DISCOURSE ANALYSIS OF LYRIC AND LYRIC-BASED CLASSIFICATION OF MUSIC

Jiakun Fang[1] **David Grunberg**[1] **Diane Litman**[2]
Ye Wang[1]
[1] School of Computing, National University of Singapore, Singapore
[2] Department of Computer Science, University of Pittsburgh, USA
fangjiak@comp.nus.edu.sg, wangye@comp.nus.edu.sg

ABSTRACT

Lyrics play an important role in the semantics and the structure of many pieces of music. However, while many existing lyric analysis systems consider each sentence of a given set of lyrics separately, lyrics are more naturally understood as multi-sentence units, where the relations between sentences is a key factor. Here we describe a series of experiments using discourse-based features, which describe the relations between different sentences within a set of lyrics, for several common Music Information Retrieval tasks. We first investigate genre recognition and present evidence that incorporating discourse features allow for more accurate genre classification than single-sentence lyric features do. Similarly, we examine the problem of release date estimation by passing features to classifiers to determine the release period of a particular song, and again determine that an assistance from discourse-based features allow for superior classification relative to single-sentence lyric features alone. These results suggest that discourse-based features are potentially useful for Music Information Retrieval tasks.

1. INTRODUCTION

Acoustic features have been used as the basis for a wide variety of systems designed to perform various Music Information Retrieval (MIR) tasks, such as classifying music into various categories. However, a piece of music is not entirely defined by its acoustic signal, and so acoustic features alone may not contain sufficient information to allow for a system to accurately classify audio or perform other MIR tasks [24]. This has led to interest in analyzing other aspects of music signals, such as lyrics [16, 22].

Although not all music contains lyrics, for songs that do, lyrics have been proven to be useful for classifying audio based on topic [17], mood [15], genre, release date, and even popularity [7]. This is a natural result since humans also consider lyrics when performing these classifications.

But while lyric features have been used in previous MIR studies, such works often use a bag-of-words or bag-of-sentences approach which considers each sentence within a set of lyrics independently. This approach sacrifices the contextual information provided by the lyrical structure, which often contains crucial information. As an example, we consider lyrics from the theme of Andy Williams' "A Summer Place":

- *Your arms reach out to me.*
- ***And*** *my heart is free from all care.*

The clause 'and' linking these two lines helps to set the mood; the listener can observe a connection between the subject reaching out to the singer, and the singer's heart consequently being at ease. But suppose the word 'and' were changed to the word 'but'. In this case, the meaning of these lyrics would be entirely different; now the singer's heart is at ease *despite* the subject reaching for him, not implicitly *because* of it. A human would no doubt observe this; however, this information would be lost with a bag-of-words or bag-of-sentences approach. We therefore hypothesize that lyrics features which operate on a discourse level, taking into account the relations between textual elements, will better represent the underlying structure of a set of lyrics, and that systems using such features will improve the performances of those using lyric features which consider each sentence independently.

In this paper we consider two classical MIR tasks: genre classification and release date estimation. Prior research has already demonstrated that lyrics-based features can improve accuracy for genre classification [22] as well as release date estimation [7]. This prior work considered individual words without taking into account how those words were linked together with discourse features or other connectors. However, it is already known that the complexity of lyrics often varies between different genres (e.g., rap music tends to have more complex lyrics than other genres [7]) as well as between different eras of music [9]. Lyrics of differing complexity are likely to have differing discourse connectors (e.g., very simple lyrics may only consist of a few unrelated elements and so have almost no discourse connectors, while dense, complicated lyrics may contain many elements which are connected together via discourse connectors), so we hypothesize that discourse

connector features may also contribute to the above tasks. As such, we investigate whether discourse features truly improve the accuracy in genre recognition, release-date estimation, and popularity analysis.

2. RELATED WORKS

Discourse analysis is a process analyzing the meaning of a text by examining multiple component sentences together, rather than each sentence on its own [26]. One dimension of it is *discourse relations*, which describes how multiple elements of a text logically relate to each other, and different discourse relation corpora and frameworks have been devised, including Rhetorical Structure Theory [21], Graphbank [27] and the Penn Discourse Treebank (PDTB) [25]. We opted to use PDTB as it is relatively flexible compared to these other frameworks [23] and more able to accommodate a wider variety of lyrics structures.

