MULTILINGUAL MUSIC GENRE EMBEDDINGS FOR EFFECTIVE CROSS-LINGUAL MUSIC ITEM ANNOTATION

Elena V. Epure
Deezer Research

Guillaume Salha
Deezer Research

Romain Hennequin
Deezer Research

research@deezer.com

ABSTRACT

Annotating music items with music genres is crucial for music recommendation and information retrieval, yet challenging given that music genres are subjective concepts. Recently, in order to explicitly consider this subjectivity, the annotation of music items was modeled as a translation task: predict for a music item its music genres within a target vocabulary or taxonomy (tag system) from a set of music genre tags originating from other tag systems. However, without a parallel corpus, previous solutions could not handle tag systems in other languages, being limited to the English-language only. Here, by learning multilingual music genre embeddings, we enable cross-lingual music genre translation without relying on a parallel corpus. First, we apply compositionality functions on pre-trained word embeddings to represent multi-word tags. Second, we adapt the tag representations to the music domain by leveraging multilingual music genres graphs with a modified retrofitting algorithm. Experiments show that our method: 1) is effective in translating music genres across tag systems in multiple languages (English, French and Spanish); 2) outperforms the previous baseline in an English-language multi-source translation task.

1. INTRODUCTION

Music genres are a key characteristic of music items [1, 2]. In music streaming services, user profiles and interests can be expressed through music genres, tracks and artists can be grouped in genre-specific collections, and content-based recommender systems frequently exploit music genres as item tags. However, music genres are difficult to infer due to their subjective nature. Based on their music preferences, musicological knowledge and culture, people inconsistently associate genres to music items [3–5]. Thus, annotating music items with genres for providing personalized recommendation and retrieval is challenging.

Acknowledging this subjectivity and the absence of a unique genre definition, recent works [6, 7] framed the music genre annotation as a *translation*. More precisely, given music items annotated with music genres originating from multiple source tag systems such as folksonomies, editorial vocabularies or taxonomies, the goal was to predict the equivalent music genres within a target tag system. In a supervised setup, the translation relied on a parallel corpus of music items jointly annotated with music genres from the source and target tag systems [6, 7]. In an unsupervised setup, when the parallel corpus was unavailable, a solution centered on taxonomy alignment was proposed [6].

However, the translation of music genres between *multilingual* sources remains unaddressed when a parallel corpus is unavailable. The only past unsupervised solution [6] relied on heuristics specific to the English language, making its adaptation to multilingual tags a challenge. Here, we propose to perform the unsupervised cross-lingual translation by leveraging multilingual music genre embeddings. Also, our method to learn these embeddings could be straightforwardly applied to new languages.

The proposed method is further summarised. First, by acknowledging the compositional nature of music genres (i.e. the meaning of multi-word music genres can be often derived from the meaning of each word), we learn music genre embeddings by applying compositionality functions to pre-trained word vectors [8–10]. Moreover, as these pre-trained vectors are often trained on language-specific text, we need to align them across languages [11, 12].

Second, we fit the obtained music genre embeddings into the music domain. The embeddings learnt on general-language corpora could sometimes be semantically ambiguous. For instance, *house* is closer to *building* than to *music* and *jazz* is more similar to *folk* than to *bebop* in fastText [8]. To tackle this problem, we create a music genre knowledge graph from multilingual DBpedia [13] that contains multilingual genres as nodes and exhibits different types of music genre relations through its edges. Then, we use *retrofitting* [14] to encode the relational knowledge from the semantic graph in the embeddings. In this work, we modify the original retrofitting algorithm [14] to distinguish between two types of relations: equivalence (e.g. *dnb* and *drum'n'bass*) and other types of relatedness such as sub-genres, derivative genres, fusion genres, stylistic origins. Besides, we use retrofitting to learn embeddings for music genres that do not exist in the pretrained embedding vocabulary by exploiting their graph relations (e.g. *ethnotronica* and *chillstep* are not in the pretrained fastText vocabulary).

We evaluate the proposed method in two experiments.

First, we collect a new parallel corpus of music items annotated with music genres in three languages (English, French and Spanish) and demonstrate the effectiveness of our method for unsupervised cross-lingual music genre translation. Second, we show that using the embeddings learnt with our method outperforms the previous baseline [6] in a music genre translation task between multiple English-language tag systems.

2. PROBLEM FORMULATION AND RELATED WORK

Annotating music items from song lyrics or audio content has concentrated significant research efforts in the music information retrieval community [7, 15, 16]. Most existing works fix a tag system and focus on general music genres like *jazz* or *pop* [3–5]. Nonetheless, the dissimilarity of music genre tag systems and their use in annotations has been recently put forward for consideration, together with the need to take into account tags with increased granularity [15, 17]. In this direction, two previous works [6, 7] have framed the music genre annotation as a tag translation task between various music genre tag systems.

