Keyword Generation for Lyrics

Bin Wei
Department of Computer Science
University of Rochester

Chengliang Zhang[1]
Microsoft Corporation

Mitsunori Ogihara[2]
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:
   a) Song lyrics are short and subtle, so the standard word frequency-based approach may not work
   b) There are lines that offer "background" that are not central to the theme
   c) Lyrics makes subtle use of words
   d) 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:

   Sim(*D*,*E*) = Sim(governor-of-*D*, governor-of-*E*) + Sim(dependent-of-*D*, 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:

   Sim(*A*, *B*) = max{ Sim(*D*,*E*) | *D* and *E* are dependencies appearing in *A* and *B* }

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

Step 3: Representative selection

1. 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.
2. Select top 10 words as the representatives of the cluster

Step 4: Candidate generation and final selection

1. 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.
2. 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:

1. 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
2. For each such lyric *M*, identify the cluster that maximizes the number of common words with *C*
3. Compute the average max-size with respect to *C*
4. Choose as the topic cluster of *L* the cluster with the largest average max-size
5. 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

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

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

Evaluation 2

1. Extract top 5 keywords from each lyric
2. 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

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

[1, 2] Work done while at U. Rochester. [2] Leave from U. Rochester