# A Data-driven Approach to the Automatic Classification of Korean Poetry

Joo Hyun Nam and Kin Choong Yow
GIST College, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
Email: harryjhnam@gist.ac.kr, kcyow@gist.ac.kr

***Abstract*—Automatic classification of text is an increasingly important area of research. It has important applications in virtual assistants and recommender systems. Among the different types of literary works, the poem is one of the most difficult to classify automatically because of the prolific use of metaphors and the short length. In this research, we propose a data-driven approach to automatically classify Korean poems. We use three different methods for finding keywords which can train the classifiers. Our results show that the proposed approach can produce better classification accuracy than using a predefined list of keywords created by a human expert.**

***Index Terms* — automatic classification, data-driven, poem, text mining, Korean text, keyword extraction**

## I. Introduction

Classifying text by its topic is useful in finding information that you want. For example, readers will be able to find poems that they want to read easily if poems are classified by its topic. However, poetry is a unique literature genre where the subject is described in an abstract and indirect manner. Therefore, it is difficult to figure out what the topic of a poem is without reading the entire poem. If a poem can be automatically classified by its topic, then it will help to organize a database of poetry systematically and improve its accessibility to the readers.

In this study, we explore the possibility of automatic classification of Korean poems by topics through keywords extracted from a database of poems. Classifying Korean text is a much more challenging task than classifying English text because the Korean grammar system uses many possible forms of word endings in a sentence, as opposed to English. We approach this problem by using a morpheme analyzing tool to separate sentences into keywords. By analyzing the occurrences of these keywords in the poems, we are able to classify Korean poems into categories based on its topic.

The rest of this paper is organized as follows. Section II describes related works in the classification of English text, and provides a brief description of the Korean basic grammar system. Section III describes our data-driven approach to generate and select the keywords used for the classification. Section IV presents the automatic classification result of our approach and Section V gives the conclusion and future work.

## II. Classification of Texts

### A. Classification of English Text

Various techniques and tools have been developed to classify English texts [1, 2]. To analyze English text, the text has to be separated into paragraphs and sentences [3]. A large corpus is usually required to use these methods, and such large corpus based text analysis methods have been used to analyze blogs [4, 5], songs [4], president's speech [4], and reactions of users to brands [6]. However, such methods are hard to be used in analyzing poems because of the much smaller corpus size.

### B. Classification of Korean Text

Unlike English, Korean text should be separated by the morphemes, not words. Morphemes are the smallest unit of words that has meaning. Generally, one independent word, such as a single noun or verb, is a morpheme. Whereas English is mostly separated into morphemes by space, in Korean there is a unit of word called 'eojeol' which is separated by space. And 'eojeol' is separated again into morphemes. For example, "나는" is the Korean translation of "I am" which '나' means 'I' and '는' means 'am'. "I am" has two morphemes 'I' and 'am' which separated by space and "나는" also has two morphemes but there is no space between them. Therefore, Korean text should be analyzed by morpheme rather than space (as opposed to English). In this study, we used the Korean morpheme analyzer tool 'KoNLPy' [7] to separate sentences into morphemes.

## III. Methodology

In this section, we describe three approaches to extract the morphemes, or keywords, that we can use to build our classifier. The poems that we use come from two books of collected poems by Kim Yongtaek, "Maybe the stars take away your sadness" vol.1 and vol.2 [8, 9]. These poems were organized into four categories, namely, "Nature", "Abstraction", "Artificial", and "Human".

The three approaches for finding keywords are: selected keywords, most frequent keywords, and 'should not appear' keywords (called as SNA keywords). These

Manuscript received June 1, 2017; revised August 11, 2017.


doi: 10.12720/jait.8.3.172-176

keywords are all noun, verb, and adjective morphemes. We only take these three word classes because these word classes contain the meaning of the sentence.

