Musical Word Embedding: Bridging the Gap between Listening Contexts and Music

Seungheon Doh 1 Jongpil Lee 1 Tae Hong Park 2 Juhan Nam 1

Abstract
Word embedding pioneered by Mikolov et al. is a staple technique for word representations in natural language processing (NLP) research which has also found popularity in music information retrieval tasks. Depending on the type of text data for word embedding, however, vocabulary size and the degree of musical pertinence can significantly vary. In this work, we (1) train the distributed representation of words using combinations of both general text data and music-specific data and (2) evaluate the system in terms of how they associate listening contexts with musical compositions.

1. Introduction
Music listeners typically rely on a combination of listening contexts to find music including elements of mood, theme, time of day, location and activity. This scenario can be handled by defining a dictionary of contextual terms and directly associating them with music as a class label (Yan et al., 2015; Ibrahim et al., 2020). However, such a music tagging approach (i.e., multi-label classification) is severely limited in considering contextual expression complexities that listeners can use from a natural language perspective. For example, a listener may use ‘club’ to search for electronic dance music, and unless a model is trained with this specific word, it is not possible to consider the word as a query string. This issue has been addressed by representing tag words with embedding vectors and associating them with music in several different settings such as zero-shot learning (Choi et al., 2019), query-by-blending (Watanabe & Goto, 2019) and multi-task music representation learning (Schindler & Knees, 2019). The aforementioned approaches were based on system training utilizing word embedding with either general text (e.g., Wikipedia or Gigaword) or music-specific corpus (e.g., tags, lyrics, artist IDs, track IDs). What is noteworthy here is that the general text training approach is limited in reflecting "musical" dimensions, whereas music-specific corpus limits incorporation of listening contexts which are not directly related to music while simultaneously suffering from small vocabulary size. In this work, we investigate various word embedding spaces trained with combinations of general and music-specific text data to bridge the gap between listening contexts and music.

2. Datasets and Method
We conducted our research using the latest Wikipedia dump 1 for general text data and a hybrid music corpus for music-specific text data. The music corpus is composed of Amazon album review, AllMusic tags 2, and artist/track IDs. The Amazon album review data contain consumer opinions about the music (He & McAuley, 2016), which was obtained from the MuMu dataset 3 (Oramas et al., 2017). The Allmusic dataset includes music tags (genre, style) and context tags (mood and theme) (Schindler & Knees, 2019). The artist/track IDs were obtained from the MSD dataset (Bertin-Mahieux et al., 2011). The IDs are also regarded as a unique word associated with the corresponding music (Watanabe & Goto, 2019). We used Word2Vec based on Continuous Bag of Words (CBOW) to learn word embedding (Mikolov et al., 2013). For music corpus, we clustered review texts, tags, and artist/track IDs for each music track using MSD track id 4 and MusicBrainz id 5 and also took the context window within the cluster. Additionally, we shuffled words within clusters to address data augmentation. Although this method broke review sentence order, it improved capture of word co-occurrences in the hybrid set with greater spread.

3. Experiments
We trained the word embedding model with vector size 100, window size 15, and five iterations. To test word embedding, we used test tags from two different datasets and thus, different characteristics. Allmusic was used to enable a balanced distribution of music terms and context terms

---

1Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea 2Steinhardt School of Culture, Education, and Human Development, New York University, New York, United States.
Correspondence to: Juhan Nam <juhan.nam@kaist.ac.kr>.

