A Parallel Corpus of Music and Lyrics Annotated with Emotions

Carlo Strapparava*, **Rada Mihalcea**°, **Alberto Battocchi**†

*FBK-irst, Trento - Italy
°University of North Texas, Denton TX - USA
† DISI - University of Trento - Italy

Abstract

In this paper, we introduce a novel parallel corpus of music and lyrics, annotated with emotions at line level. We first describe the corpus, consisting of 100 popular songs, each of them including a music component, provided in the MIDI format, as well as a lyrics component, made available as raw text. We then describe our work on enhancing this corpus with emotion annotations using crowdsourcing. We also present some initial experiments on emotion classification using the music and the lyrics representations of the songs, which lead to encouraging results, thus demonstrating the promise of using joint music-lyric models for song processing.

Keywords: emotion recognition, multimodal processing, music analysis

"Killing me softly with his song".
– C. Fox and N. Gimbel

1. Introduction

Popular songs exert a lot of power on people, both at an individual level as well as on groups, mainly because of the message and emotions they convey. Songs can lift our moods, make us dance, or move us to tears. Songs are able to embody deep feelings, usually through a combined effect of both the music and the lyrics. Song writers know that music and lyrics have to be coherent, and the art of shaping words for music involves precise techniques of creative writing, using elements of grammar, phonetics, metrics, or rhyme.

From a computational point of view, while there were several studies that dealt with music analysis (especially exploiting the popular MIDI format) (Das et al., 2000; Cataltepe et al., 2007), or with the processing of lyrics by using natural language processing techniques (Mahedero et al., 2005; Yang and Lee, 2009), a strict combination of lyrics and music dimensions has not been exploited.

In this paper, we introduce a corpus of songs with a strict alignment between notes and words, which can be regarded and used as a *parallel* corpus suitable for common parallel corpora techniques previously used in computational linguistics. The corpus consists of 100 popular songs, such as *"On Happy Days"* or *"All the Time in the World,"* covering famous interpreters such as the Beatles or Sting. For each song, both the music (in MIDI format) and the lyrics (as raw text) are included, along with an alignment between the MIDI features and the words. Moreover, because of the important role played by emotions in songs, the corpus also embeds manual annotations of six basic emotions collected via crowdsourcing.

2. Background

The computational treatment of music is a very active research field. The increasing availability of music in digital format (e.g., MIDI) has motivated the development of tools for music accessing, filtering, classification, and retrieval. For instance, the task of music retrieval and music recommendation has received a lot of attention from both the arts and the computer science communities (see for instance (Orio, 2006) for an introduction to this task).

There are several works on MIDI analysis. We report mainly those that are relevant for the purpose of the present work. For example (Das et al., 2000) describes an analysis of predominant up-down motion types within music, through extraction of the kinematic variables of music velocity and acceleration from MIDI data streams. (Cataltepe et al., 2007) addresses music genre classification using MIDI and audio features, while (Wang et al., 2004) automatically aligns acoustic musical signals with their corresponding textual lyrics.

MIDI files are typically organized into one or more parallel "tracks" for independent recording and editing. A reliable system to identify the MIDI track containing the *melody*[1] is very relevant for music information retrieval, and there are several approaches that have been proposed to address this issue (Rizo et al., 2006; Velusamy et al., 2007).

Regarding natural language processing techniques applied on lyrics, there have been a few studies that

[1] A melody can be defined as a 'cantabile' sequence of notes, usually the sequence that a listener can remember after hearing a song.

mainly exploit just the lyrics component of the songs, while ignoring the musical component. For instance, (Mahedero et al., 2005) deals with language identification, structure extraction, and thematic categorization for lyrics. (Yang and Lee, 2009) approach the problem of emotion identification in lyrics.

