Tamil lyrics corpus: Analysis and Experiments

Dhivya Chinnappa

Thomson Reuters
dhivya.infant@gmail.com

Praveenraj Dhandapani

Intellect Design Arena Ltd.
praveenraj0904@gmail.com

Synopsis

Introduction

Background

Related works

Corpus Analysis

Experimental results

Future research

Conclusion

Introduction

- Tamil - classical Dravidian language (Tamil Nadu, Srilanka)
- 2,600 years old
- Tamil movies are integral part of Tamilians across the world
- Average of 90 Tamil movies released per year (Krishnan and Sakkthivel, 2010)
- Movies are musically and lyrically rich

தமிழ்

Background - songs enhance an emotion

Song lyrics enhance an emotion in the movie

- scenes in the movie not necessarily important

Background - songs complement an emotion

Song lyrics complement an emotion in the movie

- Scenes in the song are integral part of the movie

Background - songs convey political messages

- emphasizes on emotions and sentiments outside the movie line

Related works

- Tamil lyrics dataset in Kaggle (Subramanian, 2020)
- Aspect identification
  - base, mood, style - Sundara Kanchana and Ganapathy (2017)
  - pleasantness scoring - Beulah S. E. et al. (2011)
- Automatic lyric generation
  - sequence labeling - Ramakrishnan A et al. (2009)
  - several methods from trigram approach to using images as input -Sridhar et al. (2013, 2014, 2015, 2016)
- Lyrics analysis
  - Ranganathan et al. (2013)

Corpus

- Downloaded from https://allnewlyrics.com/
  - Tamil lyrics
  - Transliterated English lyrics
  - Lyricist name
  - Movie name
  - Music director
  - Year

