, which is "Stanford CoreNLP" made by The Stanford Natural Language Processing Group. This tool provides refined results from textual data based on English grammar such as the base forms of words, the part-of-speech (POS) [15]. Then, the extractor retrieves aspects and expressions using the patterns based on what patterns are surrounding aspects or expressions based on each sentence. For diversity of extraction, the system considered the N-gram model for matching the patterns. N-gram is defined as a contiguous sequence of n items from a given sequence of words or speeches. This model is widely used for text based analysis [16]. In addition, when matching the pattern, the longest pattern has a more priority to avoid duplicate extraction. This strategy helps to less computation time. The table 1 shows examples of aspects and expressions. This model guarantees relatively higher accuracy (F-score, 82.81) than existing approaches. In this model, they suggested 3 to 7 lengths patterns for extracting aspects and 2 to 6 lengths patterns for extracting expressions, and more frequently occurred pair of aspect-expressions applied to improving efficiency and accurately. This model shows relatively higher F-score than existing researches.

3. Implementation

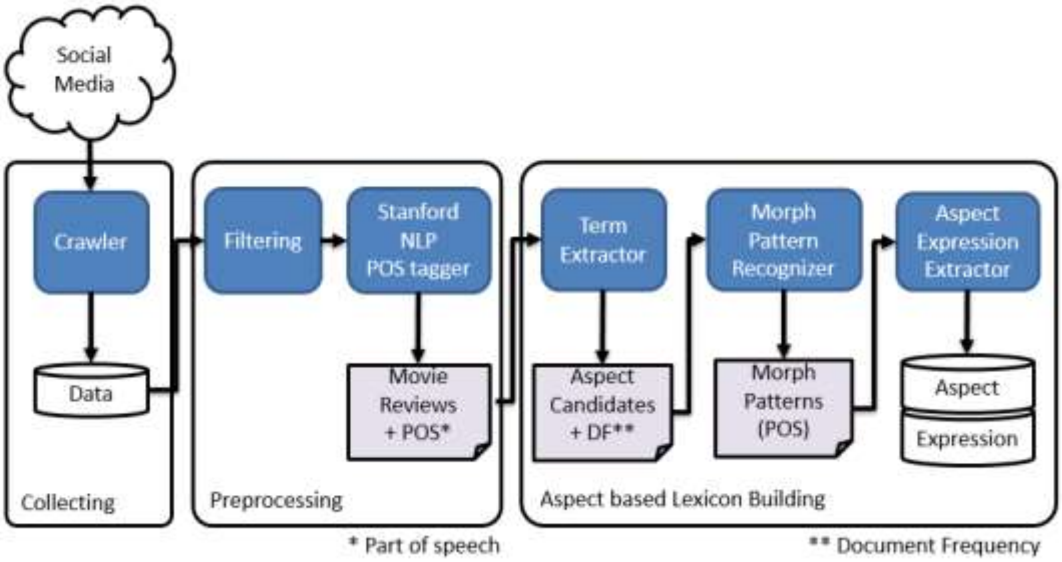

Figure 1. System Architecture and Flow

We proposed an advanced model based on the MSP model to analyze social media data which are YouTube comments and Twitter tweets because the MSP model considered only movie reviews despite demands of social media analysis. The system consists of three main parts which are data collecting, preprocessing, and aspect-based lexicon building. In the first phase, the crawler collects data from Twitter, YouTube, and movie reviews from IMDB, Rotten Tomatoes, and Metacritic. The crawler collects Twitter tweets and YouTube comments using APIs provided by Twitter and YouTube [17, 18]. The crawler collects movie reviews using a HTML parser from the online review sites. The data is a bunch of Tweets, YouTube comments, and movie reviews as documents. In the second phase, the system refines considering characteristic of social media data. Then, the system analyzes sentence parsing and part-of-speech tagging from collected documents using “Stanford core NLP” [14]. In the third phase, the system extracts aspect and expression candidates based on their ranking by the document frequency. In this case, the higher document frequency would be commonly used aspects and expressions in corpus because frequency of words may not necessarily expound topics of documents despite the frequency usually recognized as the features of textual data [19, 20]. Then, the morph pattern recognizer extracts patterns based on part-of-speech tags. Then, the aspect-expression extractor retrieves aspects and expressions using the patterns. Though this research, we suggested how to apply this model to maintain the accuracy (F-score) of the MSP model with minimizing human-coding efforts.