Specifically, given a set $\mathcal{S}$ of *source tag systems*, $S = \cup_{E \in \mathcal{S}} E$ the union of all tags across all source tag systems, $\mathcal{P}$ the partitions of S and T a *target tag system*, the goal is to define a translation scoring $f : \mathcal{P}(S) \rightarrow \mathbb{R}^{|T|}$ which estimates a score for each target tag from a set of source tags drawn from S. While in this notation a tag system refers to a set of tags, more general representations such as music genre graphs or taxonomies can also include relations between tags [18–20].

Hennequin et al. [7] proposed two translation strategies, both relying on the existence of a parallel corpus of music items annotated with music genres. Epure et al. [6] addressed also the unsupervised case, when such a parallel corpus was absent, and designed a knowledge-based method to learn tag embeddings. This method relied on multiple building blocks corresponding to tag normalization, the construction of an integrated music genre graph bringing together all source and target tag systems, and a taxonomy alignment algorithm mapping each music genre on DBpedia [13] tags. The DBpedia-related building block yielded music genre vectors quantifying the relatedness of the tag under consideration to each DBpedia music genre. For translation, considering $\{s_1, \ldots, s_K\}$ source tags and any target tag t, f was computed using cosine similarity:

$$f_t(\{s_1, s_2, \ldots, s_K\}) = \sum_{k=1}^{K} \frac{{\mathbf{s}_k}^T \mathbf{t}}{||\mathbf{s}_k||_2 ||\mathbf{t}||_2}, \qquad (1)$$

where $\mathbf{s}_k$ and $\mathbf{t}$ are the vectors corresponding to each s_k, respectively t and $||\cdot||_2$ is the Euclidian L2-norm.

In the previous unsupervised work, Epure et al. [6] focused on English-language music genres, claiming that the extension of the knowledge-based method to include multilingual tag systems was feasible since it relied on multilingual DBpedia. While we agree that it is feasible, the extent to which the introduced method could be easily changed to support other languages is questionable. Both normalizing tags and mapping music genres into the DBpedia space rely on language-specific heuristics. For instance, in normalization, heuristics referring to the length of tokens were used. However, the average word length, hence what is considered as a short or medium-length token depends on the language [21]. Then, mapping music genres on English DBpedia genres is limiting because some tags may exist in two languages but not in the English DBpedia. Computing directly the degree of relatedness of a source tag to a target tag could be a better alternative.

3. A MULTI-STEP METHOD FOR LEARNING MUSIC GENRE EMBEDDINGS

In this work, we propose a method to learn multilingual music genre embeddings that can be easily extended to new languages and support cross-lingual translation. The first step is to deduce initial embeddings for multi-word music genres by leveraging pre-trained multilingual word embeddings (Section 3.1). However, directly using these music genre embeddings in cross-lingual translation is prone to under-perform because:

- the embeddings often correspond to the most common word senses (e.g. *country* can refer to *nations* or *rock* could be closer in meaning to *stone*) and they are not disambiguated against the music domain.
- some music genres could contain rare words which are absent from the pre-trained model vocabulary, resulting in potentially unknown tag embeddings.

To address these issues, we complement distributional concept representations with semantics from knowledge bases that expose concept relations. Thus, in a second step, we assemble a multilingual music genre graph (Section 3.2). Then, we adjust the initial tag embeddings to encode the tag relations from the collected graph, ensuring the domain adaptation. For this, but also to learn embeddings for concepts with unknown vocabulary words, we use *retrofitting* [14], which we modify to reflect the different types of music genre relations (Section 3.3).

3.1 Initializing Music Genre Embeddings

Under the music genre translation framework introduced in Section 2, the main goal boils down to quantifying the degree of relatedness of two textual tags. This task is widely popular in the natural language processing (NLP) community and contemporary approaches resort to expressing the relatedness as distance between corresponding word embeddings [8–10]. The mapping of words on embeddings is guided by the distributional hypothesis [22], which states that words in similar contexts are likely to have similar meanings. Word embeddings have been proven effective in capturing word syntactic and semantic similarities and in improving downstream NLP tasks such as natural language understanding [23] and information retrieval [24].

In order to measure the relatedness of multilingual words using embeddings learnt from monolingual corpora,

an alignment between the language-specific embedding spaces is required. Through the alignment [25], we ensure that multilingual word embeddings are projected into a common space where they are comparable. Practically, a mapping function between two monolingual word embedding spaces is learnt, for instance by using a bilingual lexicon [12]. Effective alignments have been also found using orthogonal Procrustes [11, 25].