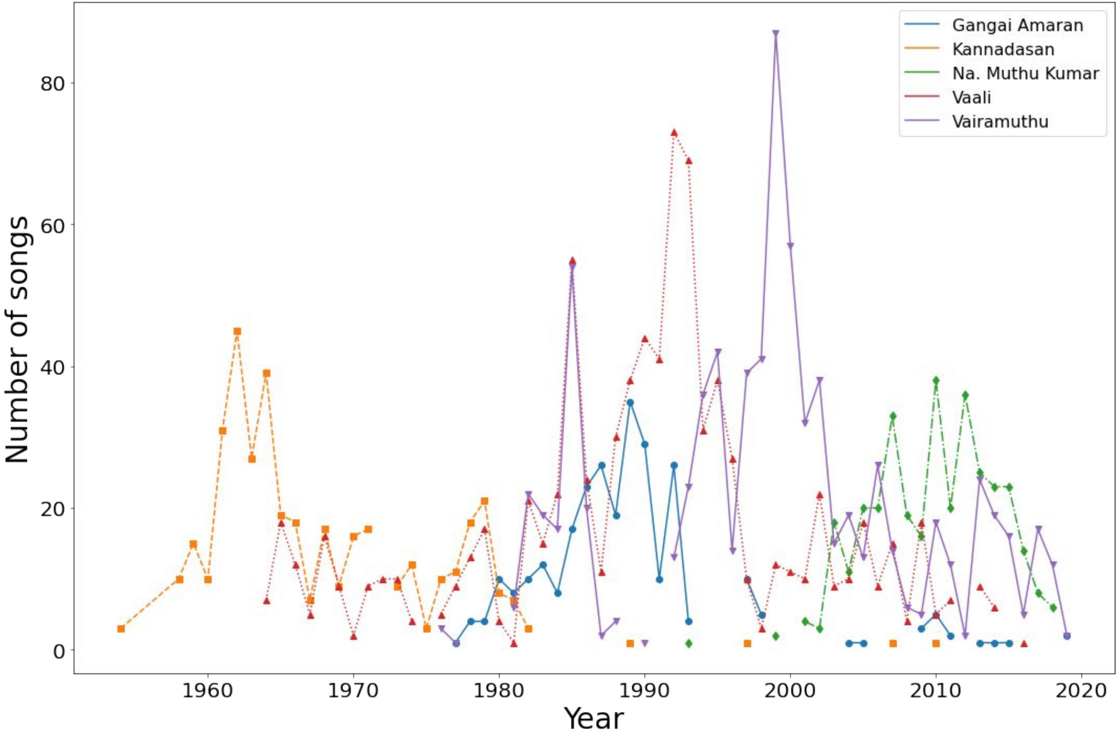
- Available at

  https://github.com/praveenraj0904/tamillyricscorpus

Lyrics	5449
Lyricists	324
Music directors	151
Movies	1147
Words	1,309,425
Unique words	129,614
Years	1954-2019
Maximum words/song	653
Maximum words/song	17
Maximum words/song	188

All statistics correspond to the corpus, and may not reflect the contribution of a lyricist in real time.

Corpus analysis - Songs by lyricists across years



Corpus analysis - Lyrics based similarity

		Jacc. sim.	cos. Sim. (TFIDF)
Vaali	Kannadasan	0.76	0.956
	Vairamuthu	0.94	0.972
	Na. Muthu Kumar	0.89	0.956
	Gangai Amaran	0.76	0.969
Kannadasan	Vairamuthu	0.71	0.948
	Na. Muthu Kumar	0.77	0.91
	Gangai Amaran	0.85	0.922
Vairamuthu	Na. Muthu Kumar	0.90	0.964
	Gangai Amaran	0.74	0.939
Na. Muthu Kumar	Gangai Amaran	0.81	0.924

Image courtesy:
https://cinema.maalaimalar.com/cinema/cinemanews/2016/08/14144932/1032533/Kavignar-Vairamuthu-condolence-to-NaMuthukumar-Death.vpf
https://www.indiaglitz.com/tuesday-trivia-shankar-a-r-rahman-replaced-vairamuthu-with-vaali-telugu-news-215063

Experiments - Lyricist identification

- Lyricist identification between top 2 lyricists
- 1670 instances (Vaali - 873, Kannadasan - 797)

	Multilingual BERT			Indic-NLP			Tamillion BERT		
	P	R	F	P	R	F	P	R	F
Vaali	0.62	0.57	0.59	0.65	0.75	0.69	0.68	0.74	0.70
Kannadasan	0.56	0.62	0.59	0.67	0.55	0.60	0.68	0.61	0.64
W. Avg	0.59	0.59	0.59	0.66	0.66	0.65	0.68	0.68	0.68

Image courtesy:
https://www.indiatvnews.com/entertainment/bollywood/vaali-s-songs-will-be-eternal-lyricist-thamarai-8752.html
https://www.amazon.in/b?ie=UTF8&node=13832121031

Research ideas

- Language trend analysis
  - Recent songs include more English words
  - Recent songs include meaningless words like jimjikka, rangu rakkara, tasaku tasaku, etc.
- Evaluating transliteration models
- Author style detection
- From related works - Emotion detection, automatic lyric generation

Image credits: https://www.tamilbot.com/en

Conclusion

- New corpus of Tamil lyrics over a range of 65 years
- Elaborate literature review and corpus analysis
  - Only one woman lyricist in the top-10 lyricists
  - Based on the similarity measures we used Vaali’s and Vairamuthu’s lyrics are similar.
- Present experimental results on identifying the lyricist for the top 2 lyricists
- Results show identifying lyricists is challenging

Image credits:
https://www.newsbugz.com/lyricist-thamarai-wiki-biography-age-movies-awards/#/?playlistId=0&videoId=0

- Dhivya Chinnappa
  - dhivya.infant@gmail.com
  - https://www.linkedin.com/in/dhivyachinnappa
  - https://twitter.com/dhivyachinnappa

- Praveenraj Dhandapani
  - praveenraj0904@gmail.com
  - https://www.linkedin.com/in/praveenraj0904/

Image courtesy: Government Oriental Manuscript Library, Madras University, Chennai