3.1. Data Collection

To collect tweets from twitter, we used a twitter collecting tool which was developed by Y. Han et al in our previous research [17]. The crawler retrieves tweets by keywords related to the target objects such as companies, products, politicians or movies with scheduled time. The crawler requests tweets with the keywords using Twitter search API with application key provided from Twitter, then Twitter gives tweets including the keyword. All keywords are stored in our database with the object information. Twitter provides the latest 9 days of tweets by each keyword from current time. The crawler collects tweets repeatedly because Twitter allows 180 requests per key in 15 minutes, and a request includes 100 tweets.

To collect YouTube comments, we used a YouTube collecting tool which was developed by Lee et al [18]. YouTube provides APIs to collect data such as video information, user profiles, and comments written by users. The crawler collects comments posted on movies that retrieve by keywords related to the target objects such as companies, products, politicians or movies. It also collects the data repeatedly within scheduled time based on user requests.

The crawler collects movie reviews with ratings of the review generated by users through Rotten Tomatoes, IMDB, and Metacritic. To collect the data, the jsoup HTML parser, an open-source Java library of methods, is designed to extract and manipulate data stored in HTML documents developed by Jonathan Hedley[1]. It automatically collects reviews using movie names as seeds such as "Jurassic World", "Avengers: Age of Ultron". There are some differences of the rating scales by sites. In the case of Rotten Tomatoes, a writer indicates their opinions weather "Fresh", or "Rotten". The Fresh means a positive and the Rotten means a negative. In the case of IMDB and the Metacritic, writer indicates their opinions from 1 to 10. The bigger number means more positive. We decided 8 to 10 are positive opinions and 1 to 3 are negative opinions to calculate positivity of expressions.

3.2. Preprocessing

To extract aspects and expressions from YouTube and Twitter accurately, we developed a preprocessing module. The module filters out when a sentence consists of less than one word such as a noun, an adjective, a verb because this sentence is a less meaningful sentence in terms of the linguistic approach. For example, when a sentence consists of not meaningful words such as special characters, prepositions or conjunctions, the sentence is not useful. Additionally, some reserved words such as hash tags (#), accounts (@) or URL formats (http://) cause errors to match morphological sentence patterns. Even though, these words contain meaningful information for the service, we decided to transform and filter out the words. Then, the system analyzes the data using "Stanford Core NLP" made by The Stanford Natural Language Processing Group [15]. This tool provides refined and sophisticated results from textual data based on English grammar such as the base forms of words, the parts of speech (POS), and the structure of sentences. In this research, we used the tool for two main reasons. The first reason is that the some online textual data contain linguistic problems such as spacing errors, idioms, and jargons. Another reason is that the system mainly uses part-of-speech information to extracts and match morphological patterns to build lexicon.

3.3. Cross-domain Analysis of Morphological Sentence Pattern Model

As we mentioned in the introduction section, we examined how the morphological sentence patterns work across domains which are movie reviews, YouTube and Twitter. This cross-domain analysis means that the system extracts aspects and expressions using patterns of each data source across other sources data. For example, the system extracts aspects and expressions from the YouTube comments and Twitter tweets using movie review patterns as shown in Figure 2. Through this comparative cross-domain analysis, we discovered how the patterns are applicable for other source data. All the results automatically calculated accuracy with the F-measure (see section 3.3).

[1] Jonathan Hedley, Jsoup HTML parser, http://jsoup.org/

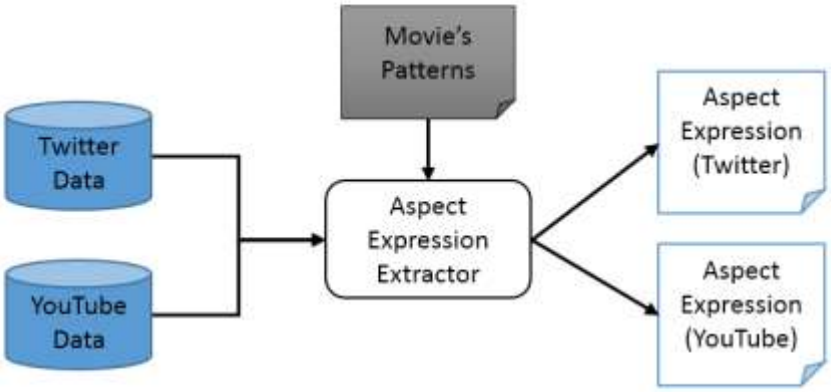
Movie's Patterns
Twitter Data
YouTube Data
Aspect Expression Extractor
Aspect Expression (Twitter)
Aspect Expression (YouTube)