Musical Word Embedding: Bridging the Gap between Listening Contexts and Music

Table 1. Compare ranking evaluation metric between 7 embedding spaces.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Unique Word</th>
<th>Unique Track</th>
<th>Unique Artist</th>
<th>AllMusic (Best) Spearman</th>
<th>SDCEG@50</th>
<th>LastFM (Best) Spearman</th>
<th>SDCEG@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>[AllMusic Tags + Amazon Music Reviews] (Augmented) + Wikipedia</td>
<td>1.98B</td>
<td>11,622,471</td>
<td>521,778</td>
<td>28,330</td>
<td>0.194</td>
<td>0.327</td>
<td>0.312</td>
<td>0.591</td>
</tr>
<tr>
<td>AllMusic Tags + Amazon Music Reviews + Wikipedia</td>
<td>1.8B</td>
<td>11,622,471</td>
<td>521,778</td>
<td>28,330</td>
<td>0.187</td>
<td>0.253</td>
<td>0.226</td>
<td>0.548</td>
</tr>
<tr>
<td>AllMusic Tags + Wikipedia</td>
<td>1.76B</td>
<td>11,163,229</td>
<td>507,435</td>
<td>25,203</td>
<td>0.157</td>
<td>0.215</td>
<td>0.183</td>
<td>0.526</td>
</tr>
<tr>
<td>[AllMusic Tags + Amazon Music Reviews] (Augmented)</td>
<td>0.27B</td>
<td>664,163</td>
<td>521,778</td>
<td>28,330</td>
<td>0.267</td>
<td>0.339</td>
<td>0.407</td>
<td>0.626</td>
</tr>
<tr>
<td>AllMusic Tags + Amazon Music Reviews</td>
<td>0.45B</td>
<td>664,163</td>
<td>521,778</td>
<td>28,330</td>
<td>0.197</td>
<td>0.232</td>
<td>0.358</td>
<td>0.600</td>
</tr>
<tr>
<td>AllMusic Tags</td>
<td>7.3M</td>
<td>1,401</td>
<td>507,435</td>
<td>25,203</td>
<td>0.252</td>
<td>0.282</td>
<td>0.551</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1.75B</td>
<td>11,163,055</td>
<td>0</td>
<td>0</td>
<td>0.088</td>
<td>0.167</td>
<td>0.162</td>
<td>0.551</td>
</tr>
</tbody>
</table>

Figure 1. Tag visualization of different type of embedding using t-SNE. The music corpus includes AllMusic Tags, Amazon music reviews, MSD Artist/Track IDs.

Results and Discussion

Table 1 shows performance results, size of the training corpus, unique words, unique tracks, and unique artists of each method. The results show that the two word embeddings including music corpus significantly outperform the model trained with Wikipedia only. This is expected as the test sets were based on music tag datasets. Between music corpus only and music corpus with Wikipedia, the result depended on how many music terms and context terms are balanced in the test sets. When music terms are concentrated (last.fm tags), word embedding trained with music corpus only outperformed that with both music corpus and Wikipedia. However, in the balanced case (AllMusic tags), word embedding trained with both music corpus and Wikipedia resulted in improved performances. Table 1 also shows that the augmented music corpus achieved notable high performance results. This suggests that the proposed data augmentation is beneficial when the order of words is not important. The t-SNE plot in Figure 1 provides a more intuitive visualization of our research results. Here, we used two music genre terms ‘electronic’ and ‘house’ and three listening context terms ‘club’, ‘club_dance’, and ‘partying’ as relevant words. In Wikipedia, the gaps between terms are significant with only ‘house’ and ‘club’ in close proximity. In the music corpus, the two genre terms and ‘club’ and ‘club_dance’ are tightly clustered while ‘partying’ is significantly beyond the cluster centroid. In the music corpus with Wikipedia, while the context term ‘partying’ is still outside of the cluster containing all of the other terms, it is substantially closer than the music corpus example. This indicates that using both general and music-specific data has the potential of capturing a more balanced correlation between music and listening context (for examples of music retrieval tasks using context words, please refer to 6).

5. Future Work

Our current plan is to expand on findings as reported in this paper and build a set of user-annotated word-to-word similarity pairs to directly measure the relationship between general words, music contexts, and music tracks. We also plan to additionally use the musical word embedding system from trained word embedding as a prototype vector for each music track in the context of audio-based music regression (Van den Oord et al., 2013), classification, and metric learning (Choi et al., 2019) settings. This will allow us to construct a more nuanced audio embedding system as conventional music classification is in the order hundreds of labels as class prototypes, while the proposed approach allows for half a million of track prototypes that are strongly reflective of millions of music context terms.

References


