DISCOURSE ANALYSIS OF LYRIC AND LYRIC-BASED CLASSIFICATION OF MUSIC

JIAKUN FANG

B.Eng., Northeastern University of China

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE

2016

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Fang Jiakun

Jiakun Fang
November 2016

Acknowledgments

First, I would like to deeply thank my supervisor, Professor Ye Wang, for providing me with valuable guidance in my research. He has given me insightful advice at every stage of the writing of this thesis and he also supported me and helped me in various aspects. His passion on academics and works enlightened me not only on my research work, but also on my everyday life. I am also grateful to Professor Diane Litman for her guidance on my research and providing many useful suggestions on my works. I enjoyed the time in Sound and Music Computing lab. I want to thank Boyd Anderson, Chitralekha, Zhiyan Duan, David Grunberg and Shenggao Zhu for their collaborations and help. And I would also like to thank the School of Computing for giving me the active study environment and providing financial support. I would like to express my special acknowledgement to my parents for their love and support. Last but not least, I'd like to thank all my friends for their encouragement.

Contents

Summary

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 collaborating 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 incorporating discourse-based features allow for superior classification relative to only using single-sentence lyric features. Finally, we perform popularity analysis by comparing Billboard Magazine's "All-Time Top 100" songs and non-top songs, and show results indicating that the patterns of discourse features are different in the two song sets. These results suggest that discourse-based features are potentially useful for Music Information Retrieval tasks.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

With the expanding number of accessible music online, people can easily and conveniently search for their favorite music on the Internet. Current music online providers, such as YouTube[1], Spotify[2], Apple Music[3] and last.fm[4], present different music search schemas according to genres, moods or release years. These high-level descriptors are summaries of some shared properties of clustering music products, and by following such organization schema users can save effort and time to search related music collections. However, even though descriptors of music are widely used in music search, the labels are mainly from human annotation. Music providers can release metadata of music, such as its title, artist and album information. These data is reliable but often hard to be used for vague music search. On the other hand, common users can provide personal tags of music, which is more likely to include useful categories like mood but is also less reliable and noisier. With millions of music online, it requires great consumption of time and human resources to manually label music. For example, a work from Dannenberg et al. reported it took 30 full-time musicologists about one year to annotate hundred-thousand songs [DFTW01]. Therefore, automatic music annotation is very important at this stage.

The automatic music annotation task can be simplified to classify music into various categories, such as by different genres, moods or release decades.

[1] https://www.youtube.com

[2] https://www.spotify.com

[3] http://www.apple.com/music

[4] http://www.last.fm

Such music classification task is tightly related to the automatic feature analysis of music, which tries to extract features that can represent meaningful information of music products. Acoustic features, which are based on musical audio, have been used as a basis for a wide variety of Music Information Retrieval (MIR) tasks, including automatic categorization. However, the audio signal cannot describe a piece of music entirely and may not contain sufficient information to allow for a system to achieve a high performance for music classification or perform other MIR tasks. For instance, a seminal work by Pachet and Aucouturier showed a "glass ceiling" effect in spectral-based method on a timbre similarity measure, indicating acoustic features alone had a limited maximal performance and "high-level" features, which can match human cognitive processes, should be underlying factors [PA04]. This result has led to interest in analyzing music in other aspects of music, such as song lyrics. Song lyrics can include information that is not contained in acoustic signals [LGH08]. For example, the mood of a song is not solely perceived from the audio. If we combine jaunty tune to an ironic lyric, the entire song may sound sad.

Although not all music contains singing and lyrics, for songs that do. A piece of lyric usually conveys the semantic content of a song which helps in better understanding of the song [LGH08]. Previous works have used lyric-based features for multiple MIR sub-tasks, such as genre classification [LOL03, MNR08, FS14], mood classification [LGH08, HDE09, HD10], artist classification [LO04] and best/worst song classification [FS14]. Furthermore, when some works compared lyric-based features and acoustic features for some specific MIR sub-tasks such as mood classification, the results showed lyric-based features can complement an audio-based system or even outperform acoustic features in some MIR tasks [LO04, HD10]. The results should not be surprising, since humans also consider lyrics when performing music tagging [PL13]. The association between mood categories and semantic terms in song lyrics, for example, is easier to be recognized than audio signals. A song which frequently references 'love', 'joy' and 'birth' would likely be classified by humans as having a different mood, for example, than a song with similar acoustic features but lyrics referencing 'death' and 'misery'.

But while lyric features have been used for multiple MIR tasks, previous works usually use a bag-of-words or bag-of-sentences approach, which considers each word or sentence within a piece of lyric separately and individually. 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 as follows:

Your arms reach out to me.
But my heart is free from all care.

The meaning of these lyrics is entirely different, as 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 lyric-based features which operate on a higher level as a discourse level [WEK12], taking into account the relations between textual elements, will better represent the structure of a set of lyrics, and that systems using such features will outperform those using lyric features which consider each sentence independently.

To verify our hypothesis, we consider three classical MIR tasks: genre classification [SZM06], release year estimation [SCSU13, BMEWL11] and popularity analysis [EXFW15, HDWE05]. Although prior research has already demonstrated that lyric-based features have the potential to provide useful information to systems performing these tasks [FS14], detailed discourse-level features have not been considered. As a corpus, we draw from a previously-collected dataset of lyrics [EXFW15] from which discourse features were accessible.

Finally, we clarify our terms in this thesis. First, although we focus on songs, which are belong to a specific type of music that contains singing and a lyric, we use "music classification" instead of "song classification" here to be consistent with MIR community. Second, we does not distinguish sub-tasks and tasks of MIR in the following chapters for automatic categorization tasks such as genre classification, since the mentioned tasks are common and crucial tasks in MIR research and some are included in Music Information Retrieval Evaluation eXchange(MIREX) [5].

1.2 Contributions

The main contributions of this thesis are listed as follows:

- To the best of our knowledge, this is the first work in MIR tasks to analyze song lyrics on a detailed discourse level, which considers contextual elements and relations between sentences within lyrics. Three discourse-level feature sets as topic structure, discourse relation and text coherence and cohesion analysis were measured on the sub-dataset of our entire dataset, which contained songs with annotated tags from music providers. The text analysis techniques were carefully selected and parameterized according to the characteristics of song lyrics.
- Experimental results from all three classical MIR tasks, namely genre classification, release date estimation and popularity analysis, showed our proposed discourse-level features were potentially useful for MIR tasks. Specifically, the discourse-level features beat the traditional bag-of-words/bag-of-sentences features in genre classification and release date estimation tasks. For popularity analysis, discourse-level features showed different patterns in the most popular songs and the rest of songs in our dataset, indicating the possibility of using such lyric-based discourse-level features for popular song annotation.

1.3 Organization

The rest of the thesis is organized as follows:

- Chapter 2 introduces related works and background knowledge from Music Information Retrieval and Natural Language Processing aspects,

[5] https://en.wikipedia.org/wiki/Music_information_retrieval

including related MIR tasks, representative features of music, and discourse analysis of texts.

- Chapter 3 describes our proposed lyric-based discourse-level features including TextTiling algorithm for topic structure analysis, PDTB-styled discourse relations, and entity-based text coherence and cohesion analysis. The exploited features and the extraction algorithms are explicitly described.
- Chapter 4 details the entire experiment design, including the task selection, the music classification process and the feature analysis systems, and experiment features and dataset.
- Chapter 5 presents results of the experiments. The experiment results indicates the usefulness of the proposed lyric-based discourse-level features in three MIR tasks: genre classification, release date estimation and popular song analysis.
- Chapter 6 summarizes the work and its limitations; and finally Chapter 7 suggests possible future works.

Chapter 2

Literature Survey

In this chapter, we present a detailed literature survey on related works and background knowledge in Music Information Retrieval and Natural Language Processing research.

People often search a set of music by categories sharing similar properties. Real music products are usually mixtures of acoustic components, such as melodic lines and instruments, which make it hard to find direct one-to-one correlations between a musical characteristic and a music category. In other words, it is not easy to tell which characteristic or which combination of characteristics makes a music product having a common summary tag. In this case, although musicians and experts can make the ground truth of music annotations, it is a difficult task for common music listeners to perform a consistent and accurate tagging [Lam08]. One of the possible reasons is that a music classification is often subjective and fuzzy and severely depends on the aesthetics of music. To make it worse, professional and systematic human annotation is unable to meet the requirement of millions of music products online. Therefore, an automatic music annotation or music classification is useful to solve this problem.

To achieve high performance of music classification, apart from classification algorithms, an accurate representation of music is another crucial factor. Irrelevant or insufficient music representative features would lead to low performances of tasks. Therefore, it is useful to find a group of suitable representations to help people to understand which characteristics claim the class correlation. Although common representative features of music are usually summarized from audio signals, describing how music audio is organized [CVK04], the seminal research from Pachet and Aucouturier showed properties from other aspects at a higher level would be complements to features extracted from audio signals [PA04]. Hence, an investigation of suitable music representation from aspects such as song lyrics is potentially useful for MIR tasks.