Figure 2. An Example of Cross-domain Analysis

3.4. Measurement

To evaluate the quality of our system, we calculated the F-measure which is broadly used to measure the performance for this type of systems [19]. The definition of the measure is:

$$\mathrm{F-measure} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (1)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (2)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (3)$$

The F-measure considers the “Recall” and “Precision” (1). The recall means the portion of relevant instances that are retrieved (2), and the precision means the portion of retrieved instances that are relevant (3). Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. For example, when an extracted aspect or expression is classified as ‘correct’ by the extractor while the aspect or expression is labelled by answer-set which is labeled by human annotators is ‘related’, the aspect or expression is considered TP (true positive). On the other hand, when an extracted aspect or expression is classified as ‘correct’ by the extractor while the aspect or expression is labelled by answer-set which is labeled by human annotators is ‘non-related’, then the aspect or expression is considered TN (true negative). Also, we compered quality of our system with existing related researches based on this measure.

4. Experiment

For the experiments, we collected 1,000 YouTube comments and 1,000 tweets using a movie title, ‘Jurassic World’ as a seed. Then, we selected 1,000 sentences for each source to compare all proposed method because the extractor retrieves based on each sentence. For cross-domain analysis, we referenced a result of movie reviews which proposed in our previous research [6].

4.1. Pattern Selection

To find optimized patterns, we examined which lengths patterns can extract most numbers of correct aspects and expressions. The pattern recognizer generated 71,178 patterns for aspects and 49,904 patterns for expressions from YouTube comments, and 47,606 patterns for aspects and 36,002 patterns for expressions from Twitter tweets.

Figure 3 and 4 shows how many correct aspects and expressions can be extracted by each length of patterns for YouTube. As shown in the figures, we used the average number of correct aspects (17.9) and expressions (9.5) from all lengths of patterns as a threshold to select patterns. From all identified result of correct aspects and expressions by the lengths of patterns, 1 to 5 lengths patterns are over the threshold and these patterns could extract 96.7% (347 out of 358) of correct aspects, and 1 to 4 lengths patterns are over the threshold and these patterns could extract 91.5% (173 out of 189) of correct expressions from YouTube.

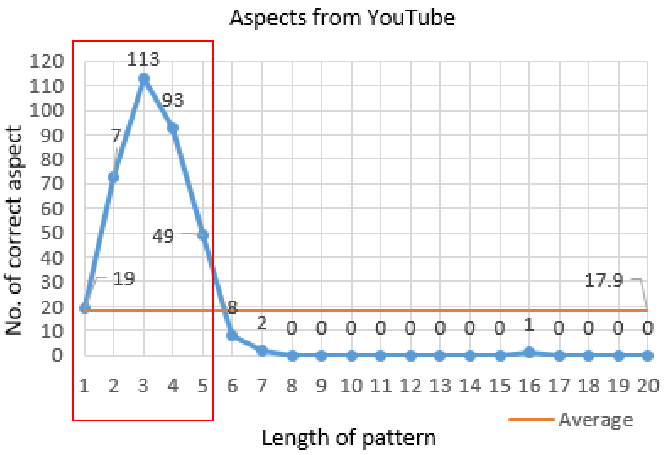

Figure 3. The Numbers of Correct Aspects by Pattern Lengths for YouTube

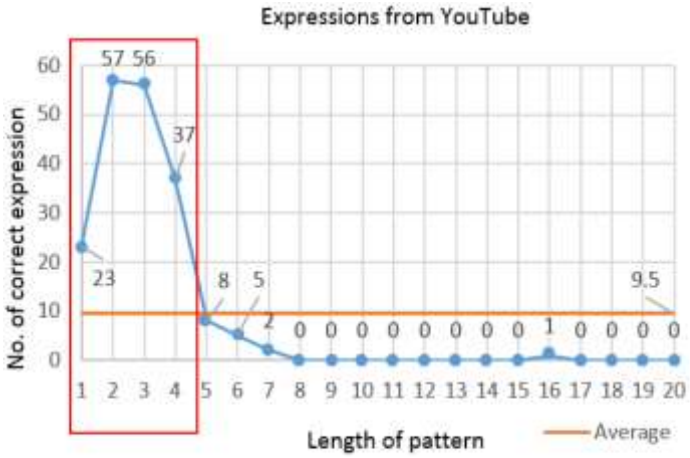

Figure 4. The Numbers of Correct Expressions by Pattern Lengths for YouTube

As shown in the Figure 5 and 6, we also used the average number of correct aspects (18.2) and expressions (16.5) as a threshold to select patterns for tweets. In this case, 2 to 6 lengths patterns could extract 89.5% (325 out of 363) of correct aspects and 2 to 5 lengths patterns could extract 80.84% (173 out of 214) of correct expressions for tweets. Therefore, we selected these lengths of patterns to extract aspects and expressions and we named these patterns as ‘Selected Pattern’.

