3rd CompMusic Workshop, IIT Madras, Chennai

Ontology for Indian Music: An Approach for ontology learning from online music forums

Supervisors
Prof. Pushpak Bhattacharyya
Prof. Preeti Rao

Joe Cheri Ross
IIT Bombay

Objective

Augment music ontology for Indian music with information extracted from online music forums

Why we need to have meta data in ontology along with audio based information ?

Better retrieval of Indian music information

Example query :

Get songs with phrase 'NDNP' and sung by a disciple of D.K. Pattammal

Outline

- Rasikas.org
- Existing work in information extraction
- Thread title processing
- Relation Extraction
  - Named Entity Recognition
- Future Work

Rasikas.org

This work focusses on Carnatic music forum: www.rasikas.org

Subforums can be generally categorized as

- Musicological, artists information
- Reviews, audience feedback

A few instances from the forum:

Ashok Madhav's talk in KGS stated KV Srinivasa Iyengar composed Natajana and Needucharanamule under Thyagaraja's mudra.

This apoorva raga is the Janya of 28th mela HK

This song also has a close resemblance of Misra Piloo or Misra Kapi(except N3).

General characteristics of the text in the forum

- Unstructured
- Ungrammatical
- Presence of sentences with less/no information
- Presence of interrogative and imperative sentences

Existing work in Information Extraction

Rasikas.org

Forum information represented as a network representation to identify popular terms within the forum, as well as relevant co-occurrences and semantic relationships. (M.Sordo, 2012)

Biomedical Domain

Manually and automatically generated pattern based approach is widely used for structured text. (Yu, H, 2002;Yu, H, 2003, Cohen, 2005)

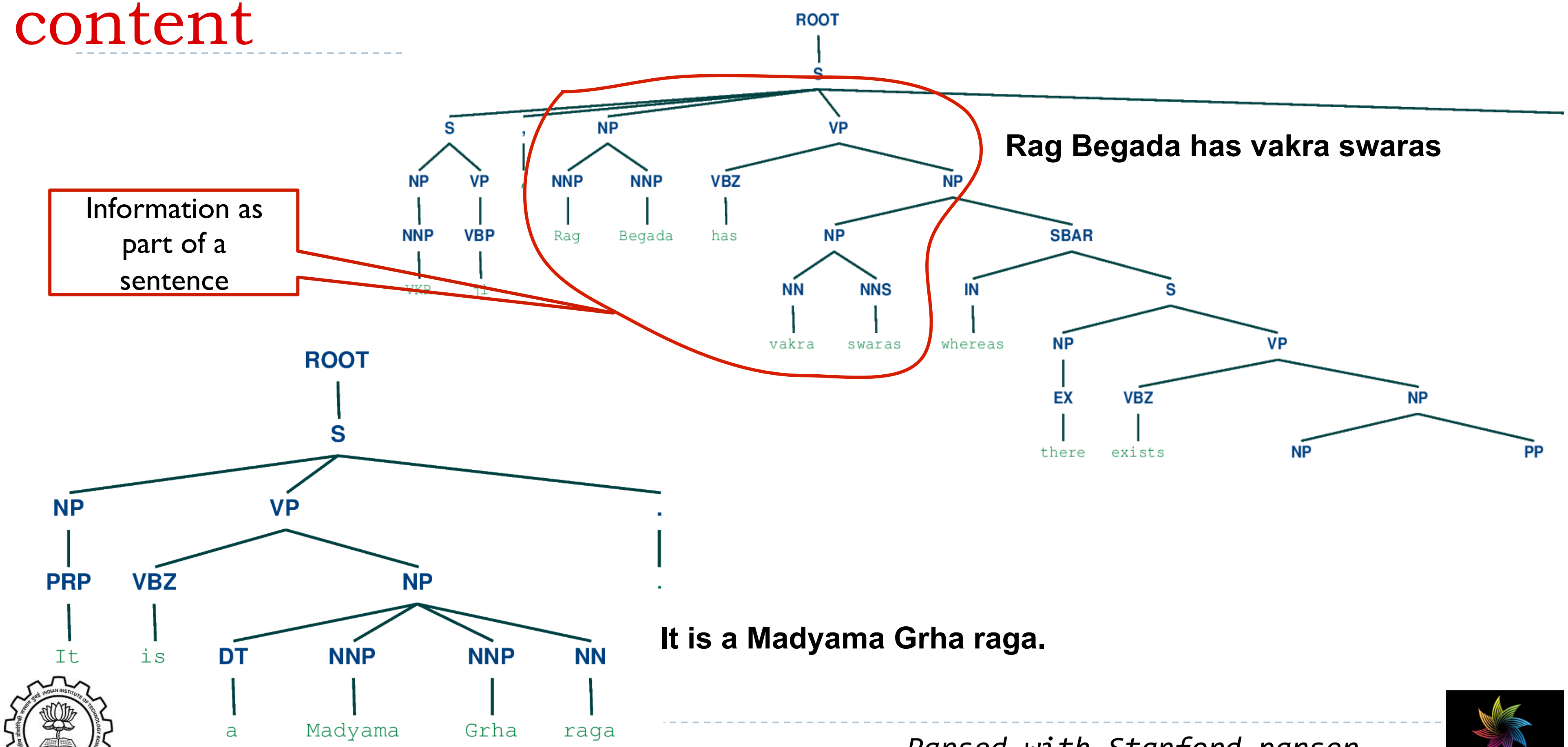
Using a shallow parser and sentence structure analysis techniques, automatic extraction of biological process functions based on Gene Ontology (GO) from text. (Koike, 2005)