3. Corpus Description

MIDI is an industry-standard protocol that enables electronic musical instruments, computers and other electronic equipment to communicate and synchronize with each other. Unlike analog devices, MIDI does not transmit an audio signal: it sends event messages about musical notation, pitch, and intensity, control signals for parameters such as volume, vibrato, and panning, and cues and clock signals to set the tempo. As an electronic protocol, it is notable for its widespread adoption throughout the music industry. MIDI files are typically created using computer-based sequencing software that organizes MIDI messages into one or more parallel "tracks" for independent recording, editing, and playback. In most sequencers, each track is assigned to a specific MIDI channel, which can be then associated to specific instrument patches. MIDI files can also contain lyrics, which can be displayed scrolling in synchronous with the music.

Given the non-homogeneous quality of the MIDI files available on the Web, we asked a professional MIDI provider for high quality MIDI files produced for singers and musicians. We collected 100 songs with the respective MIDI files, containing also lyrics that are synchronized with the notes. The genres of the songs fall mainly into pop, rock and evergreen categories. On these MIDI files, the melody channel was unequivocally decided by the provider, making it easier to extract the music and the corresponding lyrics.

SONGS	100
SONGS IN "MAJOR" KEY	59
SONGS IN "MINOR" KEY	41
LINES	4,976
ALIGNED SYLLABLES / NOTES	34,045

Table 1: Some statistics of the corpus

Figure 3. shows an example from the corpus, consisting of the first two lines in the Beatles' song *A hard day's night*. We explicitly encode the following features. At the song level, the key of the song (e.g., G major, C minor). At the line level, we represent the *raising*, which is the musical interval (in half-steps) between the first note in the line and the most important note (i.e., the note in the line with the longest duration), as well as the manual emotion annotations. Finally, at the note level, we encode the time code of the note with respect to the beginning of the song; the note aligned with the corresponding syllable; the degree of the note with relation to the key of the song; and the duration of the note. Table 1 shows some statistics collected on the entire corpus.

4. Emotion Annotations with Mechanical Turk

One of our goals in the construction of the parallel corpus of music and lyrics is to also include manual annotations that can be potentially useful for future research studies. While annotations such as part of speech tags and syntactic trees can be produced with reasonable accuracy using existing state-of-the-art tools, other annotations such as emotion and sentiment require more human effort. Thus, in this stage of our project, we focused on the manual annotation of emotions at line level, with more annotation layers to be added in future work.

Following previous work on emotion annotation of text, we use the six basic emotions proposed in (Ekman, 1993): (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE). To collect the annotations, we use the Amazon Mechanical Turk service, which was previously found to produce reliable annotations with a quality comparable to those generated by experts (Snow et al., 2008).

The annotations are collected at line level, with a separate annotation for each of the six emotions. We collect numerical annotations using a scale between 0 and 10, with 0 corresponding to the absence of the corresponding emotion, and 10 corresponding to the highest intensity. The annotators are instructed to: (1) Score the emotions from the writer perspective, not their own perspective; (2) To read and interpret each line in context, that is they were asked to read and understand the entire song before producing any annotations; (3) To produce the six emotion annotations independent from each other, accounting for the fact that a line could contain none, one, or multiple emotions. They were also given three different examples to illustrate the annotation. Each HIT (i.e., annotation session) contains an entire song, with a number of lines ranging from 14 to 110. On average, there are 50 lines per song.