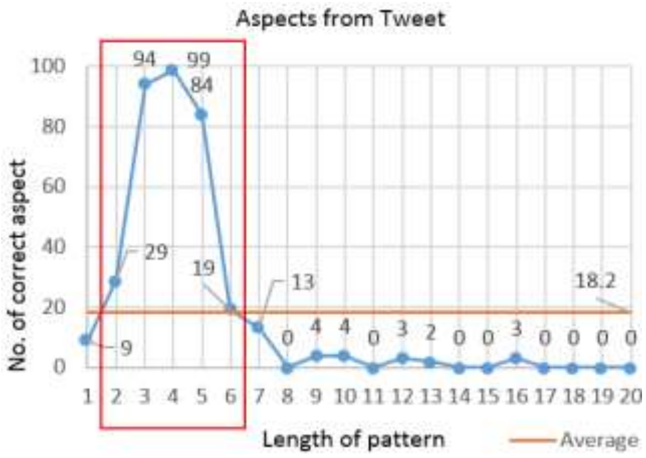

Figure 5. The Numbers of Correct Aspects by Pattern Lengths for Twitter

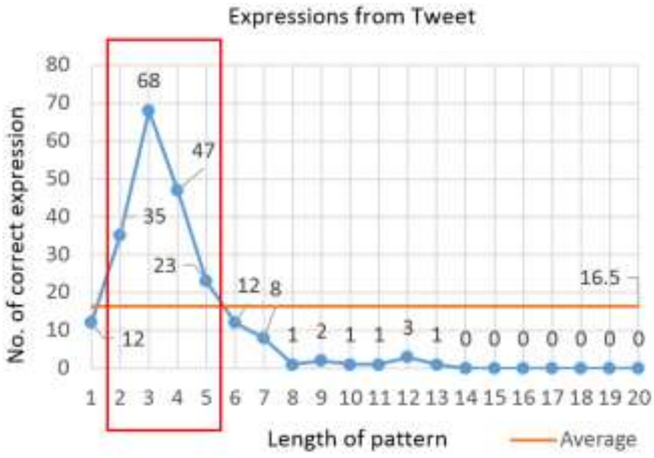

Figure 6. The Numbers of Correct Expressions by Pattern Lengths for Twitter

In addition, we used two more methods which are named “LF”. “LF” means longest-first matching. This method aims to avoid duplications of aspects and expressions because all generated patterns are expended from original patterns. If the extractor matches both generated patterns and original patterns with same target sentence, the data duplication may occur. Therefore, we decided to use “LF” method for experiments.

4.2. Performance Test

Figure 7 shows the results of extracting aspects and expressions from YouTube comments by methods. When we used selected patterns for extracting aspects from YouTube, the processing time (12,014 patterns used, 298 seconds spend) is about 6 times faster than all patterns used (71,178 patterns used, 1,801 seconds spend). Also, when we used selected the patterns for extracting expressions from YouTube (5,852 patterns used, 147 seconds spend), the processing time is about 9 times faster than all patterns used (49,915 patterns used, 1,279 seconds spend).