Another aspect of discourse analysis is text segmentation. In prior MIR studies involving lyrics, acoustic elements were used to help determine lyric segmentation points [3]. However, this approach takes the risk that errors in the audio analysis will propagate through to the lyric segmentation step. In contrast, the algorithm TextTiling takes only text as input and attempts to detect the boundaries of different subtopics within that text in order to perform meaningful segmentation [13]. Because lyrics can change topics during a song, we determined that a topic-based system like TextTiling could provide useful segmentation for MIR systems operating on lyrics.

Coherence and cohesion of a text has been proven to be important for human understanding [12] and writing quality [4]. While text coherence is a subjective property of text based on human understanding, text cohesion is an objective property of explicit text element interpretation patterns [12]. Various studies focused on elements of this specific text analysis, including entity grid [1] and coreference resolution systems [18]. A study by Feng et al. [8] showed the appearance pattern of entities may vary according to different writing style. Therefore, we hypothesize that the cohesion patterns in lyrics may vary according to different categories, and we used entity density, entity grid and coreference chain for lyric cohesion analysis.

Many music classification tasks have been investigated in the field of MIR. However, most systems which incorporate lyrics do not incorporate discourse analysis; they instead rely on approaches such as analyzing bags of words, part-of-speech tags and rhyme [7, 16, 19]. There was still little analysis of the discourse relations, topic shifts or detailed cohesion analysis.

3. FEATURES

3.1 Discourse-based Features

PDTB-styled discourse relations: We used a PDTB-styled parser [1] [20] to generate discourse relation features. In this work, we only focus on explicit discourse relations, since implicit relations are both harder to accurately determine and more subjective. In order to find such explicit relations, the parser first identifies all connectives in a set of lyrics and determines whether each one serves as a discourse connective. The parser then identifies the explicit relation the connective conveys. The system considers four general relations and 16 specific relations which are subcategories of the 4 general relations.

As an example, we consider a lyric from John Lennon's "Just Like Starting Over": "*... I know time flies so quickly/ **But when** I see you darling/It's like we both are falling in love again...*" All three of the underlined words are connectives, but the first such word, 'so,' is not a discourse connective because it does not connect multiple arguments. The parser thus does not consider this word in its analysis. The other two connectives, 'but' and 'when', are discourse connectives and so are analyzed to determine what type of relation they are; 'when' is found to convey a Temporal (general) and Synchrony (specific) relation, and 'but' is determined to convey a Comparison and a Contrast relation. In this way, the connections between the different elements of this lyric are understood by the system.

Once all the discourse connectives are found and categorized, we obtain features by counting the number of discourse connectives in each set of lyrics which corresponds to a particular discourse relation. For instance, one song might have 18 discourse connectives indicating a Temporal relation, so its Temporal feature would be set to 18. We also count the number of pairs of adjacent discourse connectives which correspond to particular relations and these adjacent discourse connectives are not necessary consecutive tokens; the same song as before might have 5 instances where one discourse connective indicates a 'Temporal' relation and the next discourse connective indicates a 'Comparison' relation, so its Temporal-Comparison feature would be set to 5. This process is performed independently for the general and the specific relations. Ultimately, we obtain 20 features corresponding to the 4 general relations (4 individual relations and 16 pairs of relations), and 272 features corresponding to the 16 specific relations (16 individual relations, and 256 pairs of relations). After removing features which are zero throughout the entire dataset, 164 features corresponding to specific relations remain. Finally, we calculate the mean and standard deviation of the sentence positions of all discourse connectives in a set of lyrics, as well as all connectives in that set of lyrics in general.

TextTiling segmentation: We ran the TextTiling algorithm to estimate topic shifts within a piece of lyric, using the Natural Language Toolkit Library [2], setting the pseudo-sentence size to the average length of a line and grouping 4 pseudo-sentences per block. Lyrics with fewer than 28 words and 4 pseudo-sentences were set as one segment, since they were too short for segmentation, and lyrics with no line splits were arbitrarily assigned a pseudo-sentence size of 7 words (average length in the dataset). Features were then calculated by computing the mean and

[1] http://wing.comp.nus.edu.sg/ linzihen/parser/

[2] http://www.nltk.org/api/nltk.tokenize.html

standard deviation in the number of words in a lyric's segments and the number of segments.

