

Keyword Generation for Lyrics

Bin Wei

Department of Computer Science
University of Rochester

Chengliang Zhang¹

Microsoft Corporation

Mitsunori Ogiwara²

Department of Computer Science
University of Miami

1. Introduction

Problem studied: Generating keywords for lyrics

Many song lyrics are available through web databases

1. Analysis of song lyrics appears to be challenging because:

- Song lyrics are short and subtle, so the standard word frequency-based approach may not work
- There are lines that offer “background” that are not central to the theme
- Lyrics makes subtle use of words
- Lines in song lyrics are often incomplete

This work

- Use sentence-level clustering to separate the topic from the background
- Use WordNet relation links to find keywords
- Test on the Digital Tradition Database

2. The Method

Step 1: Preprocessing of a lyric

- Parse each line using the Stanford parser. Each line is represented as a represented as a collection of dependencies, of the form (*relation, governor, dependent*)

Step 2: Sentence-level clustering

1. Compute similarity between two dependencies, D and E , by:

$$\text{Sim}(D,E) = \text{Sim}(\text{governor-of-}D, \text{governor-of-}E) + \text{Sim}(\text{dependent-of-}D, \text{dependent-of-}E).$$

Here the similarity on the R.H.S. is the Lesk measurement of word similarity after word disambiguation (Patwardhan et al., 2003)

2. Compute similarity between two sentences, A and B , by:

$$\text{Sim}(A, B) = \max\{ \text{Sim}(D,E) \mid D \text{ and } E \text{ are dependencies appearing in } A \text{ and } B \}$$

3. Use the data-driven clustering method of (Azran and Gaharamani, 2006) to cluster sentences

Step 3: Representative selection

- For each word (permissible categories = nouns, adjectives, adverbs, and verbs) in each cluster, compute the sum of its distance to the other words in the cluster.
- Select top 10 words as the representatives of the cluster

Step 4: Candidate generation and final selection

- From each representative word for a cluster (there are 10 of them), follow the WordNet link (even including the words appearing in the gloss) for 20 steps and produce the set of reachable words from it.
- Rank the word according to the number of representative nodes from which it is reachable and select top 15.

Step 5: Cluster selection and keyword output

- Hypothesis: The words related to the topics of a lyric are shared with many other lyrics while the words related to the background aren't.*

Given a lyric L , for each cluster C of L , do the following:

- Collect all lyrics M not equal to C , such that at least one cluster of L has at least 3 common candidate words with C
- For each such lyric M , identify the cluster that maximizes the number of common words with C
- Compute the average max-size with respect to C
- Choose as the topic cluster of L the cluster with the largest average max-size
- Output as the keywords the candidate words for the cluster

3. Evaluation

Data

- The DigiTrad database, containing about 8K folk songs with keywords
- Two groups of songs
 - SONG1: 400+ songs labeled “murder” or “marriage”
 - SONG2: 400+ songs labeled “political” or “religion”

Evaluation 1

- Extract top 5 keywords from each lyric
- Use the data-driven clustering to obtain cluster Many song lyrics are available through web databases
- Compare against the results of (Scott and Matwin, 1998) (The database has slightly changed since then)

Method	Data	Error Rate
Previous	SONG1	30.23%
New	SONG1	28.14%
Previous	SONG2	32.64%
New	SONG2	31.22%

Evaluation 2

- Extract top 5 keywords from each lyric
- Check whether the DigiTrad keyword appears

Label	Size	Appears in Top5	Appears in Top 5 for some cluster
“marriage”	195	47	64
“murder”	212	50	78
“poverty”	94	25	35
“school”	29	9	11

4. Conclusion

- Slight improvements over (Scott and Matwin, 1998)
- Need further exploration to test the efficacy of the method