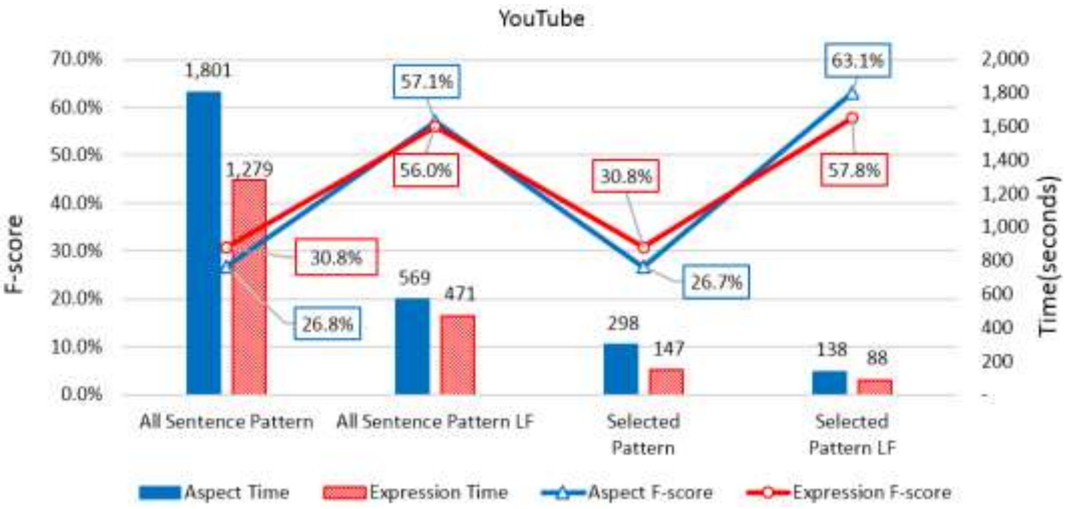

Figure 6. The Results of Extracting Aspects and Expressions by Methods for YouTube

Figure 8 shows the results of extracting aspects and expressions from Tweets by methods. When we used selected patterns for extracting aspects from tweets, the processing time (14,893 patterns used, 378 seconds spend) is about 2 times faster than all patterns used (47,606 patterns used, 872 seconds spend). And, when we used selected patterns for extracting expressions from tweets (8,627 patterns used, 221 seconds spend), the processing time is about 4 times faster than all patterns used (36,002 patterns used, 875 seconds spend).

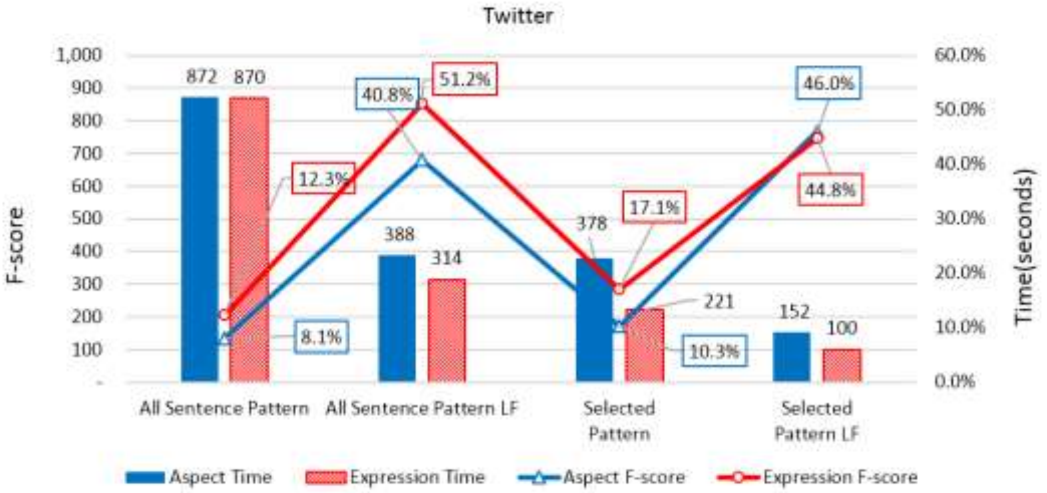

Figure 7. The Results of Extracting Aspects and Expressions by Methods for Twitter

Through these results, we found that the ‘Selected Pattern LF’ shows the highest F-score and the shortest processing time for both YouTube comments and Twitter tweets. It means that this method mostly affects both accuracy and processing time while “Selected Pattern” affects only the processing time. Therefore, we decided to use “Selected Pattern LF” for experiments.

4.3. Improvement

For further improving F-score, we used the frequency and co-occurrence of aspects and expressions. Firstly, the system used the frequency of aspects and expressions as a threshold which is named as "frequency > 1". After extracting aspects and expressions, the system filters out certain aspects or expressions when its frequency is one. As shown in the Figure 9, "Selected Pattern (frequency > 1)" method shows higher F-score (77.33% for aspects and 76.6% for expressions) than "Selected Pattern LF" (63.01% for aspects and 57.75% for expressions) for YouTube. Also, "Selected Pattern (frequency > 1)" method shows higher F-score (71.2% for aspects and 53.4% for expressions) than "Selected Pattern LF" (46.00% for aspects and 44.78% for expressions) for tweets.