Entity-density features: General nouns and named entities (i.e., locations and names) usually indicate conceptual information. Previous research have shown that named entities are useful to convey summarized ideas [11] and we hypothesized that entity distribution could vary between song categories. We implemented features including: ratio of the number of named entities to the number of all words, ratio of the number of named entities to the number of all entities, ratio of the number of union of named entities and general nouns to the number of all entities, average number of named entities per sentence, and average number of all entities per sentence. We used OpenNLP [3] to find named entities and Stanford Part-Of-Speech Tagger [4] to extract general nouns.

Coreference inference features: Entities and their pronominal references in a text which represent a same object build a coreference chain [18]. The pattern of how an entity represented by different text elements with same semantic meanings through text may vary in different song styles. We used Stanford Coreference Resolution System [5] to generate coreference chain. The total number of coreference chains, the number of coreference chains which span more than half of lyric length, the average number of coreferences per chain, the average length per chain, the average inference distance per chain and the number of active coreference chains per word were extracted. The inference distance was computed as the minimum line distance between the referent and its pronominal reference. The chain is active on a word if the chain passes its location.

Entity-grid features: Barzilay and Lapata's [1] entity grid model was created to measure discourse coherence and can be used for authorship attribution [8]. We thus hypothesized that subjects and objects may also be related differently in different genres, just as they may be related differently for artists. Brown Coherence Toolkit [6] was used to generate an entity grid for each lyric. Each cell in a grid represent one of the roles of subject (S), object (O), neither of the two (X) and absent in the sentence (-) of a entity in a sentence. We calculated the frequency of 16 adjacent entity transition patterns (i.e., 'SS', 'SO', 'SX' and 'S-') and the number of total adjacent transitions, and computed percentage of each pattern.

3.2 Baseline: Previously Used Textual Features

We selected several lyric-based features from the MIR literature to form comparative baselines against which the discourse-based features could be tested (Table 1) [7]:

Vocabulary: We used the Scikit-learn library [6] to calculate the top 100 n-grams (n = 1, 2, 3) according to their tf-idf values. When performing genre classification, we obtained the top 100 unigrams, bigrams, and trigrams for the lyrics belonging to each genre. When performing year classification, we obtained approximately 300 n-gram features evenly from three year classes. These n-grams were represented by a feature vector indicating the importance of each n-gram in each lyric. We also computed the type/token ratio to represent vocabulary richness and searched for non-standard words by finding the percentage of words in each lyric that could be found in the Urban Dictionary [7], a dictionary of slang, but not in Wiktionary [8].

Part-of-Speech features: We used Part-of-Speech tags (POS tags) obtained from the Stanford POS Tagger [9] to determine the frequencies of each super-tags (Adjective, Adverb, Verb and Noun) in lyrics.

Length: Length features such as lines per song, tokens per song, and tokens per line were calculated.

Orientation: The frequency of first, second and third pronouns as well as the ratio of self-referencing pronouns to non-self-referencing ones and the ratio of first person singular pronouns to second person were used to model the subject of given sets of lyrics. We also calculated the ratio of past tense verbs to all verbs to quantify the overall tense of songs.

Structure: Each set of lyrics was checked against itself for repetition. If the title appeared in the lyrics, the title feature for that song was given a 'True' value, which was otherwise set to false. Similarly, if there were long sequences which exactly matched each other, the 'Chorus' feature was set to 'True' for a given song. Table 1 shows the number of elements in each feature set in the classification tasks.

Dimension	Abbreviation	Length
discourse-based features	DF	250
PDTB-based discourse relation	DR	204
TextTiling segmentation	TT	3
entity density	ED	5
coreference inference	CI	5
entity grid	EG	33
textual baseline features	TF	318
vocabulary	VOCAB	303
POS tags	POS	4
length	LEN	3
orientation	OR	6
structure	STRUC	2

Table 1: Features used in classification tasks.

3.3 Normalization

Since features used for these tasks are not on the same scale, we then performed normalization on features. Each feature was normalized by its maximum value and minimum value to range from 0 to 1 (Equation 1). Then all normalized features were put into classification tasks. This normalization step was expected to improve the results of

[3] https://opennlp.apache.org
[4] http://nlp.stanford.edu/software/tagger.shtml
[5] http://nlp.stanford.edu/projects/coref.shtml
[6] http://scikitlearn.org/stable/modules/feature_extraction.html
[7] http://www.urbandictionary.com
[8] https://www.wiktionary.org
[9] http://nlp.stanford.edu/software/tagger.shtml

combination of different feature sets, as differences in variable ranges could potentially affect negatively to the performance of classification algorithm.