### A. Approach 1 – Selected Keywords by Human User

In this first approach, we build a list of keywords that are the most common for poems in each category. For example, in the category of "Nature", we would expect to find words like "나무" (which means "tree") or "눈" (which means "snow"). We will then use these keywords to build a classifier that will classify poems into categories based on the frequency of occurrences of these selected keyword.

Selected keywords are commonly referred morphemes that are the most acceptable words for each category. We generated these lists of selected keywords by ourselves. We chose a list of 40 keywords for each of the four categories. Table I shows a sample of 10 of the selected keywords in each category.

TABLE I. 10 SELECTED KEYWORDS

| *Nature* | *Abstraction* | *Artificial* | *Human* |
|---|---|---|---|
| 사과 (apple) | 사랑 (love) | 책 (book) | 당신 (you) |
| 나무 (tree) | 우울 (gloom) | 길 (road) | 나 (I) |
| 물 (water) | 친구 (friend) | 굴뚝 (chimney) | 우리 (we) |
| 바람 (wind) | 연인 (lover) | 컴퓨터 (computer) | 인간 (human) |
| 풀 (grass) | 감정 (emotion) | 차 (car) | 아이 (child) |
| 산 (mountain) | 기쁨 (joy) | 거울 (mirror) | 어른 (adult) |
| 벌레 (bug) | 슬픔 (sadness) | 시계 (clock) | 아버지 (father) |
| 눈 (snow) | 행복 (happiness) | 신문 (newspaper) | 어머니 (mother) |
| 별 (star) | 고통 (pain) | 학교 (school) | 누이 (sister) |
| 달 (moon) | 분노 (anger) | 가로등 (street light) | 사람 (people) |

### B. Approach 2 – Most Frequent Keywords

Our second approach attempts to generate the most significant keywords purely from data, hence it is a data-driven approach. We examine the frequency of occurrence of morphemes in a set of poems called the training set, and we selected the most frequent morphemes as our keywords.

To perform 10-fold cross validation, we split all poems into ten sets and take nine of them as the training set. This training set is used to build a list of the most frequent keywords for each category, which is used to build a classifier and then tested using the last set. We then repeat the process by taking a different nine sets to form the training set, leaving one out each time. The process is shown in Fig. 1.

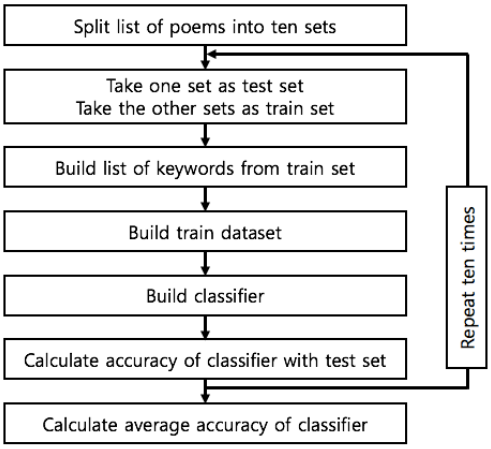


Figure 1. Finding the most frequent keywords

The most frequent keywords are morphemes that appear frequently in poems of the corresponding category. As different keywords appear different number of times in a poem, we calculate the weight for each keyword by normalizing the count by the number of morphemes in the poem and the number of poems in category. Equation (1) shows how we normalize the keywords counts. The list of most frequent keywords is the list of 40 morphemes with the highest count.

$$weight = \frac{SUM(\frac{number\ of\ appearance\ of\ morpheme\ in\ poem}{number\ of\ morphemes\ in\ poem})}{number\ of\ poems\ in\ corresponding\ category} \quad (1)$$

### C. Approach 3 – Should not Appear (SNA) Keywords

Our last approach is to augment our list of keywords with a list of "should not appear" (SNA) keywords. The SNA keywords are morphemes that should not appear in a certain category. Appearance of SNA keywords in a particular category means that the probability of the poem belonging to that category is low.