Although a variety of works have investigated MIR tasks such as genre classification, release year classification and popularity analysis, as far as we concern, little research has systematically considered a higher level of song lyrics such as features from a discourse level. Since previous works showed the usefulness of song lyrics in the three tasks, we thus hypothesize discourse-based features considering internal relations between sentences and whole viewing of a text can be helpful in these tasks.

In this section, we first introduce MIR tasks and music representation, and then present discourse analysis methods, in detail.

2.1 Music Information Retrieval Tasks

One of the basic targets of MIR is to search for music via a large amount of music release effectively [FD02]. The main components of a MIR system usually include query formation, description extraction, matching and music document retrieval [CVG+08]. However, since the diversity of music production, it requires representations from different dimensions, such as manually-produced metadata and audio-based features. This research area can be called Representation that aims to a suitable and efficient way of the representation of music. The effectiveness of musical representation can be measured by its performances in MIR tasks. In this thesis, we focus on the representation of music and estimate its contribution to MIR tasks. Since MIR includes numerous subtasks, we implemented three different subtasks, namely, genre classification, release year estimation and popularity analysis, in which song lyrics were useful [FS14]. In this section, we introduce the related works in these MIR tasks.

2.1.1 Genre Classification of Music

Music genre classification is one of the most effective organization of music products and will continue to be used to help listeners to find favourite music. Music can be described in terms of musical genres such as Classical music and Popular music. Popular music can further comprise more specific sub-genres such as Blue, Country, Pop, Reggae, R& B, and Jazz[1]. And for each sub-genres, they can be divided into more specific genres. For instance, metal music includes heavy metal and thrash metal. Works in genre classification have considered different types of features such as audio-based features as timbral textural, rhythmic and pitch features [LOL03], lyric-based features [MNR08, FS14] and a combination of features from different sources [NR07].

2.1.2 Release Year Estimation of Music

Music from a certain period probably share common ideas or convey notable events. Some people would like to listen to music released in their school time [BMEWL11]. Music providers or music magazines also organize music products by their release years. For example, BillBoard magazine published the top 20 Billboard Hot 100 Hits of each decade. On the other hand, music researchers have explored the influences of music by year periods [Str88, Pre96]. Therefore, an appropriate retrieval scheme of music by release year can be helpful. Although release year of music is often recorded by music providers, error records probably exist. An automatic year estimation can assist the correction process. Apart from year estimation, research on which features of music change over time can benefit research on music [Her05]. Release year classification can be one of methods to find possible features changing over time. A group of features that correlates to a time range and distinguishes from its value in other time periods should be a candidate set of features changing over time. Previous works have used both lyric-based features and audio-based features for music release year classification. Fell and Sporleder used lyric-based features to perform a three-class year classification [FS14]; The presentation paper for Million Song Dataset showed the release year estimation on a large dataset with features from musical audio [BMEWL11].

[1] https://en.wikipedia.org/wiki/List_of_popular_music_genres

2.1.3 Popularity Analysis of Music

Popularity is important for music. People are more likely to listen to music with a top rank. It is useful to find out whether popular songs share similar features and what characteristics make the songs successful. Popularity analysis of music tries to answer the question. Previous work on lyric analysis showed a difference of song lyrics between the good songs and the bad songs [FS14] and a lexical distinction between the most popular songs and the other songs [EXFW15]. Although the previous works indicated the usefulness of song lyrics, the correlation between the discourse organization and the popularity has not been investigated.

2.2 General Representative Features for MIR tasks

Numerous works have explored the efficient and accurate representation methods to parameterize music so that the digital representative parameters can be used to achieve best performances in MIR tasks such as music recommendation, categorization or generation. These music representations can be extracted from different sources, such as musical content-based features from audio collections and signals, and social context-based features from music annotations such as expert labels and listener-generated materials [NRSM10, CVG+08, Dow03].

2.2.1 Content-based Features

Content-based features describe the characteristics of musical audio and can be further divided into acoustic features and symbolic features. Acoustic features, extracted from audio files such as Mp3 files, are features representing properties of audio signals, which include features such as timbral features, rhythmic features, and pitch features [TC02]. Timbral texture features, calculated from the short time Fourier transform, are standard features for music-speech discrimination. Rhythmic content features represent beats and the strength. Pitch content features describe the pitch range of music. On the other hand, symbolic features are mainly in the format of MIDI files, which include musical events and parameters instead of audio samples [CYS07]. All these features represent music in an audio aspect and have been proven to be useful in multiple MIR tasks [TC02, CYS07].

2.2.2 Context-based Features

Social context-based features such as features extracted from reviews of music, social tags, user profiles and playlists [SDP12], are not directly derived from musical audio, and are mostly manually created by musicologist or common users, involving human interactions. Comparing to content-based features from music audio, context-based features are usually on a higher level that describe general concepts of music.

2.2.3 Lyric-based Features

In addition to the previous two types of music representations, researchers have begun to exploit music lyrics in different MIR tasks and hypothesized that lyrics may complement audio-based music representations to achieve higher performances in MIR tasks [MNR08, FS14, HD10].

Song lyrics are easily queried music component of songs by common listeners, since a piece of song lyric is easier to be encoded and input than signal-based features. Song lyrics usually contain semantic information and represent structures of songs and have been exploited in multiple MIR tasks such as topic detection [KKP08, Ste14], genre classification [MNR08, FS14] and mood classification [LGH08, HDE09, HD10].

Researchers tried to find the best way to use the information from song lyrics. Previous works have introduced word-level features such as n-gram model [KKP08, Ste14, MNR08, FS14], part-of-speech tags [MNR08, FS14], and words with special semantic meanings [FS14]. N-gram models extract and rank the most representative words according to different classes such as topics and genres. Part-of-speech tags show the syntactical structure in lyrics and the style may vary in different classes. Words with special semantic meanings, such as words expressing imagery, present predominant concept in lyrics. Other features such as the word similarity between segments, chorus and title occurrence, rhyme features such as repetitions of letters or words, and repetitive structures show the structure of lyrics [FS14, KKP08].

However, comparing to audio analysis of music, lyric analysis is still at an early stage. One of the possible reasons is the limitation of rich and reliable sources. Although the widely known Million Song Dataset [BMEWL11] contains 237,662 items, the lyrics are in bag-of-words format and only reference the top 5,000 word stems, which is the number far from the words that native speakers of English have learned in their childhood (about 10 thousand) [Bie05] and could fail to capture the patterns or features underlying song lyrics, such as structure and entity relationship. Another possible reason is that compared with acoustic features, lyric features are only in music that involves singing. Song lyrics are less useful while separating from melody or background music. Sometime we can enjoy music without understanding the lyric and encoding singing as an instrument. But this is not a case for some specific MIR tasks focus on lyrics, such as query by singing [WJW10], lyric generation [PAA14], and cooperating lyric-based features and audio-based features can improve MIR task performances [NR07, HD10]. The results indicated that lyric-based features representing a song in a different aspect from traditional audio-based features, and the exploration of the representative features of song lyrics can benefit the MIR community. We predict that the discourse level of song lyrics is an important factor for some MIR tasks, since it matches cognition process of humans.

2.3 Discourse Analysis

Although human usually understand a text by multi-sentences as a whole instead of each individual sentence [LNK11], previous works on lyric-based MIR tasks seldom explored discourse-level features, which convey relations between sentences and global information within the entire text [WEK12]. Previous works usually extracted features from a text in a representation of the bag of its words, ignoring the order of words and context information. However, this mechanize does not always match human regulation. For example, we often judge text quality by its coherence or referential clarity, and high-level features beyond a bag-of-word level are also important for text simplification and text summarization [WEK12], which are useful for text classification [PLUP13]. Therefore, we hypothesize that the high-level discourse-based features can also benefit music classification.

Here, we explored discourse-level features instead of bag-of-word-level or bag-of-sentence-level features to characterize a song in a different aspect from the previous works, which may better fit the suitable song retrieval tasks. Since discourse analysis is a very broad field, in our proposed work, we implemented three types of discourse-level features: topic structure, text coherence and cohesion analysis and discourse relation.

2.3.1 Topic structure

The segmentation of a text is an essential aspect of discourse analysis, which shows the boundary between meaningful units and represents the topic structure of a text. It can incorporate with other domain-specific techniques for a lecture, a meeting or a speech segmentation [GMFLJ03] and can help in text retrieval or summarization. The text segmentation task can be further divided into two sub-tasks: linear text segmentation, which aims to find subtopic shifts within a text on a coarse-grained level, and hierarchical text segmentation, which tries to build a fine-grained subtopic structures [Pur11]. However, producing a hierarchical text segmentation requires understandings of interrelations within a text and it seems an accuracy system is hard to build as the fine-grained level segmentation is even hard for humans [Pur11]. Therefore, we only focus on linear text segmentation algorithms in this work. The linear text segmentation algorithms are based on the assumption that the text segments relate to the general topic of a text and the topic density in each segments. A typical status is that lexical choice within a text segment is much more compact than between segments. For example, one segment talks about history and the next segment introduces geography. The words such as 'year', 'age' and 'heritage' in the former segment are related to the concept of the history, which does not necessarily tight to the concept of geography from the second segment.