Existing work in Information Extraction

Web information

There are approaches which learn extraction rules from corpus and use this for IE. (Muslea, 1998)

An overview of syntactic patterns in forum content



Information as part of a sentence

Rag Begada has vakra swaras

It is a Madyama Grha raga.

Parsed with Stanford parser

Thread Title Processing

Thread title processing

Title of a forum thread conveys the main topic behind the thread

Relevance of thread title processing:

It is a Madyama Grha Raga

It is predominantly an evening raga

Mostly pronoun refers to the title of the forum thread

Anaphora resolution: can be done by processing the thread title

Thread title processing

Thread Titles

Shanmukhapriya & Simhendramadhyama

Chembai Vaidyanatha Bhagavatar

Structure of Thillana

Difference between chauka varnams, pada varnams

From the parse tree of thread titles a syntactic rule can be inferred

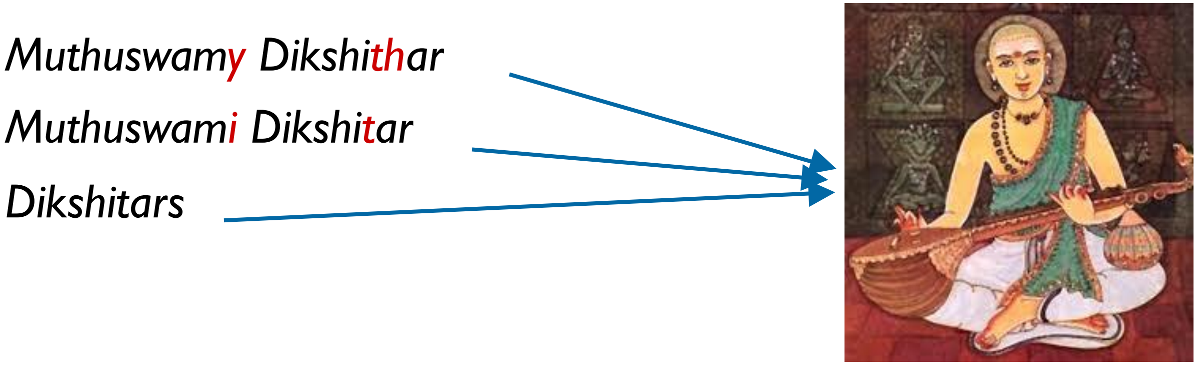
- Topics separated by CONJ*('and',',')