Starting from multilingual word vectors, we discuss strategies to initialize the music genre embeddings. Music genres can contain multiple words. We claim that the compositionality principle, stating that the meaning of a multi-word expression is dictated by the meaning of each word, often holds for our case. For instance, *Dance pop* is related to *dance* and *pop* or *Balada romántica* is a type of *ballad* which is *romantic* [1] . The contemporary approach for compositional embeddings is to learn a function which derives the embeddings of a multi-word expression from the embeddings of its words [26]. The function is learnt by minimizing the distance for each multi-word expression between its distributional embedding and its embedding computed from its word embeddings. Obtaining the distributional embedding for multi-word music genres would be however challenging because sufficiently large corpora with all tags in multiple languages are required.

For this reason, the first music genre initialization strategy we propose consists of a simple compositionality function such as averaging word embeddings (*avg*). Let $V = \{c_1, c_2, ..., c_n\}$ be the multilingual vocabulary, c_i being a concept composed of at least one word. We aim to compute $\hat{\mathbf{Q}} \in \mathbb{R}^{n \times d}$, the embedding matrix for the vocabulary V, where $\hat{\mathbf{q}}_i \in \mathbb{R}^d$ denotes the embedding of concept c_i. If c_i is composed of the following words, $\{w_1, w_2, \ldots, w_M\}$, $\hat{\mathbf{q}}_i$ can be computed as $\frac{1}{M}\sum_{m=1}^{M} \mathbf{w}_m$, where $\mathbf{w}_m$ is the embedding of the word w_m. Of note is that if c_i contains words absent from the pretrained word embedding vocabulary, the d-dimensional null vector, $\mathbf{0}_d$, is used as a default.

The second music genre initialization strategy we propose exploits the fact that some words in a compounded expression may be more illustrative than others. The more frequently a word is observed in a corpus, the more likely it is that the word is common for a language and semantically less informative (e.g. *music* in *post industrial music*). Thus, the compositional embedding computation of a multi-word expression can be modified such that the contribution of each word embedding is inversely proportional to its frequency. Pre-trained word embeddings are generally released sorted by decreasing word corpus frequency. Let z_{w_m} be the rank of w_m in this vocabulary. Then, based on the Mandelbrot's generalization [27] of the Zipf's law [28], its frequency f_{w_m} can be estimated to $f_{w_m} = 1/(z_{w_m} + 2.7)$.

Further, we rely on the smooth inverse frequency (*sif*) based averaging proposed by Arora et al. [29] to compute the multi-word expression embeddings. This method is aligned with our previous observations and proven highly effective compared to more complex neural network-based models on a large diversity of NLP tasks [29]. Given f_{w_m} the estimated frequency of the word w_m and a a fixed hyper-parameter [2] , $\hat{\mathbf{q}}_i$ is computed as:

$$\overline{\mathbf{q}}_i = \frac{1}{M}\sum_{m=1}^{M} \frac{a}{a + f_{w_m}} \mathbf{w}_m \tag{2}$$

$$\hat{\mathbf{q}}_i = \overline{\mathbf{q}}_i - \mathbf{u}\mathbf{u}^T \overline{\mathbf{q}}_i \tag{3}$$

where $\mathbf{u}$ is the first singular vector from the singular value decomposition of $\overline{\mathbf{Q}}$ obtained with the Equation 2 [30].

3.2 Assembling a Multilingual Music Genre Graph

Previous related work [6] created a music genre graph by integrating multiple English-language music genre tag systems and a crawled sub-graph of DBpedia through a node English-language based normalization step. The other music genre tag systems were Lastfm, Tagtraum and Discogs, used in the 2018 MediaEval AcousticBrainz Genre Task [17]. Here, we bypass the language-specific heuristics normalization and propose a more robust alternative. We crawl a multilingual DBpedia music genre sub-graph and use its words as basis for normalizing new tag systems.

We further detail how we assemble the DBpedia-based music genres graph. We set the seeds for crawling from: 1) DBpedia entities of type *MusicGenre*, 2) the music genres of the multilingual DBpedia-based music item corpus (described in Section 4.1), 3) synonyms of the music genres of the previous two sources, linked through the *wikiPageRedirects* relation. We discover new potential music genres by crawling DBpedia entities linked to the seeds through one of the relations: *wikiPageRedirects*, *stylisticOrigin*, *musicSubgenre*, *derivative* and *musicFusionGenre* [3] . During crawling, seeds are updated with discovered entities that were not visited before, and the crawling goes on until no seeds are left. This is applied for each language. Finally, all music genres discovered as yet are connected to their equivalent tags in other languages, when possible (the DBpedia relation *sameAs*). In a post-processing step, we remove music genre nodes written as free-style text, which do not have DBpedia pages, and the connected components which do not contain at least one high-confidence music genre (empirically, we noticed that the highest-confidence tags were those from the music item corpus).