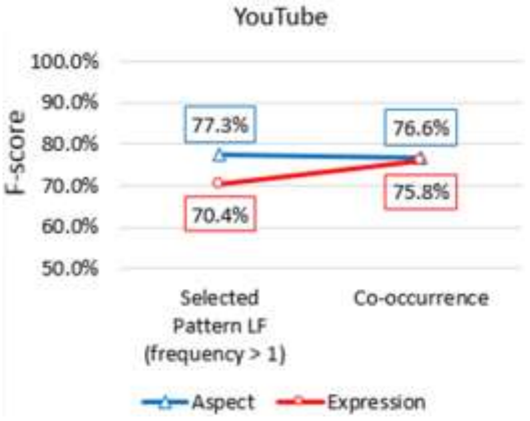

Figure 8. The Results of Extracting Aspects and Expressions with Improving Methods for YouTube

In the case of the "co-occurrence" method, the system retrieves pairs of aspects and expressions after all aspects and expressions extracted by "Selected Pattern LF" method when these pair is occurred in a sentence. As shown in the Figure 10, "Co-occurrence" method shows higher F-score (76.6% for aspect and 75.8% for expressions) than "Selected Pattern LF" method for YouTube. Also, "Co-occurrence" method shows higher F-score (77.3% for aspect and 73.8% for expressions) than "Selected Pattern LF" method for tweets. This finding suggests that frequency and co-occurrence affects the F-score for extracting aspects and expressions.

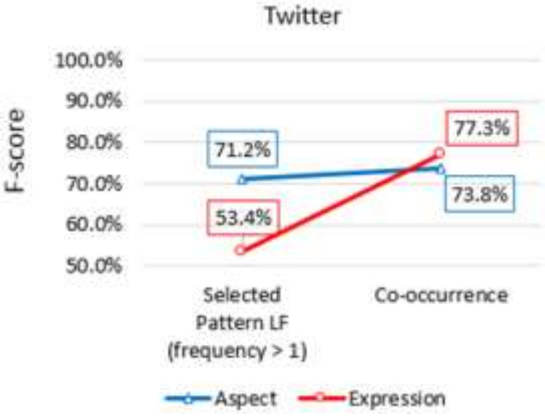

Figure 9. The Results of Extracting Aspects and Expressions with Improving Methods for Twitter

4.4. Cross-domain Analysis

Table 2. Data Sample for Experiments

Sources / Category	Movie Review	YouTube Comment	Twitter Tweet
No. of Used Sentences	1,000	1,000	1,000
No. of Documents	238	911	715
Sentences per Documents	4.2	1.1	1.4
Avg. POS per Sentences	21.1	13.3	11.3
No. of aspect*	230	283	96
No. of Expression*	250	341	154

* Human coded answer-set

The table 2 shows all used data samples for the experiments. We collected 1,000 documents for comparison between movie reviews, YouTube comments and tweets related a movie, "Jurassic World" using our crawler, and then we selected 1,000 sentences because of a fair comparison. It means that our system extracts aspects and expressions based on sentences. However, we used same amount of data, the sentences per documents are totally different between sources. In the case of a movie reviews contains about 3 times more (4.2 sentences in a document) than YouTube comments (1.1 sentences in a document) and Tweets (1.4 sentences in a document). Also, the average number of POS (part of speech) per sentences of movie reviews is about 2 times more (21.1 sentences in a document) than YouTube comments (13.1) and Tweets (11.3). This result shows the movie reviews consist of more complicated and longer sentences than others. In addition, we built answer-set that includes aspects and expressions extracted by human code to verify how accurately this system works. From all identified answer aspects and experiments, the number of answer aspects and expressions of the YouTube comments contain highest numbers of aspects and expressions. It implies that the people express opinions more variety through YouTube, while less complicated.

Table 3. The Results of Cross Domain Analysis for Extracting Aspects

Data	Pattern	Extracted Aspect	Correct Aspect	Answer	Precision	Recall	F-Score
Movie	Movie	**236**	**205**	**230**	**86.86%**	**89.13%**	**87.98%**
Movie	Twitter	238	70	230	29.41%	30.43%	29.91%
Movie	YouTube	296	67	230	22.64%	29.13%	25.48%
Twitter	Twitter	**126**	**79**	**96**	**62.70%**	**82.29%**	**71.17%**
Twitter	Movie	223	31	96	13.90%	32.29%	19.44%
Twitter	YouTube	255	33	96	12.94%	34.38%	18.80%
YouTube	YouTube	**309**	**231**	**283**	**74.76%**	**81.63%**	**78.04%**
YouTube	Movie	198	76	283	38.38%	26.86%	31.60%
YouTube	Twitter	168	66	283	39.29%	23.32%	29.27%

Table 3 and table 4 show results of extracting aspects and expressions with cross-domain analysis. From all identified results, when we used same source of data and patterns, the F-score shown significantly higher score than the other patterns were used. We assumed that the result of YouTube and Twitter could give similar F-scores when we used the patterns of YouTube and Twitter are used each other because these are social media and having similar characteristics in terms of the length and number of POSs of a sentence. However, the results shown the similar F-score as other cross-domain analysis such as between patterns of movie review with YouTube comments (25.48%) and patterns of YouTube comments with movie reviews (23.72%). This result implies that people express their opinion depending on each media.