TextTiling is a linear topic segmentation algorithm detecting boundary of different subtopic in texts [Hea97], which is based on lexical co-occurrence. The tokenizing text is first divided into pseudo-sentences of a fixed size, and the adjacent sentences are grouped into blocks. A similarity score is calculated between adjacent blocks based on lexical co-occurrence. The largest differences between similarity scores are regarded as boundaries. Based on the similar boundary detection idea, LcSeg [GMFLJ03] considered lexical chains and the similarity score between windows were weighted by tf-idf term weights and chain length. C99 [Cho00] used cosine similarity and ranking method in the local region and applied clustering to detect text boundaries. More algorithms were developed by incorporating a topic or semantic model to better represent inner cohesion in segments, such as the TopicTiling with a Latent Dirichlet Allocation model [RB12] and the C99 with a latent semantic analysis [CWHM01]. Text segmentation algorithms were used in text genres such as meeting document [BR06], dialogue [TI14] and Twitter [ISY14], but have not been applied on song lyrics, a distinguishable text genre from non-lyric texts [WYN04].

In our preliminary work investigating the usefulness of discourse-level features in MIR tasks, we used TextTiling algorithm, which served as a baseline for many text segmentation tasks [Pur11], since the TextTiling algorithm was simple and efficient. The TextTiling algorithm made few assumptions about the domain of the text [BR06] and should be suitable for the proposed segmentation task as different topics and structures existed in song lyrics. On the other hand, algorithms incorporating topic models require a suitable training dataset to best estimate the distribution of topics and make the segmentation. For instance, the TopicTiling algorithm assigns each word with a most frequent topic id based on a training set and discards the other lexical meanings. The lost meanings of words may influence the performance of text segmentation. Since our work was a seminal work for topic segmentation in song lyrics, we left the development of segmentation algorithm in the future works and focused on the investigation of usefulness of this set of representative features in MIR tasks.

2.3.2 Text Coherence and Cohesion

Given a text, we sometimes care about its compactness in overall textual signals to perform further text analysis. Text coherence and cohesion analysis opens up a pathway between linguistics and cognitive science, trying to seek the rules of relations within texts. The coherence and cohesion of a text has been proven to be important for human understanding [GMLC04] and writing quality analysis [CM10]. Although the concepts of text coherence and text cohesion are relevant to each other, there exist differences between the two concepts. While the text coherence is a subjective property of text based on human understanding, the text cohesion is an objective property of explicit text element interpretation patterns [GMLC04].

Various studies focused on sub-tasks of this specific text analysis task. In our proposed work, we only focused on entity-related measures as entity density for comprehension [FEH09], entity grid [BL08] and coreference resolution systems [LPC+11], since entities usually conveyed conceptual ideas and a study by Feng et al. [FH14] showed the appearance pattern of entities might vary according to different writing styles. Therefore, we hypothesize that the cohesion patterns in song lyrics may vary according to different categories, such as popularity and genre.

2.3.3 Discourse Relation

Discourse relation describes how two text elements logically relate to each other. Previous works have introduced various discourse relation corpora and frameworks. Rhetorical Structure Theory (RST) had a nucleus-satellite view that the satellite text span was subordinate to the nucleus, and built a tree structure presenting conceptual relations between abstract objects [MT88]. The discourse tree was constructed based on adjacent elementary discourse units, which were minimal units in the RST framework [Mar97]. Models based on a probabilistic method [SM03], a support vector machine [HPDI10] and a deep learning approach [LLH14] were developed to build the RST-styled discourse tree automatically. The evaluation of automatic RST-styled discourse parsing can be evaluated on a RST discourse treebank [COM02]. Graphbank constructed a less constrained chain graph structure allowing crossed dependencies [WG05] and Wellner et al. [WPH+09] developed an automatic classifier to identify discourse coherence relations in discourse Graphbank.

Penn Discourse Treebank (PDTB) corpus annotated discourse connectives and their arguments and defined a hierarchical scheme with three layers for the level of discourse structure [PDL+08, PMD+07]. It followed lexically grounds and predicate-argument structures described in Lexicalized Tree Adjoining Grammar for Discourse (D-LTAG) framework [Web04]. The

PDTB-styled discourse relations focused on discourse connectives as explicit discourse connectives, such as ‘since’ and ‘while’ from pre-defined syntactic classes, and implicit discourse connectives that relation situation pre-defined in the framework in which the meaning can be understood, but not explicitly stated, and used discourse connectives as predicates to detect two text elements within a sentence or between sentences as the corresponding arguments.

In our preliminary work, we used the PDTB-styled discourse relation since it was a flexible discourse relation framework allowing multi-relations in sentences and not restricted to adjacent text spans as the RST framework [NKL+13], and it was less computationally intensive and thus better for real systems.

Chapter 3

The Approach for Discourse Analysis on Lyrics

Since a text is usually understood by multiple linking sentences instead of isolate sentences independently, it raises interests on explaining such relations between sentences and internal discourse structure in texts. The applications of discourse analysis include text summarization, student essay scoring, information extraction, sentiment analysis and machine translation [WEK12]. Research on discourse analysis on multiple text genres such as essay and scientific article have been studied extensively, however, the differences between song lyrics and other text genres make the discourse analysis of lyrics becoming an explorational task.

Considering the unique characteristics of song lyrics, we adjusted the methods to fit to this specific text genre, including the encoding of PDTB-styled discourse relations and the parameter settings of the TextTiling segmentation algorithm. Three major discourse structures, namely discourse relation, topic structure and text cohesion, were analyzed. Discourse relations indicate how two text elements logically relate to each other. The pattern of discourse relations, which reflect opinion and concept relations [SWR08, NKL+13], could vary according to different music categories. For example, Rap are likely to contain more discourse relations in song lyrics than other music genres. On the other hand, topic structure can be different according to writing styles or music categories. As for text cohesion and coherence analysis, some songs may focus on a specific entity and be compact in texts, while other songs does not. The compactness of songs may influence the preference of users.

3.1 Discourse Relation

Discourse relations explore the internal logical relations between text elements. We used PDTB-styled discourse relations [PDL+08] on song lyrics for discourse analysis. Comparing to other discourse relation frameworks, such as RST-styled discourse relations [Mar97] and Graphbank-styled discourse relations [WG05], PDTB-styled discourse relations can construct a flexible discourse relation framework in an efficient way and have been used in multiple NLP tasks [NKL+13].

We used a PDTB-styled parser[1] [LNK14] 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 that links arguments. The connective classifier and explicit discourse connective classifier are trained on PDTB corpus using part-of-speech and relative-word features. If a connective is identified as a discourse connective, the parser then identifies the explicit relation the connective conveys.

The system considers four general relations that include 'Temporal', 'Comparison', 'Contingency' and 'Expansion' and sixteen specific relations which are subcategories of the 4 general relations. For example, 'Synchrony' and 'Asynchronous' are specific 'Temporal' relations. The first level and second level discourse relations in PDTB is presented in Table 3.1. For a detailed explanation, see PDTB annotation manual [PMD+07].

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

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

First Level	**Second Level**
Temporal	Synchrony
	Asynchronous
Contingency	Cause
	Pragmatic Cause
	Condition
	Pragmatic Condition
Comparison	Contrast
	Pragmatic Contrast
	Concession
	Pragmatic Concession
Expansion	Conjunction
	Instantiation
	Restatement
	Alternative
	Exception
	List

Table 3.1: First level and second level of the discourse relations in the PDTB.

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; 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 (i.e., pairs of specific relations which never occur in the corpus), 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. These features can represent general and specific discourse relation distributions in song lyrics and are probably useful for MIR tasks.

3.2 Topic Structure

Given a long text including different subjects, we can divide it into a group of shorter and more meaningful topical-coherent segments for better interpretation. Topic segmentation algorithms were proposed to solve such task. The segmentation is useful especially for information retrieval, as it can be an indicator for higher level tasks such as the text summarization

[ADBM02]. Although a song contains more than its lyric, the lyric is the most important source for semantic interpretation. For example, listeners can search a set of topics [KKP08].

Topic structure can be on a coarse-grained level, which aims to find subtopic shifts and is usually done by a linear text segmentation algorithm, or on a fine-grained level, which requires understandings of hierarchical interrelations in a text and pre-knowledge of how a category of texts should be organized. Although the latter hierarchical topic structure seems to give a clear view of how topics are placed in a song lyric, it is hard to obtain a common pattern of song lyrics. In our proposed work, we focus on the coarse-grained level of song lyrics, trying to detect how topics shift in a piece of song lyric.

Considering a presentation of a particular concept, we usually use relevant words including names, locations or referring expression. Hence, a new topic is usually showed with a group of different vocabulary or expressions from the previous one. Seeing this fact, a change in topics will be correlated to the introduction of new words. We used the TextTiling algorithm [Hea97], which was one of the most useful linear text segmentation algorithms based on the lexical change to estimate topic shifts within a piece of lyric.

Figure 3.1 shows the process of the TextTiling topic segmentation algorithm.