$$v_n = \frac{v - v_{min}}{v_{max} - v_{min}} \quad (1)$$

4. DATASET AND ANNOTATION

A previously collected corpus of 275,905 sets of full lyrics was used for these experiments and we pre-processed the dataset in 6 different types to clean up lyrics [5], including splitting of compounds or removal of hyphenated prefixes, elimination of contractions, restoration of dropped initial, abbreviation elimination, adjustment to American English spellings, and correction of misspelled words. Unlike other corpora, such as musiXmatch lyrics dataset for the Million Song Dataset [2], lyrics from the selected corpus are not bags-of-words but are stored in full sentences, allowing for the retention of discourse relations. We split song lyrics by punctuations and lines to make sentences and paragraphs to run discourse analysis algorithm in this work. We also downloaded corresponding genre tags and album release years for the songs represented in this dataset from Rovi [10]. The specific number of lyrics for each experiment is shown in Table 2.

Genre classification: We kept all 70,225 songs with a unique genre tag from Rovi for this specific task. The tags indicated that songs in the dataset came from 9 different genres: Pop/Rock (47,715 songs in the dataset), Rap (8,274), Country (6,025), R&B (4,095), Electronic (1,202), Religious (1,467), Folk (350), Jazz (651) and Reggae (446). All of these songs were then used for the genre classification experiments.

Release date estimation: Rovi provided release dates for 52,244 unique lyrics in the dataset. These release dates ranged from 1954-2014. However, some genres were not represented in certain years; no R&B songs, for instance, had release dates after 2010, and no rap songs had release dates before 1980. To prevent this from biasing our results we chose to just use one single genre and settled on Pop/Rock, for which we had 46,957 songs annotated with release dates throughout the 1954-2014 range. We then extracted all the songs labeled as having been released in one of three time ranges: 1969-1971 (536 songs total), 1989-1991 (3,027), and 2009-2011 (4,382). We put gaps of several years between each range on the basis that, as indicated in prior literature, lyrics are unlikely to change much in a single year [7].

5. GENRE CLASSIFICATION

We ran SVM classifiers using 10-fold cross-validation. These classifiers were implemented with Weka [11] using the default settings. We chose SVM classifiers because they have been proven to be of use in multiple MIR tasks [7, 15]. Because each genre had a different number of samples, undersampling [10] was performed for both training and testing to ensure that each genre was represented equally before cross-validation classification. Each song was classified in a 2-class problem: to determine if the song was of the correct genre or not. The undersampling and classification process was repeated 10 times and we present the averages of F-score for each independent classification task. The value of F-score by random should be 0.5.

Classification Task	**Number of lyric used (after undersampling)**
Genre	Pop/Rock: 45,020; Rap: 16,548; Country: 12,050; Jazz: 1,302; R&B: 8,190; Electronic: 2,404; Religious: 2,934; Folk: 700; Reggae: 892
Release Period	1,608 sets of lyrics, split evenly into three time spans

Table 2: Data sizes for experiments.

We first implemented previously-used textual features to generate a baseline for the genre classification task. Models were built based on vocabulary (VOCAB), POS tags (POS), length (LEN), orientation (OR), structure (STRUC) and all combined baseline features (TF) separately. The average F-scores are depicted in Table 3. It is apparent that using vocabulary features can achieve high performance in average, but one thing to be noted is that it heavily depends on which corpus the language model trains on to generate the n-gram vector. Here we used all lyrics from each genre to get top n-grams. Orientation features were useful for R&B recognition since we found more first pronouns in such genre. We then used these features to compare with proposed discourse-based features.

We then evaluated the utility of discourse-based features for this specific task. Table 3 presents the results from using discourse relation (DR), TextTiling topic segmentation (TT), entity density (ED), coreference inference (CI), and entity grid (EG) features to perform genre classification with the SVM classifiers. Because the discourse relation and TextTiling features showed very promising results, we also tested a system which combined those features (DR+TT). Finally, we tested all discourse features together (DF), and then all discourse and all baseline features together. Statistical significance were computed using a standard two-class t-test between the highest F-score and each result from other feature set for each genre, and each column's best result were found to be significant with $p < 0.01$.