Table 4. The Results of Cross Domain Analysis for Extracting Expressions

Data	Pattern	Extracted Expr	Correct Expr	Answer	Precision	Recall	F-Score
Movie	Movie	**283**	**206**	**250**	**72.79%**	**82.40%**	**77.30%**
Movie	Twitter	351	47	250	13.39%	18.80%	15.64%
Movie	YouTube	313	43	250	13.74%	17.20%	15.28%
Twitter	Twitter	**100**	**130**	**154**	**76.92%**	**64.94%**	**70.42%**
Twitter	YouTube	244	41	154	16.80%	26.62%	20.60%
Twitter	Movie	146	30	154	20.55%	19.48%	20.00%
YouTube	YouTube	**424**	**290**	**341**	**68.40%**	**85.04%**	**75.82%**
YouTube	Movie	207	65	341	31.40%	19.06%	23.72%
YouTube	Twitter	258	65	341	25.19%	19.06%	21.70%

4.5. Comparison with related approaches

To compare our approach with related approaches, we selected Hu and Liu [21], HashtagLex [22], Sentiment140Lex [22], and TS-Lex [23]. Hu and Liu is a traditional model with a relative small lexicon. HashtagLex, Sentiment140Lex and TS-Lex are sentiment lexicons for Twitter. MSP Model is morphological sentence pattern model for movie reviews [6]. As shown in the Table 5, our approach shows relatively higher F-score (79.64) than other approaches except MSP Model. However the model supports only movie reviews. Therefore, we suggest our model for building aspect-based sentiment lexicon for social media analysis.

Table 5. Comparison of F-score with Related Researches

Methods	F-Score	Multi Source	Social Media
HL[21]	60.49	No	No
HashtagLex [22]	65.30	No	Twitter
Sentiment140Lex [22]	72.51	No	Twitter
TS-Lex [23]	78.07	No	Twitter
MSP Model[6]	82.81	No	No
Proposed Model	**79.64**	**Yes**	**Movie, Twitter, YouTube**

5. Conclusion

In this paper, we proposed a model for building aspect-based sentiment lexicon using morphological sentence patterns. This model was designed to analyze multi-source online data including the social media data considering its characteristics. Through our experiments, we found 3 main characteristics to use this model. The first characteristic is the length of pattern. When we used certain lengths of patterns (see section 4.1.), our model spend 14 times less processing time with 2 times higher F-score than all extracted patterns used for YouTube comments and 7 times less processing time with 4 times higher F-score than all extracted patterns used for Tweets. The Second characteristic is the frequency of extracted aspects and expressions. In that case, more frequently occurred aspects and expressions tend to more accurate. The third characteristic is the more frequently co-occurred aspects and expressions tend to more accurate. Therefore we suggested thresholds considering these characteristics. In addition, we examined cross-domain analysis using YouTube comments and Twitter Tweets and Movie reviews. This examination show how the morphological sentence patterns works across other source data. Through this experiment, we discovered that the sentence consists of different structures in different sources. It implies that people share their opinions and emotions differently depending on the sources in terms of sentence structures. Therefore, we suggested that the patterns should be used for own source data to maintain the F-score. Our model shows relatively higher F-score (79.64) than existing approaches and this model can be used for multi-source data including social media data without any human-coded knowledge bases. Our future work is generalizing the patterns to be used across sources with suggested F-scores.

References

[1] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", Proceedings of the International AAAI Conference on Weblogs and Social Media, **(2010)**, pp. 122-129.

[2] A. M. Kaplan and M. Heinlein. "Users of the world, unite! The challenges and opportunities of social media", Business Horizons, vol. 53, **(2010)**, pp. 59-68.

[3] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis", Proceedings of the 2012 ACM Research in Applied Computation Symposium, **(2012)**, pp. 1-7.

[4] P. Goncalves, M. Araújo, F. Benevenuto and M. Cha, "Comparing and combining sentiment analysis methods", Proceedings of the first ACM conference on online social networks, **(2013)**, pp. 27-38.