To ensure that tag systems with different music genre spellings benefit from the multilingual graph, we define a normalization which we apply to both the graph nodes and new tags. First, we tokenize each tag by non-alphanumeric characters. Further, as in [6], we create prefix trees to split multi-word tags such as *sludgemetal* or *indierock*. Nevertheless, we do not necessarily aim at a grammatically correct split, but at one based on lemmatized DBpedia music genre words [4] . Namely, if *sludgemetal* is already among

[1] Exceptions from the principle also exist (e.g. *hard rock*).

[2] Experimentally, it has been shown that $a = 10^{-3}$ is a suitable choice when using different types of pre-trained word embeddings [29].

[3] These relation names correspond to the English-language DBpedia. They have often translated versions in DBpedia in other languages.

[4] Through lemmatization, we retrieve the base form of inflected words using spacy (`https://spacy.io`). Most genre words are generally in their base form (e.g. *jazz*). However, some other words benefit from this (e.g. *Northern* / *North* or *children* / *child*)

the DBpedia music genre words, then there is no further split and we expect its initial embedding to be corrected though the embeddings of its graph neighbors as explained in the next section (Section 3.3).

3.3 Retrofitting Music Genre Embeddings

Retrofitting [14] has been proposed as a post-processing step to improve concept embeddings by leveraging existing semantic lexicons or knowledge graphs (e.g. WordNet [31]). More precisely, concept embeddings are modified to also encode concept relations [14, 32–34].

Let $G = (V, E)$ be the graph capturing the semantic relations between the nodes in V through a set of edges $E \subseteq V \times V$. The objective of retrofitting is to learn $\mathbf{Q} \in \mathbb{R}^{n \times d}$, the new concept embeddings, such that each new embedding $\mathbf{q}_i \in \mathbb{R}^d$ does not stray too far from the initial distributional embedding $\hat{\mathbf{q}}_i$, but also becomes closer to the new embeddings of the neighbour vertices $\mathbf{q}_j \in \mathbb{R}^d, j : (i, j) \in E$. The objective function to minimize is then [14]:

$$\Phi(\mathbf{Q}) = \sum_{i \in V} \left(\alpha_i ||\mathbf{q}_i - \hat{\mathbf{q}}_i||_2^2 + \sum_{j:(i,j) \in E} \beta_{ij} ||\mathbf{q}_i - \mathbf{q}_j||_2^2 \right) \tag{4}$$

where α_i and β_{ij} are positive scalars specifying the importance given to each component, the initial embedding and each graph neighbor. As Φ is convex with respect to $\mathbf{Q}$, a solution minimizing the objective function Φ is found in [14] via an iterative strategy derived from Jacobi iteration algorithm [35] that converges for such graph-based propagation problem [35, 36]. More precisely, until convergence, $\mathbf{q}_i$ is iteratively updated as follows:

$$\mathbf{q}_i \leftarrow \frac{\sum_{j:(i,j) \in E} (\beta_{ij} + \beta_{ji}) \mathbf{q}_j + \alpha_i \hat{\mathbf{q}}_i}{\sum_{j:(i,j) \in E} (\beta_{ij} + \beta_{ji}) + \alpha_i} \tag{5}$$

where $\mathbf{Q}$ is initialized to $\hat{\mathbf{Q}}$. Equation (5) is not the same as the original one [14]. We observed that, when applying the Jacobi method to optimize equation (4), the contributing terms in the partial derivative with respect to the node i are those where i appears as source as well as target node in the inner sum, leading to a different update. In [36, 37], the same conclusion referring to a corrected update rule, different from the initial proposal, is reached.

Faruqui et al. [14] set $\alpha_i = 1$ and $\beta_{ij} = \frac{1}{\text{degree}(i)}$ for $(i, j) \in E$, where degree(i) is the number of neighbors i has in the graph G, or 0 for $(i, j) \notin E$. This choice was largely adopted in other related works [32, 38]. Speer and Chin [39] proposed to use a modified version of retrofitting to learn embeddings for unknown vocabulary concepts which are present in the knowledge graph. For this case, α_i is set to 0 for all unknown vocabulary concepts. This results in $\mathbf{q}_i$ being updated by averaging the embeddings of its neighbours at each iteration. Despite the change in the update rule we made, compared to the original work, we retain this choice of hyper-parameters as being a reasonable default, and defer the investigation of a more principled way to pick α_i and β_{ij} to future work.

We further modify retrofitting to take advantage of the different types of music genre relations. On one hand, music genres can be semantically equivalent to other music genres (the relation types *wikiPageRedirects* and *sameAs*). On the other hand, music genres can be related to other music genres, but not semantically equivalent (e.g. *stylisticOrigin*). The change we propose for computing these new embeddings ($\mathbf{Q}_{\overline{\beta}}$) is through the coefficients β_{ij}, making them dependent on music genre relation types:

$$\overline{\beta}_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E_\epsilon \subset E \\ \beta_{ij} & \text{if } (i, j) \in E - E_\epsilon \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where E_ϵ contains edges which represent equivalence relations (*wikiPageRedirects*, *sameAs*); $E - E_\epsilon$ contains edges with the remaining relation types (*stylisticOrigin*, *musicSubgenre*, *derivative*, *musicFusionGenre*).