First, we note that, for every single genre as well as the overall average, the system's classification accuracy when using DR+TT discourse features is better than its accuracy using any and all baseline features. In fact, DR features alone outperform any and all baseline features for 7 of the 9 genres as well as overall. This serves to demonstrate the utility of these particular discourse features for

[10] http://developer.rovicorp.com
[11] http://cs.waikato.ac.nz/ml/weka

Feature Set	R&B	Folk	Country	Rap	Elect.	Reli.	Jazz	Reggae	Pop	Avg.
VOCAB	58.5	51.4	59.4	**90.8**	53.7	53.5	55.3	60.7	65.7	61.0
POS	55.4	47.3	53.6	73.1	49.9	50.3	56.3	47.4	60.0	54.8
LEN	55.2	49.3	55.4	85.8	48.6	50.0	50.3	48.8	59.2	55.4
OR	66.0	54.7	58.1	84.6	54.4	52.6	58.7	54.9	63.4	60.8
STRUC	45.0	46.4	44.5	45.6	46.0	45.7	45.3	47.0	44.6	45.6
TF (All)	62.5	56.5	60.1	81.3	50.7	51.8	58.1	56.5	63.6	60.1
DR	64.9	**61.7**	65.7	89.8	**59.1**	**56.2**	**62.8**	**64.0**	66.7	**65.7**
TT	63.3	51.1	58.2	90.4	53.1	53.0	58.0	55.9	65.9	61.0
ED	55.4	58.3	53.2	76.5	53.8	53.7	46.8	57.1	61.2	57.3
CI	59.1	47.8	62.7	82.4	50.5	52.8	55.7	54.1	63.7	58.8
EG	58.7	48.3	57.1	83.9	50.5	52.6	54.9	51.4	62.9	57.8
DR + TT	**67.4**	59.1	**66.6**	**91.0**	58.3	55.3	62.3	62.3	**67.7**	**65.6**
DF (All)	58.2	53.3	60.9	75.8	49.9	54.0	57.5	49.1	61.5	57.8
All	50.0	34.5	35.7	49.6	45.2	48.3	41.1	49.4	45.8	44.4

Table 3: Accuracy of classifier using different unnormalized feature sets to estimate genre (F-Score*100).

Feature Set	R&B	Folk	Country	Rap	Elect.	Reli.	Jazz	Reggae	Pop	Avg.
VOCAB	59.3	55.6	61.0	91.3	52.7	63.0	61.8	65.1	66.5	64.4
POS	63.5	57.8	55.9	90.9	49.4	48.9	61.8	61.6	65.3	62.4
LEN	61.9	50.5	59.4	86.7	49.2	49.1	61.1	59.4	63.5	60.2
OR	68.2	55.8	55.1	85.4	47.3	46.6	60.0	55.7	64.3	60.4
STRUC	46.9	45.1	45.8	45.8	46.9	44.8	43.8	47.1	44.6	45.6
TF (All)	71.1	59.6	67.4	93.3	55.4	65.0	65.6	68.7	68.3	68.4
DR	60.9	59.0	62.3	88.4	54.9	54.6	61.1	61.0	64.7	63.1
TT	64.1	49.8	54.6	90.9	48.7	51.0	62.7	60.6	66.0	61.7
ED	37.5	45.2	38.3	65.5	45.1	45.5	47.8	47.3	51.6	48.2
CI	63.5	53.2	61.5	84.5	50.5	55.1	62.2	63.7	62.2	61.9
EG	63.7	55.5	64.5	94.1	57.8	49.5	65.5	62.1	64.4	64.1
DF (All)	71.2	**61.3**	67.3	**94.5**	**58.5**	58.5	64.5	66.5	66.3	67.7
All	**73.7**	60.6	**71.5**	**94.8**	**58.9**	**65.6**	**66.9**	**69.6**	**69.4**	**69.9**

Table 4: Accuracy of classifier using different normalized feature sets to estimate genre (F-Score*100).

this task, since they consistently outperform the baseline features. Second, we note that the entity and coreference features did not enable the classifier to achieve maximal results in this task, indicating that these features may not vary as much between genres compared to the DR and TT features. Third, we note that the system's accuracy when all features was used decreased relative to the DR+TT and DR features in every case. We then performed the normalization and each feature was normalized by its maximum value and minimum value to range from 0 to 1.

Table 4 shows the results and the combination of all feature outperformed all baseline features, while the combination of all discourse-based features can achieve higher performance than all baseline feature sets in 3 classes. Best result for each genre were found to be significant with $p < 0.01$. This further emphasized the importance of discourse-based features in this specific task.

