Journal of Engineering and Applied Sciences 15 (4): 965-969, 2020
ISSN: 1816-949X

Identifying Natural Categories of Learning Tasks using Clustering Techniques for the DeMite Game

[1]Rifki Wijaya, [2]Agus Sukoco, [3]Hashfi Rasis Hakim and [3]Ary Setijadi Prihatmanto
[1]Department of Computer Engineering, Faculty of Electrical Engineering,
Telkom University, Bandung, Indonesia
[2]Department of Informatics, Faculty of Computer Science, Universitas Bandar Lampung,
Lampung, Indonesia
[3]School of Electrical Engineering and Informatics, Institut Teknologi Bandung,
Bandung, Indonesia

Abstract: The DeMite is an educational game to learn how to pronounce English words with augmented reality, location-based and speech-to-text features presented in horror genre. This game is designed for non-English speakers, specifically Indonesian speaking user. Many people in Indonesia still cannot speak English properly, thus, there is a need to carefully select the appropriate English words as learning tasks in the DeMite game. The proposed method shows that each learning tasks can be grouped into their natural categories using clustering method.

Key words: Clustering, natural categories, learning tasks, educational, augmented, specifically

INTRODUCTION

Learning methods have evolved during these past years. Now a days, students can find any information that they need using the internet. This method makes the old learning methods such as reading the book or asking the teachers will be abandoned and will be replaced with the newer method.

Educational games are one of the new methods that applies elements of game to make the player can learn while getting the experience of playing game at the same time.

English has become the number one language that most people have used in the recent year. But there are some countries that have low proficiency level in English. Based on the studies that Education First (EF), one of English course place in Indonesia did in 2017 Indonesia has English proficiency level of 52.15 which is lower than, average English proficiency level in Asia that is 53.60.

The DeMite is an educational game with augmented reality, location-based and speech-to-text features presented in horror genre. The purpose of the DeMite is to capture or kill the ghost by speaking the right words that appear in the ghosts. When the player successfully captured or killed the ghost, the player gains experience and he get achievements to help further into the game. The DeMite game is designed for non-speaking English country, specifically, Indonesian speaking user. The inspiration came from Pokemon Go which has similar engine except for no speech-to-text feature. The ghosts appear in the DeMite game are known Indonesian ghost such as Pocong, Kuntilanak, etc.

This game has a purpose to teach Indonesian people to properly pronounce English words. To make the learning process more exciting, the player pronounces a specific word to catch or kill the ghost. The game spawns ghost geographically where the augmented reality is applied. If the player pronounces the word correctly, then, it will decrease the ghost's health and the player receives additional scores. In contrast, the player's fear increases.

However, we cannot use random words as the learning tasks in this game. Randomly generated words can cause problems (Hellekalek, 1998) in determining the decreased ghost's health and player's score as the random word does not contain any information other than the word itself.

To overcome the problem, we need a step in Procedural Content Generation (PCG) called content categorization (Togelius *et al*., 2011) should fit in. Firstly, we collected words, the learning tasks in the DeMite game, from the popular English words in Indonesia. Then, we assign features relevant to the words such as total letters, phonetic and total syllables. Finally, the parameters will be used to calculate which level the words belong into with machine learning, specifically hierarchical clustering and K-means clustering.

Services, especially, for Information of technology today are complicated and develop to become such machine learning, big data (Sukoco *et al*., 2019). Machine

Corresponding Author: Rifki Wijaya, Department of Computer Engineering, Faculty of Electrical Engineering,
Telkom University, Bandung, Indonesia

Learning is the science to enable computer to learn specific functionality via a collection of information. In the past decade, machine learning has given us many possibilities such as a self-driving car, effective web search and traffic prediction. There are two general phases in machine learning: the training phase which is the phase to collecting the data, provides features and calculates the feature values predicting phase which is the phase to predict the unlabeled data using the various feature that has already defined before.