A song lyric is firstly tokenized, and then is divided into pseudo-sentences with w tokens. The adjacent k pseudo-sentences are finally grouped into chunks. Stop words are removed before the processing. After the division process, a similarity score is then calculated between two adjacent blocks. The function of the similarity score is a formalized dot product of vectors of tokens in the right block and left block as presented in Equation 3.1 [Hea97].

$$s_i = \frac{\sum w_{block1} w_{block2}}{\sqrt{\sum w_{block1}^2 \sum w_{block2}^2}} \tag{3.1}$$

The largest differences between similarity scores are regarded as boundaries. The difference is measured by the depth of each gap, which is computed as Equation 3.2, where d is the depth score and s is the similarity score. The larger the depth score, the more likely the boundary occurs at that

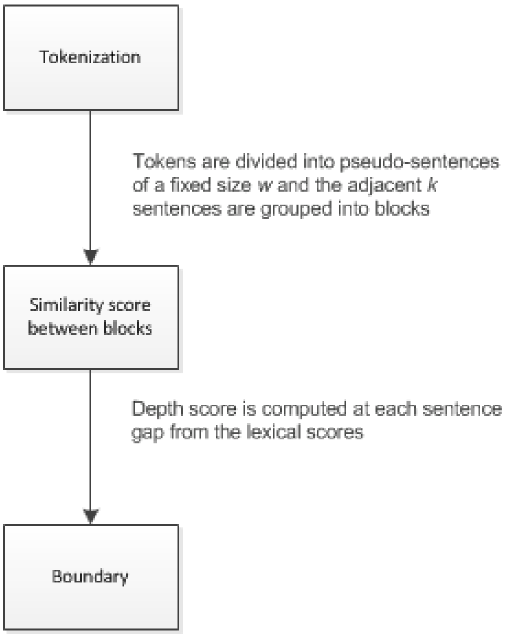
Tokenization
Tokens are divided into pseudo-sentences of a fixed size w and the adjacent k sentences are grouped into blocks
Similarity score between blocks
Depth score is computed at each sentence gap from the lexical scores
Boundary

Figure 3.1: The TextTiling algorithm process.

location. However, a pre-defined number of boundaries is arbitrary and a threshold based on the distribution of the depth scores is used to decide the boundaries. The equation of the threshold is presented in Equation 3.3, where $\overline{d}$ is the average of all depth score and σ is the standard deviation.

$$d_i = (s_{i-1} - s_i) + (s_{i+1} - s_i) = s_{i-1} + s_{i+1} - 2s_i \tag{3.2}$$

$$threshold = \overline{d} - \frac{\sigma}{2} \tag{3.3}$$

We ran the TextTiling algorithm 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, since the average line length in our lyric dataset was around 7. Features were then calculated by computing the mean and standard deviation in the number of words in a lyric's segments

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

and the number of segments. These features represented the characteristics of text segments and topic structure in song lyrics and might help in music classification tasks.

An example of the segmentation result of a song lyric is from “California Girls” by The Beach Boys [BLS13], a song with an A-A-B-A-A-B-B structure that ‘A’ represents a variable part and ‘B’ represents a repetitive part. The TextTiling segmentation result is on the topic shift basis. Although variable expressions exist, it merge the first two segments since they convey the same topic. The last two segments, which are exactly the same except one word difference, are merged as well.

3.3 Text Cohesion

3.3.1 Entity Density

General nouns and named entities including locations, organization and names usually indicate conceptual information [FJHE10]. Previous research has shown that named entities are useful to convey summarized ideas [GKMC99] and we hypothesize that entity distribution could vary according to song styles. We implemented five features including the ratios of the number of entities and the average numbers of entities per sentence, which are listed in Table 3.2. We used OpenNLP[3] to find named entities and Stanford Part-Of-Speech Tagger[4] to extract general nouns. These entity distributions can help in the summarization of concepts, and thus could be used for style classification.

3.3.2 Coreference Inference

Entities and their pronominal references in a text which represent a same object build a coreference chain [LPC+11]. The pattern of how an entity represented by different text elements with the same semantic through text may vary in different song styles. We used Stanford Coreference Resolution

[3] https://opennlp.apache.org

[4] http://nlp.stanford.edu/software/tagger.shtml

Feature Set
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 entities
average number of named entities per sentence
average number of all entities per sentence

Table 3.2: Entity density features.

System[5] to generate coreference chains.An example is from a clip of "Another One Bites the Dust" by Queen. The words as 'I', 'me' and 'my' refers the same entity and construct a coreference chain of the song.

> How do you think **I**'m going to get along without you
> when you're gone
> You took **me** for everything that **I** had
> And kicked **me** out on **my** own

The five features were extracted as listed in Table 3.3. The inference distance was the minimum line distance between the referent and its pronominal reference. The chain was active on a word if the chain passes its location. These coreference-chain-based features represented referential relations in song lyrics.

3.3.3 Entity Grid

Barzilay and Lapata's [BL08] entity grid model was created to measure discourse coherence and can be used for authorship attribution task [FH14]. We thus hypothesized that subjects and objects may also be related differently in different genres, just as they may be related differently for artists.

The entity grid method extracted local coherence of a text at the level of sentence-to-sentence transitions. A two-dimensional array was used to

[5] `http://nlp.stanford.edu/projects/coref.shtml`

Feature Set
total number of coreference chains
number of coreference chains which span more than half of lyric length
average number of coreferences per chain
average inference distance per chain
number of active coreference chains per word

Table 3.3: Coreference inference features.

Sentence	you	I	Everything
1	S	S	X
2	-	O	-

Table 3.4: Entity grid example.

represent the distribution of entities in a text. Each cell in a grid represents one of the grammatical roles of subject (S), object (O), neither of the two (X) and absent in the sentence (-) of a discourse entity in a sentence. An example is from two lines of "Another One Bites the Dust" by Queen and its entity grid is showed in Table 3.4.

> You (S) took me (O) for everything (X) that I (S) had
> And kicked me (O) out on my own (X)

Brown Coherence Toolkit [EAC07] was used to generate an entity grid for each lyric. 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 the percentage of each pattern. The number of entity-grid features are listed in Table 3.5. These entity-transition features indicated the entity distributions in sentence-to-sentence transitions.

Feature Set
total number of adjacent entity transitions
number of adjacent entity transitions
ratio of the number of adjacent entity transitions to total number of entity transitions

Table 3.5: Entity grid features.

Chapter 4

The Discourse-feature-based Approach for Three MIR Subtasks

To further validate that features on discourse level are useful in MIR tasks and even superior than features extracted from separate sentences, we then performed three MIR tasks, namely music genre classification, release year estimation and popular song analysis, with discourse-level features and a comparable lyric-based baseline feature set.

4.1 Experiment Tasks

We considered genre classification, release year estimation and popular song analysis to test the validation of discourse-level features. The annotation tags have been linked to each song in the experiment dataset.

4.1.1 Music Genre Classification

Music genre is a crucial metadata for music description and it can be useful for music retrieval and music indexing. The increasing number of music products on the Internet requires an automatic recognition for the music genre, since an accurate manual annotation is time-consuming and expertise-requiring. Music genre classification can build a model to predict the genre of a piece of music. Although automatic music genre recognition is a challenging task, numerous works have been proposed to solve the problem based on content-based musical features such as tempo, rhythmic structure and instrumentation [SZM06]. However, music genre is a combination of culture and aesthetics. A musical genre sometimes stands for a culture class [Fab81]. Here, we consider song lyrics to perform music genre classification, since song lyrics represent a unique aspect of culture including language, theme and expression context. In the experiment, we considered the sub-genres of popular songs and analyzed song lyrics on a discourse level, considering the links between text segments, which can vary according to music genres.

4.1.2 Music Release Year Estimation

The release year metadata can be used for music retrieval and recommendation system. It is suggested that listeners usually preferred music that were released in particular year periods, such as the period of their school time [BMEWL11]. Although the metadata of release year is available online or from music companies, errors and missing data still exist. An automatic music release year predictor can help in fixing such problem. On the other hand, a model of the variation in discourse-level of characteristics in song lyrics can contribute to the long-term evolution of music. Previous work has shown the style of song lyrics can change over time for Rap songs [HB10] and it is also expected to see if lyrics of other genres such as Pop/Rock can show a change. Therefore, we performed music release year classification by using discourse-level lyric-based features on Pop/Rock genre considering songs from different decades.

4.1.3 Popular Song Analysis

Song lyrics contribute to whether a song is rated as the most popular one or not. Salley presented how the interaction of alliteration with audio characteristics could make a popular song more successful [Sal11]. Another work by Cunningham et al. [CDB05] also showed a survey results from listeners that lyric quality should be one of the most important factors for disliking a song.