While the use of crowdsourcing for data annotations can result in a large number of annotations in a very short amount of time, it also has the drawback of potential spamming that can interfere with the quality of the annotations. To address this aspect, we use two different techniques. First, in each song we insert a "checkpoint" at a random position in the song – a fake line that reads "Please enter 7 for each of the six emotions." Those annotators who do not follow this concrete instruction are deemed as spammers who produce annotations without reading the content

```
<song filename=AHARDDAY.m2a>
<key time=0>G major</key>
<line pvers=1 raising=3 anger=1.5 disgust=0.7 sadness=2.5 surprise=0.8 >
<token time=5040 orig−note=B degree=3 duration=210>IT</token>
<token time=5050 orig−note=B degree=3 duration=210>′S </token>
<token time=5280 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=5520 orig−note=B degree=3 duration=210>A </token>
<token time=5760 orig−note=D′ degree=5 duration=810>HARD </token>
<token time=6720 orig−note=D′ degree=5 duration=570>DAY</token>
<token time=6730 orig−note=D′ degree=5 duration=570>′S </token>
<token time=7440 orig−note=D′ degree=5 duration=690>NIGHT</token>
</line>
<line pvers=2 raising=5 anger=3.5 disgust=2 sadness=1.2 surprise=0.2 >
<token time=8880 orig−note=C′ degree=4 duration=212>AND </token>
<token time=9120 orig−note=D′ degree=5 duration=210>I</token>
<token time=9130 orig−note=D′ degree=5 duration=210>′VE </token>
<token time=9360 orig−note=C′ degree=4 duration=210>BEEN </token>
<token time=9600 orig−note=D′ degree=5 duration=210>WOR</token>
<token time=9840 orig−note=F′ degree=7− duration=930>KING </token>
<token time=10800 orig−note=D′ degree=5 duration=210>LI</token>
<token time=11040 orig−note=C′ degree=4 duration=210>KE </token>
<token time=11050 orig−note=C′ degree=4 duration=210>A </token>
<token time=11280 orig−note=D′ degree=5 duration=330>D</token>
<token time=11640 orig−note=C′ degree=4 duration=90>O</token>
<token time=11760 orig−note=B degree=3 duration=330>G</token>
</line>
```

Figure 1: Two lines of a song in the corpus: *It-'s been a hard day-'s night, And I-'ve been wor-king li-ke a d-o-g*

of the song, and thus removed. Second, for each remaining annotator, we calculate the Pearson correlation between her emotion scores and the average emotion scores of all the other annotators. Those annotators with a correlation below 0.4 with the average of the other annotators are also removed, thus leaving only the reliable annotators in the pool.

For each song, we start by asking for ten annotations. After spam removal, we are left with about two-five annotations per song. The final annotations are produced by averaging the emotions scores produced by the reliable annotators. Figure 3. shows an example of the emotion scores produced for two lines. Overall, the correlation between the remaining reliable annotators was calculated as 0.73, which represents a strong correlation.

To illustrate the distribution of the emotions in the corpus, for each of the six emotions, Table 2 shows the number of lines that had that emotion present (i.e., the score of the emotion was different from 0), as well as the average score for that emotion over all 4,976 lines in the corpus. Perhaps not surprisingly, the emotions that are dominant in the corpus are JOY and SADNESS – which are the emotions that are often invoked by people as the reason behind a song.

Note that the emotions do not exclude each other: i.e., a line that is labeled as containing JOY may also contain a certain amount of SADNESS, which is the reason for the high percentage of songs containing both JOY and SADNESS. The emotional load for the overlapping emotions is however very different. For instance, the lines that have a JOY score of 5 or higher have an average SADNESS score of 0.34. Conversely, the lines with a SADNESS score of 5 or higher have a JOY score of 0.22.

Emotion	Number lines	Average
ANGER	2,516	0.95
DISGUST	2,461	0.71
FEAR	2,719	0.77
JOY	3,890	3.24
SADNESS	3,840	2.27
SURPRISE	2,982	0.83

Table 2: Emotions in the corpus of 100 songs: number of lines including a certain emotion, and average emotion score computed over all the 4,976 lines.

5. Preliminary Classification Experiments

To explore the usefulness of the joint music/text representation in this corpus of songs, we run a preliminary experiment for emotion recognition in songs, which relies on both music and text features. We use the corpus of 100 songs, which at this stage has full lyrics, text, and emotion annotations. We use a simple bag-of-words representation, which is fed to a machine learning classifier. We run two comparative experiments: one that uses only the lyrics and one that uses both the lyrics and the notes for a joint model of music and lyrics.