To find the list of SNA keywords, we use the same method as the most frequent keywords to find the frequency of occurrence of each morpheme. We then choose morphemes that do not appear in a category but appear in other categories. The weights for SNA keywords are calculated using the counts of the morphemes appearing in other categories, and are given a negative value.

### D. Training Classifiers

After we have obtained a list of the keywords, we can now feed it to the classifier to train it. We use two of the most popular classifier, decision trees and SVMs, in our experiments. The input data for training these classifiers consist of a list of four floating point numbers which are the summations of product between frequency of appearance of each keyword and its weight in each category. To study the dependency of the accuracy of the results on the parameters of the classifier (such as maximum depth and minimum number of leaf samples for decision trees), we repeat our experiments on various values of the classifier parameters.

## IV. Results and Discussion

### A. Dataset

From two books of collected poems by Kim Yongtaek, "Maybe the stars take away your sadness" vol.1 and vol.2, we get 154 poems. These poems are divided into four categories, "Nature", "Abstraction", "Artificial", and "Human". Table II is the distribution of poems in each category.

TABLE II. Original Distribution of Categories

| *Category* | *Number* |
|---|---|
| Nature | 47 |
| Abstraction | 43 |
| Artificial | 18 |
| Human | 46 |

### B. Implementation

We used a PC running Windows 10 on an Intel Core i5 processor with 4GB of RAM for the experiments. All our programs are written in Python and developed in IDLE.

### C. Keywords

Table III shows the top ten most frequent keywords and weights extracted from the "Nature" category in approach 2. We noticed that in this category, the top two keywords are "하" (which means "to do") and "가" (which means "to go"), which are also in the top ten most frequent keywords in the other three categories, with similar weights. The reason for this is that "하" and "가" are common action words that would appear in any poem. However, this is not a problem as these words affect the weights in each category in a similar way and thus is ineffective in influencing the classification result.

TABLE III. Top 10 Most Frequent Keywords and Weights of "Nature"

| *Keywords* | | *Weights* |
|---|---|---|
| 하 | do | 0.044093424 |
| 가 | go | 0.031182802 |
| 꽃 | flower | 0.024158324 |
| 있 | exist | 0.019521294 |
| 되 | be | 0.01939417 |
| 울 | cry | 0.01938751 |
| 밤 | night | 0.016711129 |
| 속 | inside | 0.015788395 |
| 보 | see | 0.015385867 |
| 없 | not | 0.014434752 |

Table IV shows the top ten SNA keywords also for the "Nature" category. We observe that some of these keywords, such as "도토리" (which means "acorn") and "벌레" (which means "bug"), could actually appear in some other poems in the "Nature" category. They are not included in this case because they did not appear in our training dataset. This effect is mitigated by the fact that the appearance of these SNA keywords only reduces the overall weight by a small amount, instead of forcing the poem to belong to a different category.

TABLE IV. Top 10 Should Not Appear Keywords and Weights of "Nature"

| *Keywords* | | *Weights* |
|---|---|---|
| 뜨 | rise | -0.005552159 |
| 그립 | miss | -0.005307911 |
| 도토리 | acorn | -0.004716981 |
| 벌레 | bug | -0.003939161 |
| 편지 | letter | -0.003911711 |
| 벗 | friend | -0.003911116 |
| 늦 | late | -0.003902174 |
| 죄 | sin | -0.003684178 |
| 강물 | river | -0.003569871 |
| 가도 | highway | -0.003430532 |