In the experiment, we used Billboard Magazine's lists of the "All-Time Top 100 Songs" [1] as the most popular songs. Songs were ranked with respect to

[1] www.billboard.com/articles/list/2155531/the-hot-100-all-time-top-songs

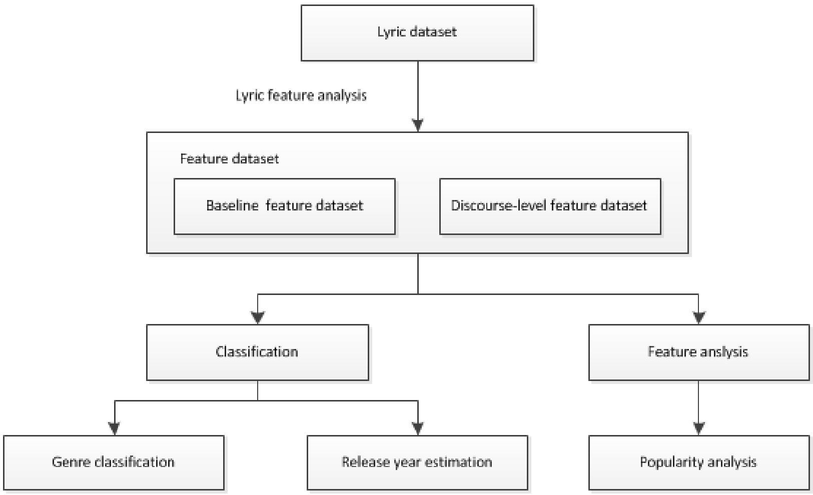

Figure 4.1: General experiment framework.

overall popularity on the magazine's "Hot 100" chart, an industry-standard ranking of top 100 popular songs in the United States, published weekly since 1958 [EXFW15].

4.2 General Experiment Framework

We treated music genre classification and release year estimation as classification tasks, and performed the empirical cumulative distribution function on popularity analysis to find out the difference patterns between the most popular songs and the other songs in our dataset. We did not perform the classification on the popularity analysis task since the classification result from extreme imbalance dataset of the most popular songs (100 lyrics) and the rest of the songs cannot draw a convincing conclusion. The entire framework is presented in Figure 4.1. The experiment features and experiment dataset are introduced in the Section 4.5 and Section 4.6 respectively.

4.3 Music Classification Framework

4.3.1 The Supervised-learning Classification of Music

Data classification is an important method for data analysis, which extracts models for important data classes. Data classification usually involves two parts: learning and classification. The learning process builds a model based on a given training dataset, while the classification process predicts data class for a test dataset or a new dataset. Considering whether a class label attribute is given for each data item tuple, the classification task can be further divided into two classes: supervised learning that the label attribute is pre-defined and unsupervised learning or clustering that the label attribute is unknown.

Given a data instance $X = (x_1, x_2, ..., x_n)$, and $x_1, x_2, ..., x_n$ represent the value of data attribute $A_1, A_2, ..., A_n$, the learning process can be seen as learning a mapping function:

$$y = f(X) \tag{4.1}$$

where y is the class label attribute for the given tuple X.

We determined our tasks as supervised-learning classification tasks and the class label attributes (e.g., music genres) were collected before learning process.

For model accuracy estimation, we decided to use k-fold cross-validation ($k = 10$) as it was suggested having a relatively low bias and variance [HK06]. During the 10-fold cross-validation process, the dataset is divided into 10 mutually disjoint subsets $D_1, D_2, ..., D_{10}$ with approximately equal sizes. The learning and testing classification process repeats 10 times. For each iteration i, D_i is selected as the test dataset, while the other subsets together are used as training dataset. The accuracy for 10-fold cross-validation is the average of accuracies for all folds. Every data tuple X is used 9 times for training and used once for testing so that data is effectively exploited in the entire process, which is possibly avoid overfitting and underfitting.

4.3.2 Representative Tuple

A data instance is represented by a set of values or a group of features showing its status. In our proposed work, each instance should be a tuple of numerical or categorical values derived from song lyrics. The discourse-based features used in the experiment are described in Chapter 3 from the aspects of discourse relations, topic structures and text cohesion and coherence. We also implemented baseline lyric-based features, which is introduced in the Experiment Features Section. To further validate the contribution of each feature group to the classification, we split each full instance to sub-instances with sub-group of features. For example, the discourse-level features are divided into three groups according to their representative aspects of lyric characteristics, namely PDTB-styled discourse relation, TextTiling topic segmentation and text cohesion. The performance of the classification task by using each sub-group of features indicates the potential of such features in MIR tasks.

4.3.3 Learning Method: Support Vector Machine

Classification can be readable by methods such as decision tree, association rule-based classifiers or mathematical formulas [HK06]. Although these classifiers can be interpreted by humans, the performance of these classifiers can be relatively low in MIR tasks [SKK08]. Therefore, we decided to use Support Vector Machines (SVMs) to solve the proposed music classification tasks, which have been widely used in multiple MIR tasks [FS14, MNR08]. Since our main target is to investigate how discourse-level features from song lyrics can benefit MIR research, we did not explore which classifier can achieve the highest performance in the specific music classification tasks in this thesis and leave it in future works.

An SVM model uses a non-linear mapping function to map the training data to a high-dimensional space, and then it constructs a Maximum Marginal Hyperplane that can best separate data points from different classes. Although the SVM model requires relatively long training time, it performs well in complex non-linear modeling. It is less likely to achieve overfitting comparing to other models.

An SVM can deal with a linear classification as well as a non-linear classification using a kernel trick. Here, we used a radial basis function (RBF) as a kernel function (Equation 4.2), where $\gamma > 0$. The RBF maps data to a high-dimensional space non-linearly so that it can be used in non-linear data classification. Linear kernel is a special case of RBF [KL03].

$$k(X_i, X_j) = \exp^{-\gamma||X_i - X_j||^2} \tag{4.2}$$

Although RBF is a reasonable choice of a kernel function, it requires hyperparameter decisions which are usually determined by experiments. In the proposed experiments, we used the default parameters in Weka and leave the parameter tuning in the future work.

4.3.4 Model Accuracy Assessment

Model accuracy assessment is an important part of the best model decision and accuracy measures usually include accuracy, error rate, recall, precision, F-score and so on. Among all these measures, we used F-score on the model assessment. F-score is a measure of model's accuracy and it reaches its best value at 1 and worst value at 0. F-score is a weighted average of precision and recall. The two measures are based on an understanding of relevance. Precision represents the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved [HK06]. The formulas for F-score is shown in Equation 4.5 [HK06].

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.5}$$

where true positive (TP) is the proportion of positives that are correctly identified and false positive (FP) is the proportion of negatives that are incorrectly identified as positives. And false negative (FN) is the proportion of negatives that are correctly identified.

4.3.5 Sampling from Imbalanced Data

Imbalanced datasets widely exist in "real-world" problem. For example, there are more pop/rock songs than blue songs in our dataset. Given an imbalanced dataset with two or more classes, quantity of data can be unequal according to their representative classes. In addition, with an imbalanced dataset, the distribution of the test dataset may be different from that of the training dataset.

Sampling techniques have been recommended for balancing the datasets to improve the performance of classification. Over-sampling and under-sampling are often used to balance the distribution of data [GC11]. Over-sampling repeatedly draw samples from a minority class to increase its samples, while under-sampling randomly delete samples from a majority class until the amount of data tuples from each class is equal.

We determined to use under-sampling method for our proposed classification cases to avoid overfitting by over-sampling, although it may ignore some information from the majority class. However, we are more likely to pay attention to a positive class, which is mostly a minority class in most of our proposed cases. For example, we concern whether a song is a rap song instead of whether a song is not a rap song.

Therefore, we performed random under-sampling to balance dataset before each 10-cross validation using a SVM. To further decrease the bias caused by random sampling, we repeated the entire sampling-classification process and took the average of values of F-measure from all classification results.

4.4 Empirical Cumulative Distribution Function

Since the extreme difference of the number of the most popular songs (100) and the number of the rest of songs in our dataset, we did not perform music classification task on popularity analysis. However, it can be useful to show different patterns on a discourse level between the two categories. To prove our hypothesis, we performed the empirical cumulative distribution function (ECDF) on discourse-level features.

ECDF is one of the simplest non-parametric estimators. Suppose $x_1, x_2, ..., x_n$ is an independent and identically distribution sample from an unknown distribution function. The ECDF is presented as Equation 4.6 and 4.7 that each data x_i puts mass $\frac{1}{n}$ [Law11]. We compared the differences of ECDFs of the discourse-level features from different music categories.

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{4.6}$$

$$I(x_i \leq x) = \begin{cases} 1 & x_i \leq x \\ 0 & otherwise \end{cases} \tag{4.7}$$

4.5 Experiment Features

4.5.1 Discourse-based Features

We used discourse features described in Chapter 3. The number of discourse features for each category is shown in Table 4.1.

Dimension	**Abbreviation**	**Number of features**
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

Table 4.1: Discourse-based features used in classification tasks.

4.5.2 Baseline Lyric-based 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 4.2) [FS14]:

Vocabulary: We used the Scikit-learn library[2] 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 frequency of each n-gram in each lyric. We also computed the type/token ratio to represent the 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, a dictionary of slang, but not Wikitionary.

Part-of-Speech features: We used Part-of-Speech tags (POS tags) obtained from the Stanford Part-Of-Speech Tagger[3] to determine the frequencies of each part of speech in lyrics. Super-tags such as Adjective, Adverb, Verb and Noun were used.

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 4.2 shows all baseline features used in the classification tasks and the number of elements in each feature set.