[5] P. Melville, W. Gryc and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2009)**, pp. 1275-1284.

[6] Y. Han, Y. Kim and I. Jang, "A Method for Extracting Lexicon for Sentiment Analysis based on Morphological Sentence Patterns", Studies in Computational Intelligence (SCI), Springer, Germany, vol. 654, **(2016)**, pp. 85-101.

[7] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", Proceedings of HLT-EMNLP, Stroudsburg, PA, USA, **(2005)**, pp. 347-354.

[8] P. Niyigena, Z. Zuping, M. Khuhro and D. Hanyurwimfura, "Efficient Document Similarity Detection Using Weighted Phrase Indexing", International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 5, **(2016)**, pp. 231-244.

[9] T. T. Thet, J. C. Na, and C. S. G. Khoo. "Aspect-based sentiment analysis of movie reviews on discussion boards", Journal of Information Science, vol. 36, **(2010)**, pp. 823-848.

[10] F. Wogenstein, J. Drescher, D. Reinel, S. Rill and J. Scheidt , "Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach", WISDOM '13, Chicago, USA, no. 5, **(2013)** August.

[11] N. Kaji and M. Kitsuregawa. "Building lexicon for sentiment analysis from massive collection of html documents", In Proceedings of EMNLP-CoNLL, **(2007)**, pp. 1075–1083.

[12] Z. Zhang and M. P. Singh., "Renew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis", In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, **(2014)**, pp. 542–551.

[13] Yen-Jen Tai and Hung-Yu Kao, "Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation", Proceedings of International Conference on Information Integration and Web-based Applications & Services, **(2013)**, pp. 53-63.

[14] J. Bross and H. Ehrig, "Automatic Construction of Domain and Aspect Specific Sentiment Lexicons for Customer Review Mining", CIKM'13, San Francisco, CA, USA, **(2013)**, pp. 1077-1086.

[15] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, **(2014)**, pp. 55-60.

[16] A. Tomovic, P. Janicic and V. Keselj, "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences", Journal of computer methods and programs in biomedicine, vol. 81, **(2006)**, pp. 137–153.

[17] Y. Han, H, Lee and K, Kim, "A Real-time Knowledge Extracting System from Social Big Data using Distributed Architecture", Proceedings of the 2015 Research in Adaptive and Convergent Systems, **(2015)**, pp. 74-79.

[18] H. Lee, Y. Han, Y, Kim and K, Kim, "Sentiment Analysis on Online Social Network Using Probability Model", In Proceedings of the Sixth International Conference on Advances in Future Internet, **(2014)**, pp. 14-19.

[19] W. Feng, “Research of Theme Statement Extraction for Chinese Literature Based on Lexical Chain”, International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 6, **(2016)**, pp. 379-388.

[20] D. Li1, X. Jin1 and L. Cui, “Text Recognition Algorithm Based on Text Features”, International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 5, **(2016)**, pp. 209-220.

[21] M. Hu and B. Liu. "Mining and summarizing customer reviews", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, **(2004)**, pp. 168-177.

[22] S.M. Mohammad and P. D. Turney, “Crowdsourcing a word–emotion association lexicon”, Computational Intelligence, vol. 29, no. 3, **(2012)**, pp. 436-465.

[23] D. Tang, F. Wei, B. Qin, M. Zhou and T. Liu, “Building Large-Scale Twitter-Specific Sentiment Lexicon”, Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, **(2014)**, pp.172–182.

Authors

Youngsub Han, He is currently a doctoral candidate in the Department of Computer and Information Sciences, Towson University. His current research interests include Big Data and data mining in social networks.

Yanggon Kim, He received his B.S. and M.S. degree from Seoul National University, Seoul, Korea in 1984 and 1986, respectively and Ph.D. degree in Computer Science from Pennsylvania State University, Pennsylvania, 1995. He is currently a professor in the Department of Computer and Information Sciences, Towson University and is a director of MS in CS program. His current research interests include Computer Network, Secure BGP network, distributed computing systems, Big Data and data mining in social networks.

Jin-hee Song, She received B.S. degree in computer science from Seoul National University of Science & Technology, South Korea, M.S. degree in computer science from Hankuk University of Foreign Studies, South Korea, and Ph.D. degree in computer science from Soongsil University, South Korea. Currently, she is a professor at School of IT Convergence Engineering, Shinhan University, South Korea. Her research interests include parallel algorithms, distributed systems, embedded system, and data mining.