4. EXPERIMENTS

We evaluate the effectiveness of the learnt music genre embeddings, first, in a new cross-lingual music genre translation scenario (Section 4.3) and, second, in an existing English-language multi-source music genre translation task [6, 17] (Section 4.4). The languages we focus on for the cross-lingual annotation are English (**En**), French (**Fr**) and Spanish (**Es**). We start by presenting the parallel corpora used in the experiments (Section 4.1). Then, we discuss the detailed evaluation setup (Section 4.2). The results show that our music genre vectors are highly effective for cross-lingual translation and lead to improved results on the past unsupervised Enligh-language translation task [6].

4.1 Datasets

For the cross-lingual translation experiment, we relied on DBpedia [13] to collect a parallel corpus. During an initial manual analysis, we noticed that DBpedia music artists or works could have associated quite different music genres across languages. We present a few examples in Table 1. Also, when the tags used in annotations were equivalent, they were sometimes partially translated (e.g. *Rock_alternatif* in **Fr**), while other times they maintained the same form as in English (e.g. *Soft_rock* in **Es**). We collected DBpedia entities of type *MusicalArtist*, *Band*, or *MusicalWork* with music genres associated in at least two languages. Then, in a post-processing step, we filtered out the music items with tags that appeared less than 16 times.

For the English-language multi-source translation, we use an existing dataset [6, 17], which contains tracks annotated with English-language music genres from three sources. Discogs (**Dc**) tags are provided by editors per album, and automatically propagated to each track [17]. Lastfm (**Lf**) and Tagtraum (**Tt**) tags are created by Internet users per track. We show in Table 2 the number of music items and unique music genres in the new multilingual and the past English-language multi-source parallel corpora.

4.2 Evaluation Setup

We evaluated our models by translating music genre tags associated with tracks from multiple source tag systems to

Title	Type	En	Fr	Es
Morning View	Album	Alternative_metal, Funk_rock Alternative_rock, Post-grunge	Rock_alternatif	Metal_alternativo Rock_experimental
Jimi Hendrix	Artist	Hard_rock, Psychedelic_rock Blues, Rhythm_and_blues	Rock_psychédélique Blues_rock, Hard_rock	Blues_rock, Rock_psicodélico Hard_rock
Julio Iglesias	Artist	Dance-pop, Latin_music Adult_contemporary_music	Pop_française	Pop_latino, Balada_romántica Soft_rock, Adult_contemporary

Table 1. Examples of DBpedia music items annotated with music genres. The tag choices are inconsistent across sources. Tags may be adapted to a language (e.g. *Pop_latino* in **es**) or may keep the same form in all languages (e.g. *Hard_rock*).

	En	Fr	Es	Dc	Lf	Tt
Music items (tracks, albums, artists)	48 146	30 611	34 918	1 098 336	686 978	589 583
Unique music genres	489	338	491	315	327	296

Table 2. Number of music items and unique music genres in the multilingual and the English-language parallel corpora.

a target tag system. The translation scoring function computes a score for each tag of the target tag system as the degree of relatedness of the target tag to the input set of source tags. Compared to Equation 1, the translation scoring function we use here averages the cosine similarities between each source and target tag embeddings:

$$\hat{f}_t(\{s_1, s_2, \ldots, s_K\}) = \frac{1}{K} f_t(\{s_1, s_2, \ldots, s_K\}) \quad (7)$$

Like in other multi-label prediction tasks [15, 40], we use a ranking metric in evaluation, namely the area under the receiver operating characteristic curve (AUC). We macro-average the scores and report their mean and standard deviations computed over 4 folds. We split the multi-label data in a stratified way, balancing the overall number of music items and tag distribution across the folds [41].

For each experiment, English-language multi-source and cross-lingual, we have three input tag systems represented as partially aligned music genre graphs. For the multi-source translation, we assemble a graph from the English-language DBpedia music genre sub-graph and the input taxonomies, Discogs, Lastfm and Tagtraum. For the cross-lingual translation, the new music genre graph, which was assembled as described in Section 3.2, has 10748 tags in **En**, 2905 in **Fr** and 3988 in **Es**. The translation is performed using annotations from combinations of two out of three tag systems to the kept-out tag system. We also retain in evaluation annotations which are only from one of the two selected source tag systems.

In each experiment, we compare the *avg* and *sif* strategies to initialize the music genre vectors. We report results when using directly the initial embeddings ($\hat{\mathbf{Q}}$) in translation or retrofitted with the original method ($\mathbf{Q}$) or with our modified version ($\mathbf{Q}_{\overline{\beta}}$). As for the choice of pre-trained word embeddings, we use multilingual fastText [10] which we align with the method proposed by Joulin et al. [42].