There are three main types of machine learning: Supervised learning, unsupervised learning and semi- supervised learning. Supervised learning can apply what has been learned in the past in order to predict the new data accurately. Starting from an analysis of the known training dataset, the learning algorithm produces a function to make the prediction about the output values. The relevant technique we apply here is unsupervised learning. Basically, it is used when the information to train the dataset is not classified or not labelled. One of the goals of unsupervised learning is to found a hidden structure from the dataset that has not been labelled (Ghahramani, 2004).

Roberts and Chen (2014) have shown that abundant and unorganized set of content were categorized quite successfully using machine learning method. Clustering algorithm helped the process became much simpler for the oracle by annotating only the representative sample in the center (called medoid) of each cluster. To address quality issue of PCG, (Shi and Chen, 2015) exploits the synergy between rule-based and learning-based methods that was easy-to-design rules to remove unappealing game segments.

Rosyid *et al*. (2018), applied similar approach in the aforementioned paragraph. He also suggested that if the learning space is very large, then the clustering techniques can be useful to identity such categories naturally.

MATERIALS AND METHODS

Purpose of this research is to make a model that can automatically label the level into the data. In this section, we describe our approach and steps to produce the level assigned to each word. Firstly, we collect the words from various websites and the syllabus used in primary school. Then, we define the features of the words collected and assigning the value for each feature. After that we use the clustering method to get to know the structural pattern of these learning tasks. We expect that by a careful selection of words, there may exist particular groups of words with respect to their features. Through this method, we can organize words in more details beyond merely set of text. This process enables more comprehensive details of English pronunciation assessment of each word in the game session. And we can use it as the tools to summarize the player's English pronunciation skills as a whole.

The raw learning tasks were collected from https://salamadian.com/kosakata-bahasa-inggris/. However, we needed to remove ambiguous and not common words. This process is still feasible using spreadsheet functions. As a result, we ended up selecting 423 words as the raw dataset.

In the next step we identified the features relevant to the collected words. From our observations and knowledge, we find out that the most suitable features for words are its total letters, total syllables, phonetics using IPA and three subcategories. The three subcategories were driven by the fact that although, Indonesian and English use the same alphabetic letters from A-Z, they have a different pronunciation that varies from each other. The letters have similar, slightly different and absolutely different pronunciation. This is one of the reasons why the people in Indonesia have a hard time learning to speak English even though the alphabetic letters are the same.

We divided the difference between textual appearance and its phonetic into three sub-categories because of the different pronunciation of the letters. The same letter may have similar pronunciation, minor different pronunciation and major different pronunciation. The first sub-category consists of alphabet letters that have the same or similar pronunciation between English and Indonesian. The letters included in this sub-category are F, L, M, N, O, Q, S, X. There is no appealing difference to pronounce these letters by Indonesian tongue.

The second sub-category consists of alphabet letters that have a minor difference pronunciation between English and Indonesian. The letters included in this sub- category are A, B, C, D, E, G, J, K, P, T, V, Z. For example in the Indonesian language the letter A is pronounced as ah and in English it pronounced as eɪ.

The third sub-category consists of alphabet letters that have significant difference of pronunciation between English and Indonesian. The letters included in this sub-category are H, I, R, U, W, Y. For example in the Indonesian language the letter H is pronounced as hah but in the English it was pronounced as eat.

After selecting the feature that we wanted to use, we proceed to assign the values to the feature of the words. Total letters can be achieved by using string length that will return the number of characters inside the string. The phonetics are obtained using the data muse Application Programming Interface (API), API for word-finding which can have parameters to set constraints to provide the desired results which in this case is the word and its phonetics. We calculate the three sub-categories by finding each letter from the word. To calculate the sub-category, we find the occurrence of the letters that are included in the corresponding sub-category.

The first approach is to perform clustering, to group words which has similar features. In this experiment, we

use hierarchical clustering and K-means clustering in order to label the data. In our experiment, the hierarchical clustering is used to label the words based on the parameters of features that already defined before. We use several linkages to find which linkage is the best for our data. The linkages used are average, centroid, median and ward.

Next step is to perform the K-means clustering. We define the method to find the corresponding number of clusters by using elbow method and average silhouette analysis. Elbow method is a method to find number of clusters by plotting the curve according to the number of clusters K and finding the location of a bend in the plot as an indicator of the appropriate number of clusters. Average silhouette method computes the average silhouette observations for different values of K clusters.