Emotion	Baseline	Textual	Musical	Textual and Musical
ANGER	89.27%	91.14%	89.63%	92.40%
DISGUST	93.85%	94.67%	93.85%	94.77%
FEAR	93.58%	93.87%	93.58%	93.87%
JOY	50.26%	70.92%	61.95%	75.64%
SADNESS	67.40%	75.84%	70.65%	79.42%
SURPRISE	94.83%	94.83%	94.83%	94.83%
AVERAGE	81.53%	86.87%	84.08%	88.49%

Table 3: Evaluations using a coarse-grained binary classification.

We transform the task into a binary classification task by using a threshold empirically set at 3. If the score for an emotion is below 3, we record it as "absent," whereas if the score is equal to or above 3, we record it as "present."

For the classification, we use Support Vector Machines (SVM), which are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin (Vapnik, 1995). Applications of SVM classifiers to text categorization led to some of the best results reported in the literature (Joachims, 1998).

Table 3 shows the results obtained for each of the six emotions, and for the three major settings that we consider: textual features only, musical features only, and a classifier that jointly uses the textual and the musical features. The classification accuracy for each experiment is reported as the average of the accuracies obtained during a ten-fold cross-validation on the corpus. The table also shows a baseline, computed as the average of the accuracies obtained when using the most frequent class observed on the training data for each fold.

As seen from the table, on average, the joint use of textual and musical features is beneficial for the classification of emotions. Perhaps not surprisingly, the effect of the classifier is stronger for those emotions that are dominant in the corpus, i.e., JOY and SADNESS (see Table 2). The improvement obtained with the classifiers is much smaller for the other emotions (or even absent, e.g., for SURPRISE), which is also explained by their high baseline of over 90%.

6. Conclusions

Popular songs express universally understood meanings and embody experiences and feelings common to everyone, usually through a combined effect of both the music and the lyrics. In this paper, we introduced a novel parallel corpus of music and lyrics, annotated with emotions at line level. This resource can be regarded and used as a parallel corpus suitable for common parallel corpora techniques previously used in computational linguistics. To explore the usefulness of the joint music/text representation, we reported on a preliminary experiment for emotion recognition in songs, with promising results.

Acknowledgments

Carlo Strapparava was partially supported by a Google Research Award. Rada Mihalcea's work was in part supported by the National Science Foundation award #0917170. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

Z. Cataltepe, Y. Yaslan, and A. Sonmez. 2007. Music genre classification using MIDI and audio features. *Journal on Advances in Signal Processing*.

M. Das, D. Howard, and S. Smith. 2000. The kinematic analysis of motion curves through MIDI data analysis. *Organised Sound*, 5(1):137–145.

P. Ekman. 1993. Facial expression of emotion. *American Psychologist*, 48:384–392.

T. Joachims. 1998. Text categorization with Support Vector Machines: learning with mny relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany.

J. Mahedero, A. Martinez, and P. Cano. 2005. Natural language processing of lyrics. In *Proceedings of MM'05*, Singapore, November.

N. Orio. 2006. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, November.

D. Rizo, P. Ponce de Leon, C. Perez-Sancho, A. Pertusa, and J. Inesta. 2006. A pattern recognition approach for melody track selection in MIDI files. In *Proceedings of 7th International Symposyum on Music Information Retrieval (ISMIR-06)*, pages 61–66, Victoria, Canada, October.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

S. Velusamy, B. Thoshkahna, and K. Ramakrishnan. 2007. Novel melody line identification algorithm for polyphonic MIDI music. In *Proceedings of 13th International Multimedia Modeling Conference (MMM 2007)*, Singapore, January.

Y. Wang, M. Kan, T. Nwe, A. Shenoy, and J. Yin. 2004. LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of MM'04*, New York, October.

D. Yang and W. Lee. 2009. Music emotion identification from lyrics. In *Proceedings of 11th IEEE Symposium on Multimedia*.