4.3 Results on Cross-Lingual Genre Translation

In Table 3, we present the results of the cross-lingual music genre translation. The baseline we propose estimates the relatedness of two tags to be the length of their shortest path in the multilingual DBpedia-based music genre graph. As a reminder, this graph is partially aligned, meaning that some music genres have equivalent tags in other languages. As shown in Table 3 in parentheses, the baseline scores are quite high proving that the graph is fairly effective for cross-lingual translation on this dataset. Even so, we are able to exceed these scores by a large margin with our music genre embeddings, initialized with *sif* and retrofitted to take into account the music genre relations.

When comparing the initialization strategies, we can observe that directly using *sif* embeddings in translation outperforms the baseline, while *avg* yields lower AUC scores. For all languages as targets, the *sif* initialization is consistently more effective than the *avg* initialization. A significant difference between the two types of retrofitting applied to both initialization strategies exists, our version resulting in higher AUC scores. By differentiating between the two relation types, equivalence and other relatedness, the music genre embeddings appear to encode more accurately their relations within and across languages.

4.4 Results on Multi-Source Genre Translation

In Table 4, we present the results of the English-language multi-source music genre annotation. We re-compute the baseline [6] results using Equation 7. Compared to the previously reported AUC scores [6], the re-computed ones are higher showing that the modified translation scoring function does not disadvantage the knowledge-based music genre embeddings derived with the baseline. In contrast to the baseline, our most effective method, using *sif* initialization and our version of retrofitting, yields consistently higher AUC scores. The increase in performance is of 11.3 percentage points for **Dc** as target, 5.9 points for **Lf** as target and 9.3 points for **Tt** as target.

The *sif* initialization of tag embeddings results in higher AUC scores than *avg* both when the embeddings are used directly as they are or retrofitted, in particular, when **Dc** is target. Also, let us notice that directly using the embeddings initialized with *sif* leads to an increased performance compared to the baseline for **Dc** and **Tt**. Retrofitting the embeddings significantly increases the AUC scores for all

	Baseline	$\hat{\mathbf{Q}}$ (*avg*)	**Q** (*avg*)	$\mathbf{Q}_{\overline{\beta}}$ (*avg*)	$\hat{\mathbf{Q}}$ (*sif*)	**Q** (*sif*)	$\mathbf{Q}_{\overline{\beta}}$ (*sif*)
En + Es $\Longrightarrow$ Fr	85.4 ± 0.4	73.7 ± 0.2	77.5 ± 0.1	87.0 ± 0.2	85.9 ± 0.1	87.7 ± 0.1	$\mathbf{92.3 \pm 0.1}$
En + Fr $\Longrightarrow$ Es	84.3 ± 0.2	73.6 ± 0.3	76.1 ± 0.2	84.9 ± 0.2	85.6 ± 0.0	86.3 ± 0.1	$\mathbf{91.3 \pm 0.1}$
Fr + Es $\Longrightarrow$ En	80.4 ± 0.1	76.7 ± 0.4	84.2 ± 0.2	87.0 ± 0.3	84.5 ± 0.2	88.4 ± 0.3	$\mathbf{90.2 \pm 0.2}$

Table 3. Macro-AUC (%) in cross-lingual music genre translation with standard deviation computed over 4 folds. Results are shown for different embedding initialization (*avg* and *sif*), used directly ($\hat{\mathbf{Q}}$) or retrofitted with the original retrofitting (**Q**) or with our version ($\mathbf{Q}_{\overline{\beta}}$). The baseline is built on the shortest paths in the DBpedia-based multilingual graph.

	Baseline	$\hat{\mathbf{Q}}$ (*avg*)	**Q** (*avg*)	$\mathbf{Q}_{\overline{\beta}}$ (*avg*)	$\hat{\mathbf{Q}}$ (*sif*)	**Q** (*sif*)	$\mathbf{Q}_{\overline{\beta}}$ (*sif*)
Lf + Tt $\Longrightarrow$ Dc	76.2 ± 0.1	75.2 ± 0.2	82.0 ± 0.2	83.0 ± 0.2	81.3 ± 0.2	87.3 ± 0.1	$\mathbf{87.5 \pm 0.0}$
Dc + Tt $\Longrightarrow$ Lf	84.5 ± 0.2	81.6 ± 0.2	87.2 ± 0.1	88.0 ± 0.1	84.6 ± 0.1	90.1 ± 0.1	$\mathbf{90.4 \pm 0.1}$
Lf + Dc $\Longrightarrow$ Tt	82.5 ± 0.3	82.2 ± 0.3	87.8 ± 0.2	88.1 ± 0.2	86.4 ± 0.1	$\mathbf{91.6 \pm 0.2}$	$\mathbf{91.8 \pm 0.2}$

Table 4. Macro-AUC (%) in English multi-source music genre translation with standard deviation computed over 4 folds. Results are shown for different embedding initialization (*avg* and *sif*), used directly ($\hat{\mathbf{Q}}$) or retrofitted with the original retrofitting (**Q**) or with our version ($\mathbf{Q}_{\overline{\beta}}$). The baseline consists in tag alignment against English DBpedia music genres [6].

tag systems as targets. Compared to the experiments reported in Section 4.3, this time, we observe only a marginal difference between the original retrofitting and our version.