Because clustering didn't have significant result, the next approach that we take is to do parameter priority sorting. With parameter priority sorting, we select and sort parameter based on its importance. In this study, the writer sorted parameters starting from the most important to least important are total letters, total syllables, total ambiguity, third sub-category, second sub-category and first sub-category. Based on this parameter, we can distribute the words evenly for each level. To calculate each sub-category for a certain word, we can do equation as following. For first sub-category, we use this following equation.

$$s1 = n(F)+n(L)+n(M)+n(N)+n(O)+n(Q)+n(S)+n(X)$$

Where:
s1 : The result of the first sub-category
n (F) : The total occurrences of letter F in the word
n (M) : For total occurrences of letter M
n (N) : For total occurrences of letter N
n (O) : For total occurrences of letter O
n (Q) : For total occurrences of letter Q
n (S) : For total occurrences of letter S and
n (X) : For the total occurrences of letter X

$$s2 = n(A)+n(B)+n(C)+n(D)+n(E)+n(G)+n(J)+ n(K)+n(P)+n(T)+n(V)+n(Z)$$

Above equation is for the second sub-category where s2 is the result of the second sub-category and n (letter) is the total occurrences for each letter corresponding with their alphabetical letter.

$$s3 = n(H)+n(I)+n(R)+n(U)+n(W)+n(Y)$$

The equation above is similar with other two equations which s3 is the results for the third sub-category and the n (letter) is the total occurrences for each letters corresponding with their alphabetical letters.

Table 1: Collected raw words statistics

Collected words statistics	Value
Total words	423
Shortest word	Go, He
Longest word	Congratulations
Shortest word length	2
Longest word length	15
Total noun	150
Total verb	137
Total adjective	136

Table 2: Parameter priority sorting results

Level	Total words
1	44
2	45
3	42
4	44
5	44
6	43
7	43
8	43
9	43
10	32
Total	432

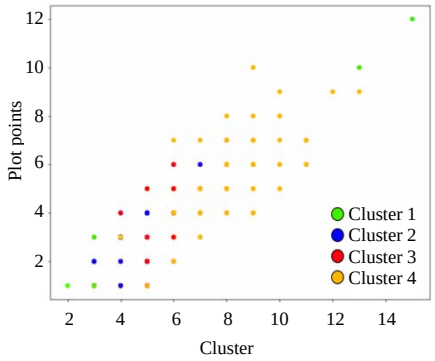

Fig. 1: Plot points with cluster dependent colours

RESULT AND DISCUSSION

In this section, we report the implementation and testing of the clustering techniques in order to determine the level for each word. We provide the comparison of the results between each clustering methods.

We collected 423 words that is popular in Indonesia as the learning task, Table 1 shows the summary of the dataset. We annotated the values of the features of each word using a computer program of which input is the strings of the word.

After determining feature values for each word, we applied clustering techniques: Firstly, we used hierarchical clustering to group the data into a tree-root-shaped structure. Hierarchical clustering uses different linkage in order to calculate distance between clusters. The linkage types that we are using in this research are average, centroid, median and ward.

In Fig. 1, we can see the data point with the colors as its clusters with each color represents the cluster.

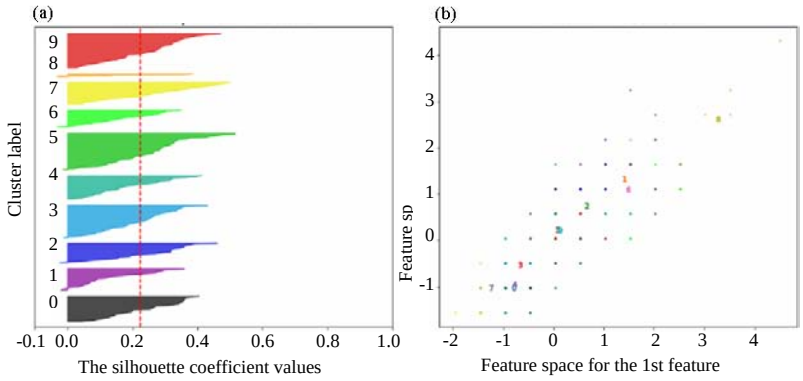

Fig. 2(a, b): Silhouette analysis for n_clusters = 10

Because the data set are not that big and the data is too diverse, the plot points was scattered and not really grouped together.