- <Noun> PP <Topics separated by CONJ>

* Conjunction

Thread title processing

Resolve Topic with Dictionary/Ontology



Muthuswamy Dikshithar

Muthuswami Dikshitar

Dikshitars

(Source : www.indianetzone.com)

All the mentioned names refers to the same Muthuswamy Dikshitar

Problem: *From a thread title in the forum how do we identify the identity(semantics) of a topic*

1. Get the topics of the title using the syntactic patterns
2. Identify the corresponding entry in the dictionary/ontology

[1] http://www.indianetzone.com

Thread title processing

Resolve Topic with Dictionary/Ontology

1. Get all the combinations of adjacent words in the phrase.(unigram, bigram, trigram)

words phrase: 'w1 w2 w3'

combinations: w1, w2, w3, w1 w2 , w2 w3 , w1 w2 w3

2. Perform fuzzy matching with the concepts and concepts instances in the dictionary/ontology

3. Tag the words combination with the identified concepts/concept instances

only if the similarity with any concept is greater than a predefined threshold

Relation Extraction

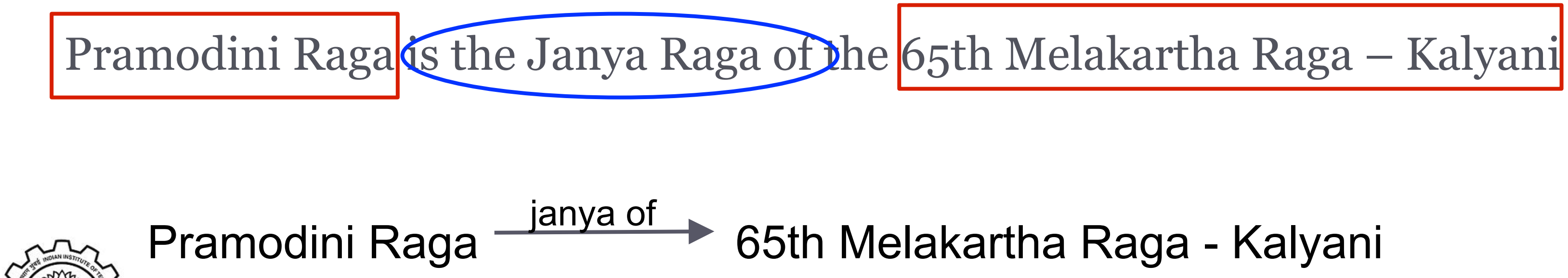
Using Natural Language Processing

Relation Extraction: Basic steps

(i) Identify named entities (Named Entity Recognition- NER)

Ex: Sruti, Hamsadhvani, Purandara Dasa

(ii) Find relation between named entities



Pramodini Raga is the Janya Raga of the 65th Melakartha Raga – Kalyani

Pramodini Raga → janya of → 65th Melakartha Raga - Kalyani

First step to relation extraction: Named Entity Recognition(NER)

Issues with NLTK NER (Bird, 2006)

- Person
- Organization
- GPE

They learnt under Alathur Venkatesa Iyer, the father of Sivasubramania Iyer

The finesse and authority with which they handled compositions like Vidulaku Mrokeda (Mayamalavagowla, Tyagaraja), …….

The effect of such a training is evident in the music of the Alathur Brothers

Alathur Venkatesa Iyer (1895–1958) was a teacher of Carnatic music.

Venkatesa Iyer was instrumental in bringing out a large number of krithis of Maharaja Swathi Thiruna of Travancore.

Trichy J. Venkatraman, Chengelput Ranganathan, Clarinet A. K. C. Natarajan, who though numbering few, have in good stead been the torch bearers of the Alathur style.

Named Entity Recognition
Possible approaches

1. Dictionary based
2. Rule based
3. Machine Learning based

(Ananiadou, 2006)

Named Entity Recognition

Dictionary based approach

We follow dictionary based approach

Why this approach is better for our purpose ?

- **Purpose: Ontology learning**

All the relations involving named entities are to be mapped to the same corresponding instance in the ontology

- **Indian Terminologies**

Carnatic and Hindustani music concepts, instruments etc. This set is limited except for person names.