One interesting trend in these results is in the 'Rap' column, which shows that not only was the classification accuracy for Rap songs far higher than the other classes, but it was also the one genre where TT features outperformed DR features. Although the discourse-based features did not outperform the baseline features in this genre, it should be noted that the TextTiling segmentation features did obtain virtually identical performance to the best baseline features with only a 3-dimensional feature vector; the VOCAB features, by contrast, encompassed hundreds of dimensions. We investigated this further and found that Rap music tended to have more topic segments (5.9/song on average, while the average for other genres was 4.9), and more varied adjacent discourse relations as well (for instance, each rap song had on average 6.6 different types of adjacent discourse relations; non-rap songs averaged 4.0). This suggests that TextTiling segmentation features may be a more compact way to accurately represent topic-heavy lyrics, such as those commonly found in rap music.

We finally analyzed the portion of each type of discourse connective for the four first-level PDTB-styled discourse relations of all discourse connectives in each genre. We found that Religious songs use more expansion relations than other genres (42% and 37% in average), while less expansion relations are written in Rap songs (34%).

Connectives standing for temporal relations present more in Rap songs (26% and 23% in average). R&B songs contains more contingency connectives (24% and 26% in average).

6. RELEASE DATE ESTIMATION

We investigated whether discourse-based features can help to estimate the release date of a song, on the basis that the lyric structure of song texts is likely to change over time [7, 14]. We first formed a subset of all the Pop/Rock songs in our dataset, since as mentioned before these songs spanned a greater time period than the other genres. We then extracted all the songs labeled as having been released in one of three time ranges: 1969-1971 (536), 1989-1991 (3,027), and 2009-2011 (4,382). Based on the idea from prior study [7], we made gaps since that the lyrics would be unlikely to change very much in a single year. Undersampling was used to balance the dataset building a sub-dataset before each classification with an SVM with 10-fold cross validation for three-class classification. The process was repeated 10 times.

Table 5 shows results. As can be seen from the table, discourse relation features alone outperformed the baseline feature sets in average F-score for each three year class ($p < 0.001$), which indicates that the sentence relations in lyrics likely vary over years, and that discourse relation features are useful at indicating this. Although not as much as the discourse relation features, the topic segments and coreference inference features contribute to this specific classification task as well, showing topic presentation and cohesion structure changed over time. TextTiling features proved to increase accuracy for one year range, 2009-2011, indicating that the number and relations of topics of music released in this era likely varied as compared to previous eras, and also that text segmentation-based features are useful in noting this change. The number of topics and the number of words in each topics in average increases over time. As for the coreference inference features, the number of coreference chains and the number of long coreference chains showed raising values according to release periods. More coreference chains and long coreference appeared more often in the recent years, indicating a fluent and centric content. The other discourse features were again shown to be less useful than these ones. Finally, the early ages and recent ages were more likely to be recognized, while the middle ages generally achieved the lowest F-scores among all feature sets except structure features. This result is intuitive; music will likely be more similar to music that were produced closer together.

We then normalized to 0 to 1 for all features and repeated the task to show whether discourse features can improve the performance of baseline features for this task. Table 6 shows that the combination of all features outperformed the other feature sets in this three-class classification task ($p < 0.001$).

Feature	1969-1971	1989-1991	2009-2011	Avg.
VOCAB	46.8	33.7	34.9	38.5
POS	30.0	24.5	52.8	35.8
LEN	34.6	26.7	50.6	37.3
OR	43.4	32.0	50.6	42.0
STRUC	0.00	29.1	50.7	26.6
TF (All)	42.2	27.6	53.6	41.2
DR	**59.7**	**43.0**	55.0	**52.6**
TT	46.5	34.8	47.6	43.0
ED	40.4	29.5	41.7	37.2
CI	47.7	29.3	53.8	43.6
EG	41.2	32.5	44.3	39.4
DR + TT	58.5	40.7	**56.3**	51.8
DF (All)	43.3	28.3	53.8	41.8
All	36.2	30.6	30.4	32.4

Table 5: Accuracy of classifier using different unnormalized feature sets to estimate release date (F-Score*100).

