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

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

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

[1] https://dumps.wikimedia.org/enwiki/20200601/
[2] https://www.allmusic.com
[3] https://www.upf.edu/web/mtg/mumu
[4] http://millionsongdataset.com/
[5] https://musicbrainz.org/

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

Corpus	Size	Unique Word	Unique Track	Unique Artist	AllMusic (Seen)		LastFm (Unseen)	
					spearmanr	nDCG@30	spearmanr	nDCG@30
[AllMusic Tags + Amazon Music Reviews] (Augmented) + Wikipedia	1.98B	11,622,471	521,778	28,330	0.194	0.327	0.312	0.591
AllMusic Tags + Amazon Music Reviews + Wikipedia	1.8B	11,622,471	521,778	28,330	0.187	0.233	0.226	0.548
AllMusic Tags + Wikipedia	1.76B	11,163,229	507,435	25,203	0.157	0.215	0.183	0.526
[AllMusic Tags + Amazon Music Reviews] (Augmented)	0.27B	664,163	521,778	28,330	**0.267**	**0.339**	**0.407**	**0.626**
AllMusic Tags + Amazon Music Reviews	45.3M	664,163	521,778	28,330	0.187	0.232	0.358	0.600
AllMusic Tags	7.1M	1,401	507,435	25,203	0.252	0.242		
Wikipedia	1.75B	11,163,055	0	0	0.098	0.167	0.162	0.551

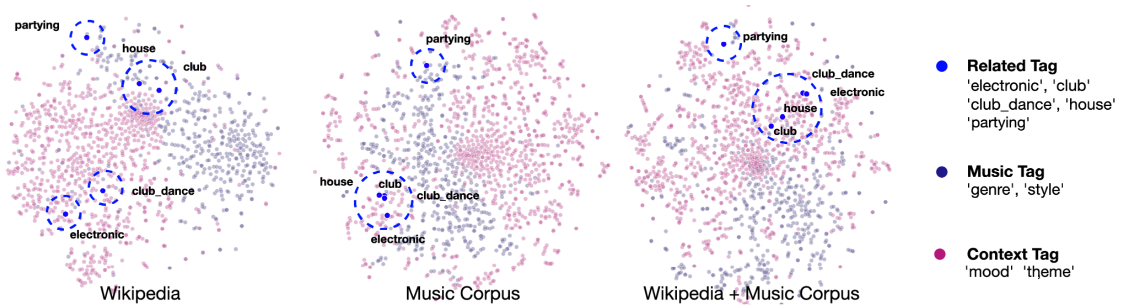

Figure 1. Tag visualization of different type of embedding using t-SNE. The music corpus includes *AllMusic* Tags, *Amazon* music reviews, MSD Artist/Track IDs.

consisting of 1,401 genre/style music tags and mood/theme context tags. This dataset was also used for training (seen) as part of the music corpus. The *last.fm* dataset with genre, mood, and eras tags was also used. We selected the top 100 tags with maximal occurrence frequency. The latter dataset which focuses on music terms was not included in the training (unseen) phase. To measure word-to-word similarity performance of the proposed word embedding system, we employed a co-occurrence of tags scheme for ground truth creation. We then measured spearman's rank correlation and normalized discounted cumulative gain at k (nDCG@k) between ground-truth co-occurrence and word-to-word similarity of word embeddings. For the nDCG evaluation, we use the top k retrieved words ($k = 30$).

4. 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.

[6] https://dohppak.github.io/MusicWordVec

References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

Choi, J., Lee, J., Park, J., and Nam, J. Zero-shot learning for audio-based music classification and tagging. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 67–74, 2019.

He, R. and McAuley, J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 507–517, 2016.

Ibrahim, K. M., Royo-Letelier, J., Epure, E. V., Peeters, G., and Richard, G. Audio-based auto-tagging with contextual tags for music. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16–20. IEEE, 2020.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *Proc. International Conference on Learning Representations,Workshop Track Proceedings,ICLR*, 2013.

Oramas, S., Nieto, O., Barbieri, F., and Serra, X. Multi-label music genre classification from audio, text, and images using deep features. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 23–30, 2017.

Schindler, A. and Knees, P. Multi-task music representation learning from multi-label embeddings. In *Proc. International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6. IEEE, 2019.

Van den Oord, A., Dieleman, S., and Schrauwen, B. Deep content-based music recommendation. In *Advances in neural information processing systems*, pp. 2643–2651, 2013.

Watanabe, K. and Goto, M. Query-by-blending: a music exploration system blending latent vector representations of lyric word, song audio, and artist. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 144–151, 2019.

Yan, Q., Ding, C., Yin, J., and Lv, Y. Improving music auto-tagging with trigger-based context model. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 434–438. IEEE, 2015.