[2] http://scikit-learn.org/stable/modules/feature_extraction.html

[3] http://nlp.stanford.edu/software/tagger.shtml

Dimension	Abbreviation	Number of features
vocabulary	VOCAB	303
POS tags	POS	4
length	LEN	3
orientation	OR	6
structure	STRUC	2

Table 4.2: Baseline features used in classification tasks.

4.6 Experiment Dataset

Subsets of our song lyric dataset [EXFW15] were used for our experiments. Unlike other corpora, such as musiXmatch lyrics dataset for the Million Song Dataset [BMEWL11], lyrics from the selected corpus are not bags-of-words, but are stored in full sentences, allowing for the retention of discourse relations.

We also downloaded corresponding genre tags and album release years for the songs presented in this dataset from Rovi[4].

Genre classification: We only kept 70,225 songs with a unique genre tag 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 analyzed the change of song lyrics by decades. We then extracted all the songs labeled

[4] http://developer.rovicorp.com

Music Genre	**Number of lyrics**
Pop/Rock	45,020
Rap	16,548
Country	12,050
R& B	8,190
Religious	2,934
Electronic	2,404
Jazz	1,302
Folk	700
Reggae	892

Table 4.3: Data sizes for music genre classification after under-sampling.

Release Year Class	**Number of lyrics**
1969-1971	536
1989-1991	536
2009-2011	536

Table 4.4: Data sizes for release year classification after under-sampling.

in one of three time ranges: 1969-1971 (536 songs), 1989-1991 (3,027), and 2009-2011 (4,382).

Top song analysis: We labeled songs as 'popular' or not based on whether they were present on Billboard Magazine's list of "All-Time Top 100" songs and propose a comparison between 'popular' songs and the other songs on discourse-based features. 87,278 total songs were used in this experiment.

The specific number of lyrics for each experiment is shown in Table 4.3, Table 4.4 and Table 4.5.

Release Year Class	**Number of lyrics**
The most popular songs	100
Other songs	87,178

Table 4.5: Data sizes for popular song analysis.

Chapter 5

Experiment and Result

5.1 Genre Classification

We ran SVM classifiers using 10-fold cross-validation. These classifiers were implemented with Weka[1] using the default settings. We chose SVM classifiers because they have been proven to be of use in multiple MIR tasks [FS14, HD10, XMS05]. Because each genre had a different number of samples, under-sampling [GC11] was performed 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. A separate classifier was used for each genre. The under-sampling and classification process was repeated 10 times and we present the average F-score for each classification task. The value of F-score by random should be 50%.

5.1.1 Classification with Non-normalized Features

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 (%) from separate classifiers for each genre are depicted in Table 5.1. Since vocabulary features are heavily dependent on which corpus the language model trains on to generate the n-gram vector, we note that we used all lyrics from each genre to obtain top n-grams.

[1] http://cs.waikato.ac.nz/ml/weka

Feature	R& B	Folk	Country	Rap	Electronic	Religious	Jazz	Reggae	Pop	Average
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 5.1: Comparison of different feature sets for genre classification by F-score (%).

We then evaluated the utility of discourse-based features for this specific task. Table 5.1 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, and then all discourse and all baseline features together. Statistical significance was 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 was 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 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. Apart from the importance of choosing only the best features for this classification task to avoid lowering classification accuracy, we then performed normalization on the feature sets. The normalization step was expected to improve the results of combination of different feature sets.

One final 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.

5.1.2 Classification with Normalized Features

We then normalized all numeric values in the feature dataset which included all songs with genre tags. The value of each feature ranged from 0 to 1. The same group of features except the combination of discourse relation features and TextTiling topic segmentation features were used for the same genre classification tasks. The average F-scores (%) from separate classifiers for each genres were showed in Table 5.2. A standard two-class t-test between the highest F-score and each result from other feature set for each genre was computed, and each column's best result was found to be significant with $p < 0.01$.

As can be seen from Table 5.2, the feature set of the combination of all features including both baseline features and discourse-based features (All) outperforms all baseline features for each genre. In addition, the feature set of the combination of all discourse-based features (DF) alone is better than all baseline feature sets in 'Folk', 'Rap' and 'Electronic' genre classification. The presented results show that the proposed discourse features assist the baseline features to achieve higher accuracy in every genre, and discourse features can be better used for some specific genre classification tasks. We note that each feature set alone did not enable the classifier to achieve maximal results in this task, indicating the importance of using features from multiple aspects from lyrics to achieve a higher performance.

Both normalized and non-normalized genre classification show the usefulness of the proposed discourse-based features in this particular MIR task. We then investigated stability of the proposed model when input changed.

Feature	R& B	Folk	Country	Rap	Electronic	Religious	Jazz	Reggae	Pop	Average
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 5.2: Comparison of different normalized feature sets for genre classification by F-score (%).

Music Genre	**Number of lyrics**
Pop/Rock	58,422
Rap	23,358
Country	18,682
R& B	17,334
Religious	4,840
Electronic	15,394
Jazz	2,478
Folk	4,344
Reggae	1,698

Table 5.3: Data sizes for music genre classification after under-sampling.

5.1.3 Classification with songs belong to multiple genres

We then performed genre classification on a dataset with songs that belong to multiple genres. For example, a song from the dataset can from both 'Pop/Rock' and 'Rap'. The number of lyrics after under-sampling for each genre classification is presented in Table 5.3. We then performed genre classification using feature sets of vocabulary (VOCAB), all baseline features (TF), discourse relation and topic structure (DR+TT) and all features (All) as they are representative feature sets according to the previous sections.

Table 5.4 shows the binary classification results and the best performances are bolded. As can be seen from Table 5.4, the discourse-based features alone can outperform baselines in 6 genres: Pop, Folk, Rap, Religious, Reggae and Jazz. It is statistical significant ($p < 0.01$) compared to the other feature models. Discourse-based features improved the classification results in the other 3 genres: R& B, Country and Electronic and the difference existed in the best results and the others ($p < 0.05$).

Feature	R& B	Folk	Country	Rap	Electronic	Religious	Jazz	Reggae	Pop	Average
VOCAB	51.4	50.3	55.5	74.5	50.1	51.4	51.6	50.4	53.8	54.3
TF(All)	65.9	53.1	64.0	75.9	50.4	52.4	54.6	52.0	59.8	58.7
DR + TT	63.3	**60.1**	63.0	**86.7**	55.8	**55.8**	**59.8**	**60.2**	**63.0**	**63.3**
All	**69.4**	55.1	**66.0**	76.6	**57.2**	52.9	54.8	55.5	62.8	61.0

Table 5.4: Comparison of different feature sets for genre classification by F-score (%).

5.2 Release Year Estimation

We investigated whether discourse-based features can help to estimate the release date of a song, on the basis that the lyrical structure of song texts is likely to change over time [FS14, HB10]. 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 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 [FS14]. Under-sampling was again used to balance the dataset building a sub-dataset with 1603 lyrics before each classification with an SVM with 10-fold cross validation for three-class classification. The process was repeated 10 times. The F-score by random should be 33%.

5.2.1 Classification with non-normalized Features

Table 5.5 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. 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 other discourse features were again shown to be less useful than the DR and TT ones. In addition, 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. This result is intuitive; music in the earliest and latest ranges can be as much as 42 years removed from other songs in the dataset, but music in the middle range can be no more than 22 years removed from the rest of the songs, and so will likely be more similar to them since they were produced closer together. Finally, we observed a remarkable classification result 0.00 for 1969-1971 class with structural features, which include song title and chorus detection. No instance was classified to the earliest year period with the SVMs. One of the main reasons is that only two nominal features with values of 0 or 1 representing the occurrence of a song title or a chorus constructed the structural feature tuple and an occurrence of a song title or a chorus may not change over years in our dataset for release year classification.

5.2.2 Classification with normalized Features

We then run classifications with normalized features using the same settings as in the previous sub-section. Table 5.6 shows all average F-scores for each feature set. The combination of all features outperformed all baseline features in average F-score ($p < 0.001$). We note that structure features achieve high accuracy in the class of latest ages, but none of songs was classified as a song from earliest ages. This pattern shows by only using structure features the classification performance can not achieve a maximum accuracy. On the other hand, the classification accuracy for the class of earliest ages using the combination of all baseline features (TF) is better

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 5.5: Comparison of different feature sets for release year estimation by F-score (%).

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 5.6: Comparison of different normalized feature sets for release year estimation by F-score (%).

than others, while the average F-scores indicate this group of features cannot outperform the combination of all features in this multi-class classification task. By combining the proposed discourse features, the year classification performance can be improved.

5.3 Popular Song Analysis

Since lyrics can influence a song's popularity, we analyzed the connectives and discourse connectives of 'most popular' or 'top' songs (defined as being songs on Billboard Magazine's "All Time Top 100 Songs" list) compared to songs which were not on that list. In order to avoid the problem of very long lyrics biasing the results, the number of connectives was normalized by the number of words in each song. We also analyzed text cohesion and coherence by analyzing whether the coreference inference pattern presents differently in 'top' songs and the rest of songs by using the average number of coreferences per chain. 87,278 total songs remained to be used in this experiment, of which 100 were labelled as 'top' songs and the remainder were non-top songs.