Feature	1969-1971	1989-1991	2009-2011	Avg.
VOCAB	51.4	41.6	42.3	45.1
POS	58.7	24.5	46.7	43.3
LEN	61.4	27.9	45.8	45.0
OR	58.1	17.4	48.3	41.3
STRUC	0.0	22.0	**87.3**	36.4
TF (All)	**63.4**	42.0	53.1	52.8
DR	57.6	34.5	47.7	46.6
TT	59.9	29.9	37.8	42.5
ED	30.0	16.3	47.4	31.2
CI	62.0	27.2	52.3	47.2
EG	57.4	46.6	42.0	48.7
DF (All)	57.0	44.9	48.8	50.3
All	61.0	**48.8**	54.7	**54.7**

Table 6: Accuracy of classifier using different normalized feature sets to estimate release date (F-Score*100).

7. CONCLUSION AND FUTURE WORK

We investigated the usefulness of discourse-based features and demonstrated that such features can provide useful information for two MIR classification tasks. Genre classification and release date estimation were all enhanced by incorporating discourse features into the classifiers. However, since discourse-based features rely on passages with multiple text elements, it may be noisy when used on music with short lyrics. As this work is an exploration work, further analysis is required. For instance, we split song lyrics by lines and punctuations in this work, which fitted most of the cases in our dataset. The split rules of sentences can influence the results from discourse analysis algorithms.It will be potentially useful to use these features for other MIR tasks such as keyword extraction and topic classification. In the future, we will explore all these discourse-based features on other MIR tasks and find sensible sets of features and fusion strategies for further improving performance for these tasks.

8. REFERENCES

[1] R. Barzilay and M. Lapata. Modeling local coherence: an entity-based approach. *Computational Linguistics*, 34(1):141–148, 2008.

[2] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 591–596, 2011.

[3] H. T. Cheng, Y. H. Yang, Y. C. Lin, and H. H.Chen. Multimodal structure segmentation and analysis of music using audio and textual information. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 1677–1680, 2009.

[4] S. A. Crossley and D. S. Mcnamara. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 984–989, 2010.

[5] R. J. Ellis, Z. Xing, J. Fang, and Y. Wang. Quantifying lexical novelty in song lyrics. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 694–700, 2015.

[6] M. Elsner, J. Austerweil, and E. Charniak. A unified local and global model for discourse coherence. In *Proceedings of the Conference on Human Language Technology and North American Chapter of the Association for Computational Linguistics*, pages 436–447, 2007.

[7] M. Fell and C. Sporleder. Lyrics-based analysis and classification of music. In *Proceedings of International Conference on Computational Linguistics*, pages 620–631, 2014.

[8] V. W. Feng and G. Hirst. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29(2):191–198, 2014.

[9] Y. Gao, J. Harden, V. Hrdinka, and C. Linn. Lyric complexity and song popularity: Analysis of lyric composition and relation among billboard top 100 songs. In *SAS Global Forum*, 2016.

[10] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Chapman & Hall, London, 2011.

[11] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 121–128, 1999.

[12] A. C. Graesser, D. S. Mcnamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers*, 36(2):193–202, 2004.

[13] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[14] H. Hirjee and D. G. Brown. Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 5(4):121–145, 2010.

[15] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 619–624, 2010.

[16] X. Hu, J. S. Downie, and A. F. Stephen. Lyric text mining in music mood classification. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 411–416, 2009.

[17] F. Kleedorfer, P. Knees, and T. Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 287–292, 2008.

[18] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, 2011.

[19] T. Li and M. Ogihara. Music artist style identification by semi-supervised learning from both lyrics and content. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 364–367, 2004.

[20] Z. Lin, H. T. Ng, and M. Y. Kan. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184, 2014.

[21] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[22] R. Neumayer and A. Rauber. Integration of text and audio features for genre classification in music information retrieval. In *Proceedings of the European Conference on Information Retrieval*, pages 724–727, 2007.

[23] J. P. Ng, M. Y. Kan, Z. Lin, V. W. Feng, B. Chen, J. Su, and C. L. Tan. Exploiting discourse analysis for article-wide temporal classificatio. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12–23, 2013.

[24] F. Pachet and J. J. Aucouturier. Improving timbre similarity: How high is the sky. *Journal of Negative Results in Speech and Audio Sciences*, 1(1):1–13, 2004.

[25] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. The penn discourse treebank 2.0. In *Proceedings of Language*

Resources and Evaluation Conference, pages 2961–2968, 2008.

[26] B. Webber, M. Egg, and V.Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 18(4):1–54, 2012.

[27] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.