Further, we give more details about the translations enabled by the baseline and our retrofitted *sif* embeddings. Often, we yield better music genre mappings (e.g. we map *Discogs:uk garage* on *Tagtraum:garagerock*, while the baseline maps it on *Tagtraum:dubstep*). However, there are also cases where the baseline leads to more accurate mappings (e.g. *Discogs:modal* is mapped on *Lastfm:cooljazz* compared to our best mapping on *Lastfm:jazz*). Finally, the baseline could not map at all some music genres, while we could (e.g. we map *Discogs:crunk* on *Tagtraum:gangstarap* and on *Lastfm:rap*).

To sum up, exploiting the semantics of the music genre graph edges leads to marginally improved results w.r.t. the original retrofitting in the English-language multi-source translation and significantly higher AUC scores in the cross-lingual translation. The *sif* initialization yields better translations than the *avg* initialization. Lastly, we outperform the baselines by large margins in both experiments.

5. CONCLUSION

In this paper, we presented a new multi-step method for multilingual music genre embeddings learning. This method combines pre-trained word embeddings, music genre graphs and a retrofitting method leveraging different types of music genre relations to adapt embeddings to the music domain and learn embeddings for music genres with unknown words in the pre-trained word embeddings vocabulary. Our experiments demonstrate the effectiveness of the proposed method, both in the English-language multi-source and the new cross-lingual translation tasks.

For future work, we plan to learn embeddings for each music genre relation type. Fang et al. [33] consider that each edge represents a linear translation from the embedding of one node to the embeddings of its neighbour. In a generalized setup, functional retrofitting proposed by Lengerich et al. [32] defines a linear relational penalty function for each type of relation in the graph.

Then, we aim to address the incremental updates of the music genre graph in order to avoid re-applying retrofitting every time the graph is updated. Instead of relying on the constraints represented by the graph edges directly in retrofitting, Glavaš and Vulič [43] use them as training instances to learn an explicit retrofitting function, which can be after applied to new node embeddings.

Further, we want to apply our method to new languages, especially from other language families, as well as to investigate other pre-trained word embeddings and alternatives to embed multi-word concepts [26]. For this, the current music genre graph needs to be populated with new multilingual music genres and their relations, and a parallel corpus of music items covering new languages should be collected if further evaluation is required. Continuing to rely on the multilingual DBpedia is an option, though a limiting one, given that only some world languages are supported. Thus, music genre translation involving resource-poor languages remains a challenge. However, for the supported languages, our approach allows generating cross-lingual music genre annotations, which could be useful for other music information retrieval and recommendation tasks such as language-aware music genre autotagging, localized playlist captioning and music genre-driven recommendations, cross-cultural music genre perception modeling for user studies.

Finally, the multilingual data and the code to learn and evaluate music genre embeddings are made available to the community [5]. Also, a demo to visualize the music genre vector space and the cross-lingual translation results for DBpedia music items is available [44].

6. ACKNOWLEDGEMENTS

We would like to thank Manuel Moussallam, Marion Baranes, Anis Khlif and the ISMIR reviewers for their insightful and helpful comments on the paper.

[5] https://github.com/deezer/MultilingualMusicGenreEmbedding

7. REFERENCES

[1] M. I. Mandel, D. Eck, and Y. Bengio, "Learning tags that vary within a song," in *Conference of the International Society for Music Information Retrieval*, 2010.

[2] M. Schedl and B. Ferwerda, "Large-scale analysis of group-specific music genre taste from collaborative tags," in *IEEE International Symposium on Multimedia*, 2017.

[3] A. J. Craft, G. A. Wiggins, and T. Crawford, "How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems," in *Conference of the International Society on Music Information Retrieval*, 2007.

[4] J. H. Lee, K. Choi, X. Hu, and J. Downie, "K-pop genres: A cross-cultural exploration," in *Conference of the International Society on Music Information Retrieval*, 2013.

[5] M. Sordo, O. Celma, M. Blech, and E. Guaus, "The Quest for Musical Genres: Do the Experts and the Wisdom of Crowds Agree?" in *Conference of the International Society on Music Information Retrieval*, 2008.