In order to determine the potential of discourse analysis for indicating whether a given song was likely to be popular or not, we calculated the empirical cumulative distribution function (ECDF) of the normalized number of discourse connectives and the average number of coreferences per chain for the set of 100 Top Songs as well as the remaining songs (Figure 5.1 and Figure 5.2).

The curve corresponding to top songs is higher than the curve corresponding to other songs as shown in Figure 5.1, indicating that top songs tend to have fewer discourse relations. The opposite pattern shows in the number of coreferences per chain for the two sets as shown in Figure 5.2, suggesting top songs contain more coreferences per chain. A non-parametric Wilcoxon rank sum test showed that these results had very high statistical significance ($p < 0.005$), further validating these results. The same comparison was run on the normalized number of all connectives as well, but the result showed there was no significant difference between top songs and other songs.

Our preliminary discourse analysis on the most popular songs and the rest of songs shows the different patterns in the most popular songs and the rest of songs, revealing the trend of popular songs such as less complex sentence relations and relative concentric entities, which can be further used as a predictor for popular song estimation.

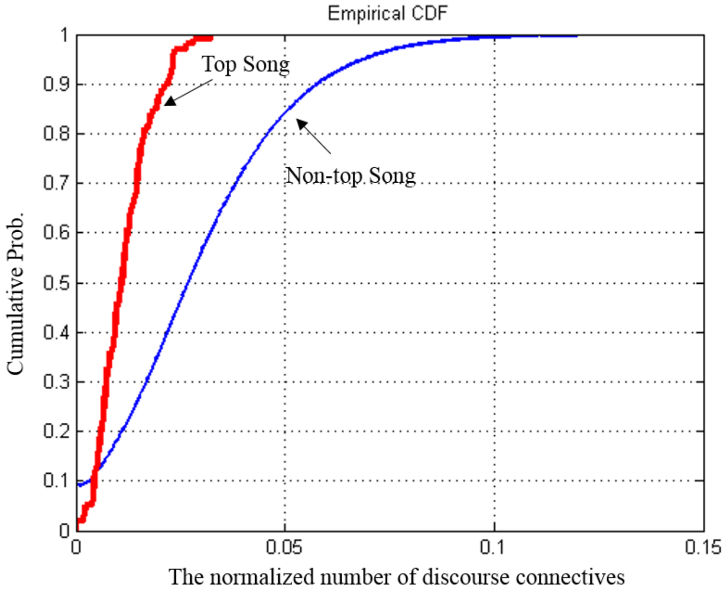

Figure 5.1: ECDF of the normalized number of discourse connectives for the 100 Top Songs and non-top songs in the corpus.

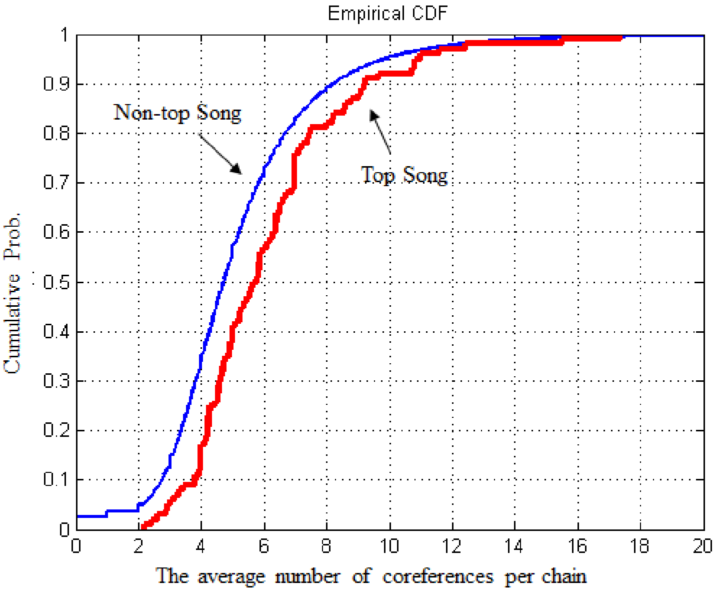

Figure 5.2: ECDF of the average number of coreferences per chain for the 100 Top Songs and non-top songs in the corpus.

Chapter 6

Conclusion

In this thesis, we investigated the usefulness of discourse-based features from song lyrics and demonstrated that such features can provide useful information for three MIR classification tasks. The word "music" is specifically referred to "song with lyric" in this thesis, but we follow the convention from MIR community and use "music" in the content.

We analyzed song lyrics on a higher level than a bag-of-words/bag-of-sentences level in three different discourse aspects: discourse relation, topic structure and text cohesion. The PDTB-styled discourse relations were extracted from song lyrics and the TextTiling algorithm was used for linear topic segmentation. Text cohesion was analyzed by using entity-density features, coreference-inference features and entity-grid features. To further analyze how discourse-level features can benefit MIR tasks, we then performed three MIR tasks with all these features. Genre classification and release year estimation showed that by incorporating discourse-level features into the classifiers the performances were higher than the performances only using previously used word-level features in both non-normalized and normalized classifications. When the input data for gerne classification changed to songs that belong to multiple genres, the classification results showed the stability of the model that discourse-level features can outperfered the baseline features. The popularity analysis of song lyrics indicated the discourse-level differences between the most popular songs and the rest of songs in our dataset. Our discourse analysis of song lyrics shows the potential of this group of features in MIR tasks. Since our main target is to investigate suitable discourse-level lyric-based features for MIR tasks instead of finding the most suitable classifiers in this thesis, we used SVM classifiers for all these classification tasks as the SVM model can provide relatively high performances in MIR tasks [SKK08].

However, this work is an exploration work and further sophisticate 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, such as PDTB-styled parser, coreference resolution system and entity grid. The representative vector will probably vary if another split scheme is implemented or can be less useful when it is applied on another dataset.

Our target is to investigate the suitable discourse-level lyric-based features for MIR tasks and therefore used SVM classifers for all classification tasks.

As for the performances of classification tasks, the F-scores of genre classification were not as high as described in some previous works might due to the difference between lyric datasets [FS14]. Our dataset contains a variety of songs and it is difficult to find a common pattern from such a big dataset. Another reason is that the probability of a song belongs to a musical genre was simplified from an original ten-scale value to a binary value to reduce the sparsity in the experiment.

To sum up, in this thesis, we proposed a group of discourse-level features on song lyrics and then experimented on three MIR tasks. Although more investigations can be done in this research direction, the presented results have indicated the potential of these features on MIR tasks.

Chapter 7

Future Work

This thesis presents a pilot study on discourse analysis of song lyrics and its application in MIR tasks. In the future, we will extend our work in the following ways:

Further Feature Analysis: we will make a deeper investigation on the influence of discourse-level features obtained from different text pre-processing schemes, such as sentence split schemes. A good scheme can better represent the characteristics of song lyrics and may achieve higher performances in the MIR tasks.

Discourse Feature Extension: we will further explore other dimensions of discourse-level features on song lyrics, such as function structure and eventuality structure [WEK12]. The pattern of these features on a large lyric dataset with songs from different categories can benefit MIR society by showing how song lyrics change over classes.

MIR Task Extension: our experiments on genre classification, release year estimation and popularity analysis showed the usefulness of our proposed discourse-level lyric-based features. The proposed features can be applied in other MIR tasks such as mood classification and keyword extraction.

References

[ADBM02] R. Angheluta, R. De Busser, and M. Moens. The use of topic segmentation for automatic summarization. In *Proc. of the Association for Computational Linguistics Workshop on Automatic Summarization*, pages 11–12, 2002.

[Bie05] A. Biemiller. Addressing developmental patterns in vocabulary: implications for choosing words for primary grade vocabulary instruction. *Teaching and learning vocabulary: Bringing research to practice*, pages 223–242, 2005.

[BL08] R. Barzilay and M. Lapata. Modeling local coherence: an entity-based approach. *Computational Linguistics*, 34(1):141–148, 2008.

[BLS13] A. Barate, L.A. Ludovico, and E. Santucci. A semantics-driven approach to lyrics segmentation. In *Proc. of International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 73–79, 2013.

[BMEWL11] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of the International Society of Music Information Retrieval*, volume 2, pages 591–596, 2011.

[BR06] S. Banerjee and A.I. Rudnicky. A texttiling based approach to topic boundary detection in meetings. In *Proc. of the International Conference on Spoken Language Processing*, 2006.

[CDB05] S. J. Cunningham, J. S. Downie, and D. Bainbridge. "the pain, the pain": modelling music information behavior and the songs we hate. In *Proc. of the International Society of Music Information Retrieval*, pages 474–477, 2005.

[Cho00] F. Y. Choi. Advances in domain independent linear text segmentation. In *Proc. of the 1st North American Chapter* *of the Association for Computational Linguistics Conference*, pages 26–33, 2000.

[CM10] S. A. Crossley and D. S. Mcnamara. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proc. of the 32nd annual conference of the Cognitive Science Society*, pages 984–989, 2010.

[COM02] L. Carlson, M.E. Okurowski, and D. Marcu. Rst discourse treebank. *Linguistic Data Consortium*, 2002.

[CVG$^+$08] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. In *Proc. of IEEE*, volume 96, pages 668–696, 2008.