### D. Classification

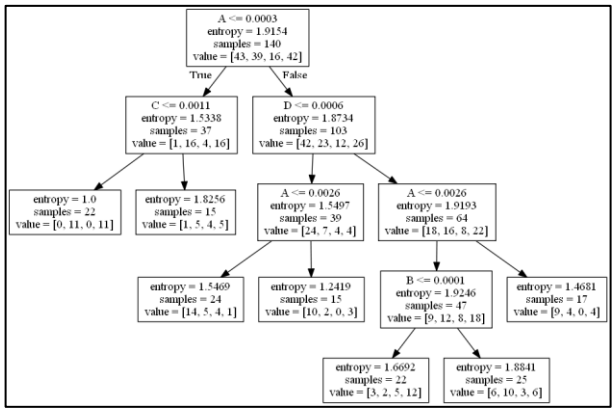


Figure 2. Decision tree by selected keywords

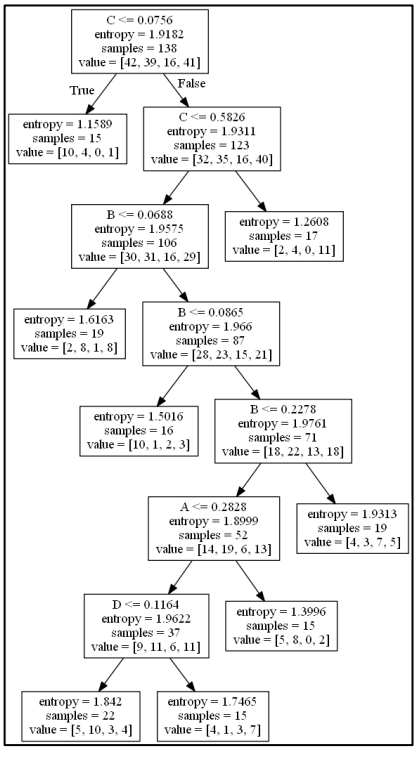


Figure 3. Decision tree by most frequent keywords

We build classifiers using the Scikit-learn toolkit [10]. We use two of the most popular classifier, decision trees and SVMs (linear, RBF, and polynomial) in our experiments.

Figs. 2, 3, and 4 shows the decision trees obtained in each of the three approaches.

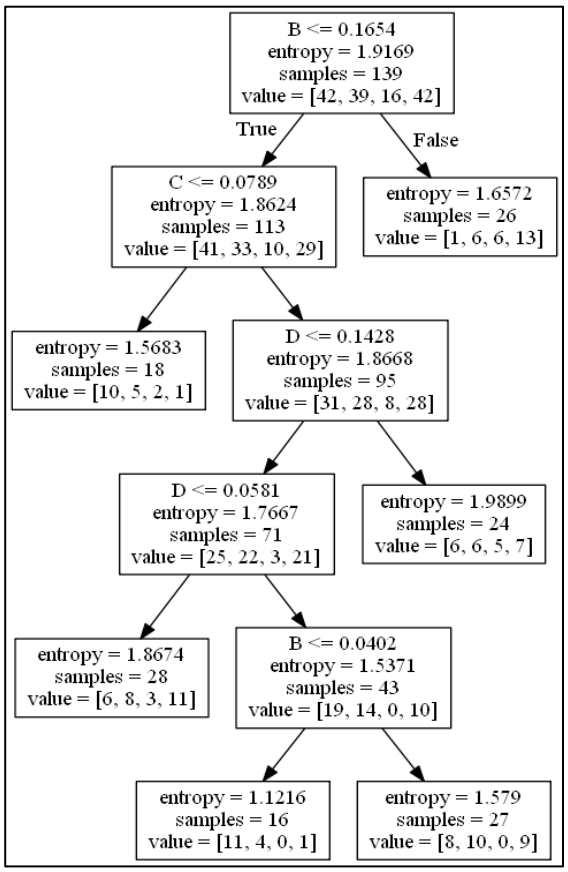


Figure 4. Decision tree by should not appear keywords

## E. Classification Accuracy

Tables V and VI show the accuracy of the decision tree classifier used in our experiments. By repeating the experiments for different parameters of the classifier, we found that for the decision tree, the optimum value is 10 for maximum depth, and 15 for minimum leaf samples.