- **Named entity categories**

NE categories are specific to music domain (Artists, Instruments, Music concept, Location). This is different from the standard NER categories which includes person, location, organization etc.

Named Entity Recognition

Dictionary based approach

Method

1. Get the NP(noun phrases) from the parse tree of sentence in the forum
2. Using n-gram approach* identify the corresponding instance in the ontology

*Mentioned in thread title processing

Named Entity Recognition

Relevance of n-gram string comparison

Solution to :

Given an NP phrase what combination of the words contributes to a name in the dictionary

Example:

Padma Bhushan T. N. Seshagopalan

Sangeet Samrat Chitravina N. Ravikiran

NER: How to develop/add the dictionary ?

- Since the entities related music concepts, instruments is a limited set, dictionary expansion primarily targets artist names
- A good source is wikipedia pages under categories related to Carnatic music

How to expand artists names in dictionary ?

- Identify the named entities of persons from wikipedia sources and add it to the dictionary

NER:How to develop/add the dictionary ?

Approaches

1. Bootstrap approach

2. Score based on the frequency of the component names in the names corpus

NER: How to develop/add the dictionary ?

1. Bootstrap approach

For a given category of named entities, from a set of seed words identify other named entities in the same category

Method (Thelen, 2002)

```
0. Define a set of seed words for the category
   Extract all verb patterns in the corpus
1. Score the verb patterns
2. Get the top (20 + i) verb patterns
3. Identify candidate words through top ranked verb patterns
4. Score candidate words and add top ranked words to lexicon
5. Repeat from step 1
```

NER:How to develop/add the dictionary ?

Bootstrap approach- Results

Wikipedia corpus:

seed words: ['Dikshithar', 'D. K. Pattammal', 'M. L.Vasanthakumari', 'M. S. Subbulakshmi', 'Muthiah Bhagavathar', 'Mysore Vasudevachar', 'Kanchipuram Nayana Pillai', 'Kanchipuram N.S.Krishnaswamy Iyengar', 'Chembai Vaidyanatha Bhagavathar', 'Ariyakudi Ramanuja Iyengar', 'Musiri Subramania Iyer', 'Maharajapuram Viswanatha Iyer', 'Semmangudi Srinivasa Iyer', 'Alathur Brothers', 'G. N. Balasubramaniam', 'Madurai Mani Iyer', 'Alathur Venkatesa Iyer', 'Ramnad Krishnan', 'M. D. Ramanathan', 'S.Ramanathan', 'Mysore V. Ramarathnam', 'K.V. Narayanaswamy', 'Sirkazhi Govindarajan', 'Maharajapuram Santhanam', 'Tanjore S. Kalyanaraman', 'D. K. Jayaraman', 'T. K. Rangachari', 'Vairamangalam Lakshminarayanan', 'Madurai Somu', 'Mavelikkara Prabhakara Varma', 'Neyyattinkara Vasudevan']

Additional relevant candidate words extracted:

'Rangarajan', 'Sastry Sankara', 'Bhagavatar Muthiah', 'Mahadevan Nithyashree', 'Saketharaman S', 'Purushothaman Suguna', 'Vaidhya Rajhesh', 'Jayaraman', 'Pillai',

NER: How to develop/add the dictionary ?

2. Scoring an NE based on existing NEs in corpus

Scoring Names in Database

- Get NEs associated with carnatic music from infobox of wil

- Score the components of the words based on it's occurren

Ex: Muthuswamy Dikshithar → 'Muthuswamy', 'Dikshithar'

Ramaswamy Dikshithar → 'Ramaswamy', 'Dikshithar'

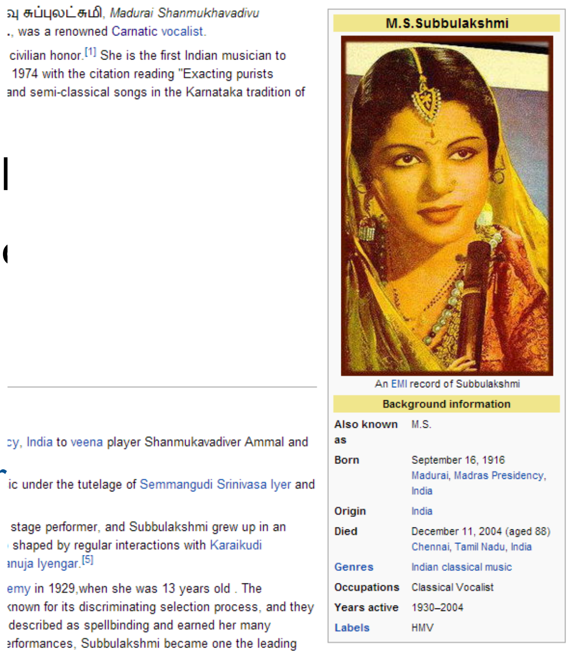
Here score('Dikshithar') greater than score of other 2 words as 'Dikshithar

Deciding a NP(noun phrase) is a name

NP= "n_1 n_2...n_l "

Score(NP)= $\sum_{i=1}^{l} score(n_i)$

NP acceptable as a name if Score(NP) > threshold



வு சுப்புலட்சுமி, *Madurai Shanmukhavadivu*
., was a renowned Carnatic vocalist.

civilian honor.[1] She is the first Indian musician to
1974 with the citation reading "Exacting purists
and semi-classical songs in the Karnataka tradition of

cy, India to veena player Shanmukavadiver Ammal and

ic under the tutelage of Semmangudi Srinivasa Iyer and

stage performer, and Subbulakshmi grew up in an
shaped by regular interactions with Karaikudi
anuja Iyengar.[5]

emy in 1929,when she was 13 years old . The
known for its discriminating selection process, and they
described as spellbinding and earned her many
erformances, Subbulakshmi became one the leading

M.S.Subbulakshmi

An EMI record of Subbulakshmi

Background information	
Also known as	M.S.
Born	September 16, 1916 Madurai, Madras Presidency, India
Origin	India
Died	December 11, 2004 (aged 88) Chennai, Tamil Nadu, India
Genres	Indian classical music
Occupations	Classical Vocalist
Years active	1930–2004
Labels	HMV

Conclusion & Future work

- We discussed approaches to enable dictionary based NER
- The specifics of the domain helps to reduce the complexity of NER.
- Dictionary expansion approaches to be compared

Future Work

- Relation extraction and mapping extracted relations to the pre-defined relations.
- Augmenting to existing music ontology.

References

1. Ananiadou S, McNaught J, (eds), Text Mining for Biology and Biomedicine. Artech House, London 2006.
2. Thelen, M., Riloff, E., A bootstrapping method for learning semantic lexicons using extraction pattern contexts. Proc. of the ACL-02 conference on Empirical methods in NLP, 2002
3. Yu, H., Hatzivassiloglou, V., Friedman, C. et al., ‘Automatic extraction of gene and protein synonyms from MEDLINE and journal articles’, in ‘Proceedings of the AMIA Symposium’, 9th–13th November, San Antonio, TX 2002.
4. Yu, H., Agichtein, E., ‘Extracting synonymous gene and protein terms from biological literature’, Bioinformatics, Vol. 19, Suppl., 2003.
5. Cohen, A. M., ‘Using symbolic network logical analysis as a knowledge extraction method on MEDLINE abstracts’, BMC Bioinformatics, 2005
6. Koike, Asako, Yoshiki Niwa, and Toshihisa Takagi. "Automatic extraction of gene/protein biological functions from biomedical text." Bioinformatics 21.7, 2005
7. Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006.

BACKUP

Thread title processing: Method

Each thread title is processed to find concepts involved

1. Identify different sections of title separated by certain special characters
2. If the title has PP phrase
   - i. get the components of the title

     <prop> of|with|between <topics>
   - ii. Check if there are different topics in <topics> separated by ',' or 'and|&'
   - iii. form a semantic relation from the components identified
3. If the title doesn't have PP phrase
   - i. get the topics in the title
   - ii. Check if there are different topics in <topics> separated by ',' or 'and|&'