[CVK04] C. H. L. Costa, J. D. Valle, and A. L. Koerich. Automatic classification of audio data. In *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 562–567, 2004.

[CWHM01] F. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *Proc. of the Conference on Empirical Methods on Natural Language Processing*, volume 4, pages 109–117, 2001.

[CYS07] Z. Cataltepe, Y. Yaslan, and A. Sonmez. Music genre classification using midi and audio features. *EURASIP Journal on Advances in Signal Processing*, 1:1–8, 2007.

[DFTW01] R. Dannenberg, J. Foote, G. Tzanetakis, and C. Weare. Panel: new directions in music information retrieval. In *Proc. of the International Computer Music Conference*, pages 295–340, 2001.

[Dow03] J. S. Downie. Music information retrieval. *Annual review of information science and technologyg*, 37(1):295–340, 2003.

[EAC07] M. Elsner, J. Austerweil, and E. Charniak. A unified local and global model for discourse coherence. In *Proc. of the Conference on Human Language Technology and North Amer-*

ican Chapter of the Association for Computational Linguistics, pages 436—447, 2007.

[EXFW15] R.J. Ellis, Z. Xing, J. Fang, and Y. Wang. Quantifying lexical novelty in song lyrics. In *Proc. of the International Society of Music Information Retrieval*, pages 694–700, 2015.

[Fab81] F. Fabbri. A theory of musical genres: two applications. *Popular Music Perspectives*, pages 52–81, 1981.

[FD02] J. Futrelle and J. S. Downie. Interdisciplinary communities and research issues in music information retrieval. In *Proc. of the International Society of Music Information Retrieval*, volume 2, pages 215–221, 2002.

[FEH09] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237, 2009.

[FH14] 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.

[FJHE10] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proc. of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284, 2010.

[FS14] M. Fell and C. Sporleder. Lyrics-based analysis and classification of music. In *Proc. of the International Conference on Computational Linguistics*, pages 620–631, 2014.

[GC11] J. D. Gibbons and S. Chakraborti. Nonparametric statistical inference, 2011.

[GKMC99] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 121–128, 1999.

[GMFLJ03] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 562–569, 2003.

[GMLC04] 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.

[HB10] Hirjee H. and D. G. Brown. Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 5(4):121–145, 2010.

[HD10] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: a feature analysis. In *Proc. of the International Society of Music Information Retrieval*, pages 619–624, 2010.

[HDE09] X. Hu, J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. In *Proc. of the International Society of Music Information Retrieval*, pages 411–416, 2009.

[HDWE05] X. Hu, J. S. Downie, K. West, and A. F. Ehmann. Mining music reviews: promising preliminary results. In *Proc. of the International Society of Music Information Retrieval*, pages 536–539, 2005.

[Hea97] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[Her05] D. Herd. Changes in the prevalence of alcohol use in rap song lyrics, 1979–97. *Addiction*, 100(9):1258–1269, 2005.

[HK06] J. Han and M. Kamber. Data mining concept and techniques, 2006.

[HPDI10] H. Hernault, H. Prendinger, D. A. Duverle, and M. Ishizuka. Hilda: a discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3), 2010.

[ISY14] T. Inui, M. Saito, and M. Yamamoto. Automatic news source detection in twitter based on text segmentation. In *Proc. of the 28th Pacific Asia Conference on Language, Information, and Computation*, pages 195–203, 2014.

[KKP08] F. Kleedorfer, P. Knees, and T. Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Proc. of the International Society of Music Information Retrieval*, pages 287–292, 2008.

[KL03] S. S. Keerthi and C. J. Lin. Asymptotic behaviors of support vector machines. *Neural Computation*, 15(7):l667–1689, 2003.

[Lam08] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[Law11] J. F. Lawless. Statistical models and methods for lifetime data, 2011.

[LGH08] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Proc. of the IEEE International Conference on Machine Learning and Applications*, pages 688–693, 2008.

[LLH14] J. Li, R. Li, and E. H. Hovy. Recursive deep models for discourse parsing. In *Proc. of the Conference on Empirical Methods on Natural Language Processing*, pages 2061–2069, 2014.

[LNK11] Z. Lin, H.T. Ng, and M.Y. Kan. Automatically evaluating text coherence using discourse relations. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 997–1006, 2011.

[LNK14] 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.

[LO04] T. Li and M. Ogihara. Music artist style identification by semi-supervised learning from both lyrics and content. *ACM Multimedia*, pages 364–367, 2004.

[LOL03] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–289, 2003.

[LPC+11] 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 *Proc. of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, 2011.

[Mar97] D. Marcu. From discourse structures to text summaries. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, volume 97, pages 82–88, 1997.

[MNR08] R. Mayer, R. Neumayer, and A. Rauber. Rhyme and style features for musical genre classification by song lyrics. In *Proc. of the International Society of Music Information Retrieval*, pages 337–342, 2008.

[MT88] B. 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.

[NKL+13] 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 classification. In *Proc. of the Conference on Empirical Methods on Natural Language Processing*, pages 12–23, 2013.

[NR07] R. Neumayer and A. Rauber. Integration of text and audio features for genre classification in music information retrieval. In *Proc. of the European Conference on Information Retrieval*, pages 724–727, 2007.

[NRSM10] A. Nanopoulos, D. Rafailidis, P. Symeonidis, and Y. Manolopoulos. Musicbox: personalized music recommendation based on cubic analysis of social tags. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 18, pages 407–412, 2010.

[PA04] 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.

[PAA14] S. Pudaruth, S. Amourdon, and J. Anseline. Automated generation of song lyrics using cfgs. In *Proc. of the International Conference on Contemporary Computing*, pages 613–616, 2014.

[PDL$^+$08] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A.K. Joshi, and B.L. Webber. The penn discourse treebank 2.0. *Language Resources and Evaluation*, 2008.

[PL13] F. L. Parra and E. Leon. Unsupervised tagging of spanish lyrics dataset using clustering. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 130–143, 2013.

[PLUP13] J. M. Pereaortega, E. Lloret, L. A. Urenalopez, and M. Palomar. Application of text summarization techniques to the geographical information retrieval task. *Expert Systems With Applications*, 40(8):2966–2974, 2013.

[PMD$^+$07] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, and A. Joshi. The penn discourse treebank 2.0 annotation manual. *Language Resources and Evaluation*, 24(1):2961–2968, 2007.

[Pre96] A. J. Prevos. The evolution of french rap music and hip hop culture in the 1980s and 1990s. *French Review*, pages 713–725, 1996.

[Pur11] M. Purver. Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317, 2011.

[RB12] M. Riedl and C. Biemann. Topictiling: a text segmentation algorithm based on lda. In *Proc. of the Annual Meeting on Association for Computational Linguistics Student Research Workshop*, pages 37–42, 2012.

[Sal11] K. Salley. On the interaction of alliteration with rhythm and metre in popular music. *Popular Music*, 30(3):409–432, 2011.

[SCSU13] S. Scardapane, D. Comminiello, M. Scarpiniti, and A. Uncini. Music classification using extreme learning machines. In *Proc. of the 8th International Symposium on Image and Signal Processing and Analysis*, pages 377–381, 2013.

[SDP12] Y. Song, S. Dixon, and M. Pearce. A survey of music recommendation systems and future perspectives. In *Proc. of International Symposium on Computer Music Modeling and Retrieval*, 2012.

[SKK08] N. C. Silla, A. L. Koerich, and C. A. Kaestner. Feature selection in automatic music genre classification. *Information Systems Management*, pages 39–44, 2008.

[SM03] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156, 2003.

[Ste14] L. Sterckx. *Topic detection in a million songs*. PhD thesis, Ghent University, 2014.

[Str88] W. Straw. Music video in its contexts: popular music and post-modernism in the 1980s. *Popular Music*, 7(3):247–266, 1988.

[SWR08] S. Somasundaran, J. Wiebe, and J. Ruppenhofer. Discourse level opinion interpretation. In *Proc. of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 801–808, 2008.

[SZM06] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[TC02] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, volume 10, pages 293–302, 2002.

[TI14] K. Takahashi and M. Inoue. Multimodal dialogue segmentation with gesture post-processing. In *Proc. of the Inter-*

national Conference on Language Resources and Evaluation, pages 3433–3437, 2014.

[Web04] B. Webber. D-ltag: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779, 2004.

[WEK12] B. Webber, M. Egg, and V. Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490, 2012.

[WG05] F. Wolf and E. Gibson. Representing discourse coherence: a corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

[WJW10] C. C. Wang, J. S. R. Jang, and W. Wang. An improved query by singing/humming system using melody and lyrics information. In *Proc. of the International Conference on Contemporary Computing*, pages 45–50, 2010.

[WPH$^+$09] B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky, and R. Sauri. Classification of discourse coherence relations: an exploratory study using multiple knowledge sources. In *Proc. of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 117–125, 2009.

[WYN04] Z. L. Wang, H. Yu, and F. Nishino. Automatic special type website detection based on webpage type classification. In *Proc. of the First International Workshop on Web Engineering*, 2004.

[XMS05] C. Xu, N. C. Maddage, and X. Shao. Automatic music classification and summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450, 2005.