TABLE V. DECISION TREE ACCURACY BY MAXIMUM DEPTH

| *Maximum Depth* | *Keyword* | | |
|---|---|---|---|
| | *Selected* | *Most Frequent* | *Should Not Appear* |
| 5 | 33.49% | 27.02% | 26.74% |
| 10 | 34.41% | 31.47% | 29.23% |
| 15 | 33.01% | 26.16% | 25.95% |

TABLE VI. DECISION TREE ACCURACY BY MINIMUM LEAF SAMPLES

| *Minimum Leaf Samples* | *Keyword* | | |
|---|---|---|---|
| | *Selected* | *Most Frequent* | *Should Not Appear* |
| 10 | 35.52 | 27.61 | 26.97 |
| 15 | 36.23 | 33.72 | 29.87 |
| 20 | 35.05 | 31.10 | 30.71 |

Table VII show the accuracy of the Support Vector Machine (SVM) classifier using Radial Basis Functions (RBF) used in our experiments. By repeating the experiments for different parameters of the classifier, we found that the optimum value is 0.2 for gamma.

TABLE VII. RBF SVM ACCURACY BY GAMMA

| *Gamma* | *Keyword* | | |
|---|---|---|---|
| | *Selected* | *Most Frequent* | *Should Not Appear* |
| 0.1 | 31.88% | 35.70% | 34.40% |
| 0.2 | 31.88% | 36.60% | 37.85% |
| 0.3 | 31.88% | 35.98% | 36.60% |
| 0.4 | 31.88% | 35.98% | 37.23% |

Table VIII shows the classification accuracy by classifiers in the different approaches. From Table VIII, we can see that there is a greater degree of accuracy using the approach of most frequent keywords as compared to the approach using selected keywords. The approach of using SNA keywords also has a higher accuracy than the approach of using the most frequent keywords.

TABLE VIII. AVERAGE ACCURACY OF EACH APPROACH BY CLASSIFIERS

| *Classifier* | *Keyword* | | |
|---|---|---|---|
| | *Selected* | *Most Frequent* | *Should Not Appear* |
| Decision Tree | 42.93% | 33.72% | 28.69% |
| Linear SVM | 31.88% | 37.85% | 37.85% |
| RBF SVM | 31.88% | 36.60% | 37.85% |
| Polynomial SVM | 29.88% | 31.17% | 31.17% |

We observe that for the case of Decision Trees, the accuracy of using the approach of SNA keywords is actually worse. Let's look at the confusion matrix to determine why the accuracy of decision tree is worse in the approach for SNA keywords.

Tables IX, X, and XI are the confusion matrices of decision trees for the different approaches. From these tables, we can see that the classification accuracy of poems in the "Artificial" category is extremely poor. It may be due to the number of poems in the "Artificial" category. It makes it harder to find the meaningful most frequent keywords and should not appear keywords.

Also, in Tables IX, X, and XI, in the case of the "Nature" category, we also observe that the number of correct classifications is getting smaller from the approach using selected keywords to the approach using should not appear keywords. As shown in Tables III and IV, the most frequent keywords list seems reasonable but the SNA keywords list includes some words that are relevant to nature, such as acorn, bug and river. This may be the reason of the low accuracy. This kind of invalid keywords may appear because of the lack of number of poems.

TABLE IX. CONFUSION MATRIX FOR DECISION TREE FOR THE APPROACH USING SELECTED KEYWORDS

| *Real Category* | *Predicted Category* | | | |
|---|---|---|---|---|
| | Nature | Abstraction | Artificial | Human |
| Nature | 34 | 9 | 0 | 4 |
| Abstraction | 14 | 17 | 1 | 11 |
| Artificial | 5 | 8 | 0 | 5 |
| Human | 12 | 19 | 0 | 15 |

TABLE X. CONFUSION MATRIX FOR DECISION TREE FOR THE APPROACH USING MOST FREQUENT KEYWORDS

| *Real Category* | *Predicted Category* | | | |
|---|---|---|---|---|
| | Nature | Abstraction | Artificial | Human |
| Nature | 24 | 13 | 3 | 7 |
| Abstraction | 16 | 9 | 4 | 14 |
| Artificial | 9 | 1 | 4 | 4 |
| Human | 15 | 9 | 8 | 14 |

