Word Clustering Using Word Embedding Generated by Neural Net-based Skip Gram

Hosung Park, Minkyu Lim, Donghyun Lee, and Ji-Hwan Kim*
Department of Computer Science and Engineering
Sogang University
Seoul 04107, South Korea
{hosungpark & limkki & redizard & kimjihwan }@sogang.ac.kr

Jeong-Sik Park
Department of Information and Communication Engineering
Yeungnam University
Gyeongsan, South Korea
parkjs@yu.ac.kr

Gil-Jin Jang
School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
gjang@knu.ac.kr

Abstract—This paper proposes word clustering using word embedding. We used a neural net-based continuous skip-gram method for generating word embedding in continuous space. The proposed word clustering method represents each word in the vector space using a neural network. The K-means clustering method partitions word embedding into predetermined K-word clusters.

Keywords—Word clustering; Word embedding; Skip gram; Neural Network; N-gram; Language Model;

I. INTRODUCTION

A speech recognition system suggests the word series yield the greatest product of $P(O|W)$ and $P(W)$, as in (1). $P(W)$, generated from the language model of the speech, denotes the word probability. $P(O|W)$ is calculated from the acoustic model[1].

$$\bar{W} = \arg\max_W P(W|O) = \arg\max_W P(O|W)P(W) \quad (1)$$

The application of language models includes not only speech recognition but also machine translation, natural language processing, and typo detection. Conventional language models express relations between words with a probability model by using statistic data based on n-gram word frequency. In this approach, the prediction is rather unstable because the probability of unseen word sequence is calculated with smoothing such as back-off or interpolation. It also has a limitation that the restriction of $n$ does not allow expressing the longer history data. Unseen word sequence refers to the word sequence not occurring in the training data. Increasing the size of $n$ allow for expressing longer history but it also increases the number of unseen word sequences.

The class-based language model can be useful for dealing with the problem of unseen word sequence[2]. If the training data contained ‘Monday’ and ‘Tuesday’, we knew that there could be in all day of week. Therefore, we could predict the language model probability of a day of week, such as ‘Wednesday’, ‘Thursday’, ‘Friday’, ‘Saturday’, and ‘Sunday’. The group of words refers to ‘word class’ or ‘cluster’. Word clustering algorithms such as frequency-based, cross entropy-based, and similarity-based clustering algorithms are highlighted for solving the problem of unseen word sequences.

Word embedding is the feature learning techniques in natural language processing that words from the vocabulary are mapped to arrays of real numbers in a low dimensional continuous space, relative to the vocabulary size. The representative method for word embedding is neural net-based continuous skip-gram method.

In this paper, we investigate word clustering algorithms via previous researches. This paper is organized as follows. Section 2 introduces word clustering methods. Section 3 describes word clustering algorithms Section 4 concludes this paper.

II. RELATED WORK

The frequency-based clustering method creates clustering information using frequency-binning factorization, and each word is assigned to their class according to their word frequency (Mikolov and Kombrink 2011). For example, in the case of a news text corpus with a one million words level, unigram cumulative probability distribution according to word frequency clearly shows the Zipf’s law. Zipf’s law is a theory that shows that the frequency of any word is inversely proportional to the rank of the word when words represented in a natural language corpus expression are listed according to their frequency of use in descending order. Thus, the highest frequency word is approximately twice as high in terms of usage as the second highest frequency word and three times as high as the third highest frequency word. This shows that several high frequency words account for most of the text in the corpus. Since the number of high frequency words is small, they tend to comprise a single small word cluster. On the other hand, low frequency words tend to comprise a word cluster that includes a large number of words.

A cross entropy-based clustering method is a data-driven method that is based on entry criterion, which is widely used in natural language processing (Shi et al. 2013). The input to this method is a text sentence composed of word strings, whereas the output is in binary tree format, with each leaf node composed of
words. Leaf nodes that have the same parent node in this tree are created as a single word cluster. The most general criterion is to maximize the log likelihood of the learning data or to minimize cross entropy. Bigram statistical information is a practical implementation and a single word is assumed to be mapped into a single cluster.

The continuous skip-gram method predicts the current input and output words when the context is used in the continuous bag-of-words model. That is, it predicts the front and rear words within a certain range based on a specific word in a single sentence. In our implementation, five front and rear words are predicted and experiments conducted.

The input and output values in the continuous skip-gram method are as follows: First, an index was assigned to all vocabularies in the vocabulary and |V| was set as the size of the vocabulary dictionary. The input word embedding was set to zero for all the words in the |V|-dimension vocabulary and the value of dimension that correspond to each currently word index was set to one. The output word embedding was set to zero for all the words in the |V|-dimension vocabulary and the values of word indexes that correspond to five word history and five words future words were set to one. A word embedding table is included in the projection layer. In a word embedding table, every word in the vocabulary dictionary has its own arbitrary P-dimension value. The word embedding table is assumed to have a continuous uniform distribution within a certain range (e.g., [-1, 1]). Word indexes that have one in the input word embedding were searched and the P-dimension values of the searched word indexes are found from the word embedding table. Then, the word embedding values of the input words are summed and log linear function is applied to them. The generated word embedding was assumed to be a goal word embedding. The values corresponding to five-word history and five-word future words in the goal word embedding become one. The word embedding values of the input words in the word embedding table are adjusted by using a back-propagation algorithm.

The proposed clustering method creates a word class by representing each word in the vector space using a neural network and measuring the similarity between words. The core words in this method are represented as a word embedding in continuous space, thereby employing K-means clustering. The K-means clustering method is one of the simplest unsupervised learning algorithms that solve a clustering problem. It is a method of classifying word embedding given via predetermined K word clusters. Its input, K, the number of clusters excluding data, is called a seed point. A seed point is chosen randomly; however, knowledge about preferred cluster structure may also be used to select the seed point. This means that as soon as a single word embedding is joined to a single word cluster, the centroid of the word cluster is recalculated. Only two learning steps in a data set are carried out in the K-means algorithm. The learning process is as follows: First, the algorithm starts with a K word cluster. For the remaining word embedding, the closest centroid is searched. A word embedding is then included in a cluster that is identified as having the closest centroid to the above centroid. Once every word embedding is assigned, the centroid of the assigned word cluster is recalculated. In the next step, the input word embeddings are processed again. For every word embedding, the closest centroid is searched. A sample is placed in the word cluster identified as having the closest centroid.

IV. CONCLUSIONS

In this paper, we proposed word clustering method using word embedding generated by neural net-based skip gram. We used neural net-based continuous skip-gram method for generating word embedding in continuous space. For clustering word embedding, the proposed word clustering method represents each word in the vector space using a neural network. The K-means clustering method partitions word embedding into predetermined K word clusters.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NO. NRF-2014R1A1A1002197).

REFERENCES