[6] E. V. Epure, A. Khlif, and R. Hennequin, "Leveraging knowledge bases and parallel annotations for music genre translation," in *Conference of the International Society for Music Information Retrieval*, 2019.

[7] R. Hennequin, J. Royo-letelier, and M. Moussallam, "Audio based disambiguation of music genre tags," in *Conference of the International Society of Music Information Retrieval*, 2018.

[8] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *International Conference on Language Resources and Evaluation*, 2018.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.

[10] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *International Conference on Language Resources and Evaluation*, 2018.

[11] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

[12] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*. Springer, 2007, pp. 722–735.

[14] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.

[15] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text and images using deep features," in *Conference of the International Society on Music Information Retrieval*, 2017.

[16] E. Coviello, R. Miotto, and G. R. Lanckriet, "Combining content-based auto-taggers with decision-fusion," in *Conference of the International Society on Music Information Retrieval*, 2011.

[17] D. Bogdanov, A. Porter, J. Urbano, and H. Schreiber, "Mediaeval 2017 acousticbrainz genre task: content-based music genre recognition from multiple sources," in *MediaEval 2017 AcousticBrainz*, 2017.

[18] H. Schreiber, "Genre ontology learning: Comparing curated with crowd-sourced ontologies." in *Conference of the International Society for Music Information Retrieval*, 2019.

[19] M. Achichi, P. Lisena, K. Todorov, R. Troncy, and J. Delahousse, "Doremus: A graph of linked musical works," in *International Semantic Web Conference*, 2018.

[20] P. Lisena, K. Todorov, C. Cecconi, F. Leresche, I. Canno, F. Puyrenier, M. Voisin, T. Le Meur, and R. Troncy, "Controlled vocabularies for music metadata," in *Conference of the International Society on Music Information Retrieval*, 2018.

[21] R. D. Smith, "Distinct word length frequencies: distributions and symbol entropies," *Glottometrics*, vol. 23, pp. 7–22, 2012.

[22] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[23] J. Li and D. Jurafsky, "Do multi-sense embeddings improve natural language understanding?" in *Conference on Empirical Methods in Natural Language Processing*, 2015.

[24] C. V. Gysel, M. De Rijke, and E. Kanoulas, "Neural vector spaces for unsupervised information retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 4, pp. 1–25, 2018.

[25] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," in *Conference on Empirical Methods in Natural Language Processing*, 2018.

[26] V. Shwartz, "A systematic comparison of English noun compound representations," in *Joint Workshop on Multiword Expressions and WordNet*. Association for Computational Linguistics, 2019, pp. 92–103.

[27] B. Mandelbrot, "An informational theory of the statistical structure of language," *Communication theory*, vol. 84, pp. 486–502, 1953.

[28] G. K. Zipf, *Human behavior and the principle of least effort*. MA, Addison-Wesle: Cambridge, 1949.

[29] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *International Conference on Learning Representations*, 2017.

[30] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Springer, 1971, pp. 134–151.

[31] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[32] B. J. Lengerich, A. L. Maas, and C. Potts, "Retrofitting distributional embeddings to knowledge graphs with functional relations," in *International Conference on Computational Linguistics*, 2018.

[33] L. Fang, Y. Luo, K. Feng, K. Zhao, and A. Hu, "Knowledge-enhanced ensemble learning for word embeddings," in *World Wide Web Conference*, 2019.

[34] T. Scheepers, E. Kanoulas, and E. Gavves, "Improving word embedding compositionality using lexicographic definitions," in *World Wide Web Conference*, 2018.

[35] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003.

[36] Y. Bengio, O. Delalleau, and N. Le Roux, "Label propagation and quadratic criterion," *Semi-Supervised Learning*, pp. 193–216, 2006.

[37] T. K. Saha, S. Joty, N. Hassan, and M. A. Hasan, "Dis-s2v: Discourse informed sen2vec," *arXiv preprint arXiv:1610.08078*, 2016.

[38] D. Hayes, "What just happened? Evaluating retrofitted distributional word vectors," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[39] R. Speer and J. Chin, "An ensemble method to produce high-quality word embeddings," *arXiv preprint arXiv:1604.01692*, 2016.

[40] K. Ibrahim, J. Royo-Letelier, E. Epure, G. Peeters, and G. Richard, "Audio-based auto-tagging with contextual tags for music," in *International Conference on Acoustics, Speech, and Signal Processing*, ser. ICASSP, 05 2020, pp. 16–20.

[41] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2011.

[42] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

[43] G. Glavaš and I. Vulić, "Explicit retrofitting of distributional word vectors," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

[44] E. V. Epure, G. Salha, F. Voituret, M. Baranes, and R. Hennequin, "Muzeeglot: annotation multilingue et multi-sources d'entités musicales à partir de représentations de genres musicaux," in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4: Démonstrations et résumés d'articles internationaux*. ATALA, 2020, pp. 18–21.