TABLE XI. CONFUSION MATRIX FOR DECISION TREE FOR THE APPROACH USING SHOULD NOT APPEAR KEYWORDS

| *Real Category* | *Predicted Category* | | | |
|---|---|---|---|---|
| | Nature | Abstraction | Artificial | Human |
| Nature | 18 | 13 | 1 | 15 |
| Abstraction | 22 | 6 | 1 | 14 |
| Artificial | 6 | 2 | 0 | 10 |
| Human | 17 | 9 | 0 | 20 |

## V. CONCLUSION AND FUTURE WORKS

In this research, we tried three different approaches for finding keywords and weights for the classification of Korean Poetry. In SVMs, data-driven approaches have better accuracy then using user selected keywords. However, in decision trees, the selected keywords approach has better accuracy than data-driven approaches. This may be due to a lack of data in the training set and the number of keywords we choose (we only chose the top 40 keywords per category). Thus, in our future work, we will try to use all words from the poems and not just the top 40. Also we need more poems to train the classifiers and we will try other classifiers such as Bayes' net which is more efficient for datasets with uncertainty.

## ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1A02019235)

## REFERENCES

[1] K. Ahmad, *Affective Computing and Sentiment Analysis Emotion, Metaphor and Terminology*. Springer, GE: Text, Speech and Language Technology Series, 2011

[2] M. Feidakis, T. Daradoumis, and S. Caballe, "Emotion measurement in intelligent tutoring systems: What, when and how to measure," in *Proc. 3rd Int. Conference on Intelligent Networking and Collaborative Systems (INCoS)*, 2011, pp. 807-812.

[3] A. Osherenko, "Towards semantic affect sensing in sentences," in *Proc. the AISB 2008 Symposium on Affective Language in Human and Machine*, 2008, pp. 41-44.

[4] P. Dodds and C. Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, vol. 11, pp. 441-456, 2010.

[5] R. Mihalcea and H. Liu, "A corpus-based approach to finding happiness," in *Proc. the AAAI Spring Symposium on Computational Approaches to Weblogs*, 2006, pp. 19.

[6] K. Voll and M. Taboada, "Not all words are created equal: Extracting semantic orientation as a function of adjective relevance," in *AI 2007: Advances in Artificial Intelligence*, M. Orgun and J. Thornton, Eds. Springer Berlin Heidelberg, 2007, pp. 337-346.

[7] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in *Proc. the 26th Annual Conference on Human & Cognitive Language Technology*, 2014.

[8] Y. Kim, *Maybe the Stars Take away Your Sadness*, Korea: Wisdom House, 2015

[9] Y. Kim, *Maybe the Stars Take away Your Sadness Plus, Korea: Wisdom House, 2016*

[10] Scikit-learn: Machine Learning in Python. [Online]. Available: http://scikit-learn.org/stable/index.html, accessed 2 June 2017

**Joo Hyun Nam** is a B.S student in GIST college, Gwangju Institute of Science and Technology, Republic of Korea. She is now majoring in Mechanical Engineering. Her current research interests include deep learning, natural language system and recommender system.

**Kin Choong Yow** obtained his B.Eng. (Elect) with First Class Honours from the National University of Singapore in 1993, and his Ph.D. from the University of Cambridge, UK, in 1998. Prior to joining GIST college, he was Professor at the Shenzhen Institute of Advanced Technology, P.R.China, and Associate Professor at the Nanyang Technological University of Singapore. He runs the Generic Intelligence and Smart Environment Lab (GISEL) in GIST, and his current research interests includes cognitive models, biologically inspired vision, artificial consciousness, commonsense reasoning, deep learning, intelligent CCTV systems, and autonomous robot operations in smart environments. Prof. Yow has published over 80 research papers, and has held a number of leadership roles in various international journals and conferences including JAIT, IntJIT, CCGrid2013, ICPAD2012 and MobileHCI2007.