Figure 2 explains the silhouette analysis for k-means with 10 clusters. There are some clusters that incorrectly group the learning tasks.

Because clustering didn't provide the result that we want, we use parameter priority sorting. The result of parameter priority sorting can be seen in Table 2. In Table 2 we can see that the words are grouped evenly and can be used as learning task for the DeMite game. In this section, we discuss the issues arising from our work.

Creating rule-based for learning tasks is easier and better than generated with machine learning. However with the numbers of growing data to be analyzed and checked if they are suitable for the learning task, the machine learning approach deemed to be necessary. Using the machine learning, the developer can save many times and can focus into the further developing the game.

We need further investigations whether word we categorized are appropriate to be used in game session. There are factors such as the ambiguity of the word that cannot be checked automatically without proper assessment. When we develop the model for the first time, we are not aware considering the phonetic ambiguity of the word could impact on determining the level. We find out the ambiguity problem when we are doing assessment to calculate the score.

To counter the problem, we are planning to apply an active learning (Prince, 2004) techniques to mark the quality of a small number of representative learning tasks as the filtering model. We take sample of the words randomly using a Sampler and then we as an Oracle, determining the label if the learning tasks are suitable are not. Then a supervised learning model is trained using these labelled samples. Next, we use the cross-validation (Arlot and Celisse, 2010) to validate the training set and the test set. We recursively do this action until the cross-validated satisfy a certain threshold. After finding the suitable model, we can use it to determine the remaining learning tasks whether the word can be used or not. If the word cannot be used, we will replace the word with the non-ambiguous candidate that have relatively similar feature values.

CONCLUSION

Rule-based level determining needs a specific feature to determine the levels correctly. It also needs to be assessed if the feature that we selected feasible to determining the level. Using clustering techniques can help determining the level automatically but sometimes the learning tasks is grouped on the wrong cluster. By using these, we achieve the generated levels for each word, although, there is ambiguity problem that needs to be tackled of. However with clustering techniques the learning tasks are not grouped evenly and it cannot be used for the DeMite game. Using parameter priority sorting we can achieve the result to group the learning tasks into clusters evenly.

ACKNOWLEDGEMENT

The researchers would like to thank the action editor, the DeMite team members and people involved in creating The DeMite game for the valuable comments that improve the presentation of this manuscript.

REFERENCES

Arlot, S. and A. Celisse, 2010. A survey of cross-validation procedures for model selection. Stat. Surv., 4: 40-79.

Ghahramani, Z., 2004. Unsupervised Learning. In: Advanced Lectures on Machine Learning, Bousquet, O., U. von Luxburg and G. Ratsch (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-23122-6, pp: 72-112.

Hellekalek, P., 1998. Good random number generators are (not so) easy to find. Math. Comput. Simul., 46: 485-505.

Prince, M., 2004. Does active learning work? A review of the research. J. Eng. Educ., 93: 223-231.

Roberts, J. and K. Chen, 2014. Learning-based procedural content generation. IEEE. Trans. Comput. Intell. AI. Games, 7: 88-101.

Rosyid, H.A., M. Palmerlee and K. Chen, 2018. Deploying learning materials to game content for serious education game development: A case study. Entertainment Comput., 26: 1-9.

Shi, P. and K. Chen, 2015. Learning constructive primitives for online level generation and real-time content adaptation in Super Mario Bros. Artif. Intell., Vol. 1,

Sukoco, A., A.S. Prihatmanto, R. Wijaya, Il. Sadad and R. Darmakusuma, 2019. SEMUT: Next generation public transportation architecture in the era IoT and big data. J. Eng. Appl. Sci., 14: 4052-4059.

Togelius, J., G.N. Yannakakis, K.O. Stanley and C. Browne, 2011. Search-based procedural content generation: A taxonomy and survey. IEEE. Trans. Comput. Intell. AI. Games, 3: